Philipp Stöckle, Wolfgang Koppensteiner,
Katharina Korecky-Kröll, Maria Luisa Krapp, Tanja Wissik (eds.)

LEXICOGRAPHY IN THE AGE OF AI

PROCEEDINGS
OF THE XXII EURALEX
INTERNATIONAL CONGRESS 2026

29 September - 3 October 2026, Vienna | Austria



EURALEX
EURALEX AUSTRIA 2026
AUSTRIAN ACADEMY OF SCIENCES PRESS
ÖAW Austrian Academy of Sciences
acdh austrian centre for digital humanities

Editors: Philipp Stöckle, Wolfgang Koppensteiner, Katharina Korecky-Kröll, Maria Luisa Krapp, Tanja Wissik

Cover & title page design: Sandra Lehecka

Some rights reserved.
Copyright © Austrian Academy of Sciences, Vienna 2026
Typesetting: Sunrise Setting, Brixham, UK
ISBN: 978-3-7001-5137-1
https://epub.oeaw.ac.at/5137-1
https://verlag.oeaw.ac.at

FOREWORD

The **XXII EURALEX International Congress** was held on **29 September – 3 October 2026 in Vienna, Austria**.

The main topic of the XXII EURALEX International Congress was **Lexicography in the Age of AI**. The rapid development of artificial intelligence (AI) and natural language processing (NLP) has begun to transform the ways in which lexicographic data are compiled, analysed, presented, and used. Large language models, advanced corpus tools, and other machine learning methods offer unprecedented opportunities to accelerate tasks such as the extraction of lexical information, the induction of word senses, or the drafting of definitions. They also make possible new forms of user interaction, including conversational and personalized dictionaries that can adapt to different audiences or integrate multimodal content such as text, sound, and images.

At the same time, these technologies pose serious challenges that go beyond purely technical concerns. Automatically generated content raises questions of reliability, bias, and transparency, while the integration of opaque systems into lexicographic workflows risks undermining scholarly standards of quality and accountability. The ecological impact of large-scale AI models must also be taken into consideration, raising doubts about their long-term sustainability in resource-intensive projects such as lexicography. Finally, the introduction of AI into the field invites reflection on the role of the lexicographer: how can human expertise remain central when automation is increasingly pervasive, and what kinds of collaborations between humans and machines are both effective and responsible?

The conference provided a forum for discussing these promises and risks. We brought together professional lexicographers, publishers, researchers, software developers, and anyone interested in dictionaries to consider both the potential of AI for creating more dynamic, interactive dictionaries, and the ethical and ecological responsibilities that come with it. While many contributions engaged directly with the overarching theme of AI and its implications for lexicography, the conference programme also reflected the full breadth of the field of lexicography, encompassing diverse theoretical, methodological, and applied perspectives beyond the conference's main focus.

All in all, 144 presentations were held at the congress: five keynote talks, 114 paper presentations, 25 poster presentations, and five software demonstrations. This book contains 93 full papers submitted by presenters of talks, posters, and software demonstrations, as well as by presenters of the pro-conference workshops; not every presenter chose to submit a full paper for these proceedings. Two of the 93 full papers were contributed by presenters of the pre-conference workshop *Multi-Word Expressions and Phraseology: Corpus-Based and Computer-Processed*.

All submissions were reviewed in a double-blind peer review process by at least two members of the Scientific Committee (see page xv) to whom we are very grateful for their support. All decisions to accept or reject submissions for presentation at the congress and full papers for publication in the conference proceedings were based on the average score from reviews and in many cases on further evaluation by members of the Programme Committee (see page xv). We are very grateful to the EURALEX Board members who supported us as members of the Programme Committee: **Kristina Štrkalj Despot** (Croatian Academy of Sciences and Arts, Croatia), **Carolina Flinz** (University of Pavia, Italy), **Iztok Kosem** (Jožef Stefan Institute/University of Ljubljana, Slovenia), **Veronika Lipp** (ELTE Research Centre for Linguistics, Hungary), and **Geraint Paul Rees** (Universitat Pompeu Fabra, Spain). Without the expertise and commitment of all colleagues who served on the Scientific and the Programme Committee, we would

not have been able to maintain the high academic standard of presentations at EURALEX congresses and of their proceedings – thank you!

Preceding the main conference, four workshops were held on Tuesday, 29 September: the **8th Globalex Workshop on Lexicography and Neology (GWLN-8)**, organised by Ilan Kernerman and Kris Heylen in cooperation with the ENEOLI COST Action, which addressed the shifting role of the lexicographer in the age of AI and the detection, evaluation, and description of neologisms; **Constructicography in Practice: Constructions, Challenges, and Opportunities**, organised by Voula Giouli, Athina Sioupi, Nina Böbel, and Alexander Ziem, which brought together theoretical, descriptive, and computational perspectives on representing constructions in lexicographic resources; **Postediting Lexicography with Sketch Engine and Lexonomy**, organised by Ondřej Matuška, a hands-on tutorial on combining corpus- and AI-generated content with expert post-editing using Sketch Engine, Lexonomy, and the Dictionary Express method; and **Multi-Word Expressions and Phraseology: Corpus-Based and Computer-Processed**, organised by Lian Chen and Besim Kabashi, which examined how corpus linguistics, NLP, and AI are transforming the modelling, resource creation, and pedagogical and translational treatment of phraseology.

The conference programme was framed by five keynote talks that, together, mapped the current terrain of lexicography from computational, historical, cognitive, institutional, and technological angles. **Robert Lew** (*With Bots Like These, Who Needs Dictionaries?*) asked how generative AI is reshaping the reference landscape, arguing from recent studies that dictionaries – bilingual ones especially – remain competitive with chatbots for language learning. **Kate Wild** (*Beyond the Entry: The Evolution of Historical Lexicography at the Oxford English Dictionary*) traced a century of change at the OED, from strictly alphabetical revision to a data-driven practice treating the dictionary as a linked digital dataset. **Andreas Baumann** (*Polysemy: Operationalization, Emergence, Perception, and the Human in the Loop*) examined how polysemy can be measured by combining lexicographic and corpus data, extending the discussion to lexical sentiment and the continuing need for human judgement alongside NLP. **Ana Salgado** (*Dictionaries as Pipelines: Lexicographic Work at the Lisbon Academy of Sciences*) presented dictionary-making as a governed pipeline, linking a continuously updated dictionary, participatory lexicography, and a cross-national lexicographic platform of Portuguese. **Jan Hajič** (*Large Language Models, Symbolic Meaning Representations and Ontologies: Perspectives and Questions*) closed the series by bridging large language models and symbolic meaning representations, discussing European LLM development and how lexical resources can be unified across annotation frameworks. Across these five talks, the congress placed AI and computational methods in sustained dialogue with the historical, cognitive, and institutional foundations of lexicographic practice.

The conference itself was organised into nine thematic sections: **Computational & AI-Based Lexicography**, covering NLP, generative AI and large language models as method or subject of lexicographic work, from AI-assisted generation of definitions and entries to the evaluation of chatbots and LLMs; **Corpus Linguistics & Corpus-Based Resources**, addressing corpus compilation, analysis and exploitation as the basis for lexicographic description, including monitor corpora, frequency studies and trend-word detection; **Digital Lexicography: Design, Data Modelling & Theory**, concerned with e-lexicography, database and interface design, linked-data/OntoLex modelling, and metalexicographic theory; **Historical & Diachronic Lexicography**, focusing on etymology, historical and diachronic dictionaries, and the digitisation of historical sources and manuscripts; **Learner's Lexicography & Dictionary Use**, dedicated to learner's dictionaries, foreign/second-language pedagogy, and dictionary-use research; **Lexical Semantics, Neology & Phraseology**, treating word sense, polysemy, semantic change, neologisms, collocations, phraseology and idioms; **Morphology, Word Formation & Grammatical Lexicography**, covering derivation, valency, inflection and grammatical de-

scription in the dictionary; **Multilingualism, Language Varieties & Contact Lexicography**, spanning bilingual and multilingual dictionaries, language contact and borrowing, translation lexicography, and dialect, regional-variety and minority/low-resource-language lexicography; and **Specialised & Terminological Lexicography**, encompassing specialised dictionaries, terminology, language for special purposes and glossaries. The papers collected in this volume follow the same thematic organisation.

Alongside the parallel paper sessions, the congress also featured five software demonstrations and 25 poster presentations on Wednesday afternoon, showcasing tools and early findings spanning AI-assisted terminology and lexicography, historical and digital dictionary resources, and pedagogical applications, and giving work in progress a dedicated forum for direct exchange with participants.

The congress was organised by the Austrian Centre for Digital Humanities (ACDH) at the Austrian Academy of Sciences (ÖAW) in Vienna. Our sincere thanks go to all colleagues at the ACDH who supported the organisation of the congress and the publications of the Book of Abstracts as well as the present Conference Proceedings. We would also like to thank our colleagues from the ÖAW publishing house for the great collaboration and their flexibility in creating the publications.

Furthermore, we would like to thank our sponsors (see page xiv) who financially supported EURALEX 2026 and without whose generous support the congress could not have taken place.

As Chair of the XXII EURALEX Organising Committee, I would like to express my deep gratitude to all members of our Local Organising Committee (see page xv). Without their help and their great ideas along the way, the preparation of the conference would have been much more difficult (and much less fun). Finally, a huge thank you goes to my co-editors, **Wolfgang Koppensteiner**, **Katharina Korecky-Kröll**, **Maria Luisa Krapp**, and **Tanja Wissik**, whose tireless efforts were invaluable in bringing this publication into life.

Philipp Stöckle
Chair of EURALEX 2026
September 2026

TABLE OF CONTENTS

FOREWORD iii

ACKNOWLEDGEMENTS xiii

Sponsors xiv

Programme Committee xv

Scientific Committee xv

PART I CONFERENCE OVERVIEW xix

Keynotes xx

Pre-Conference Workshops xx

Talks xx

Posters xxvi

Software Demonstrations xxviii

PART II PROCEEDINGS

Computational & AI-Based Lexicography

Andrea Abel, Luca Ducceschi & Federico Villa 3
Detecting Semantic Variation between German Standard Varieties through Word Embedding Comparison

Iván Arias-Arias, María José Domínguez Vázquez, María Teresa Sanmarco Bande & Carlos Valcárcel Riveiro 16
From Content to Code and Back: Evaluating Agentic Coding in LLMs for Building R Shiny Dictionary Interfaces

Igor Boguslavsky, Vyacheslav Dikonov, Evgeniya Inshakova, Alexandre Lazursky, Svetlana Timoshenko & Tatiana Frolova 34
Lexicographic Resources for Inference-Based Semantic Text Processing System

Li Cui, Seol Namkung, Jun Lee, Hae-Yun Jung & Kilim Nam 46
What LLMs Learn and What They Miss: Frequency Effects in the Automated Description of Korean Verbal Inflection

Gilles-Maurice de Schryver 58
Chatbot Lexicography: Are GenAI LLMs Simply 'Dictionaries' in their Own Right and 'Lexicographers' Plain and Simple?

Gilles-Maurice de Schryver 88
The Ninth Art in Lexicography: The Power of Comic Strips

Hanna Fischer, Alfred Lameli, Nathalie Mederake & Nico Urbach 103
Linking Dictionaries Through Meaning: A Sense-Level Evaluation of Taxonomy-Guided LLM Classification in Lexicography

Katharina Korecky-Kröll, Philipp Stöckle, Daniel Elsner, Wolfgang Koppensteiner & Elisabeth Eder 117
LLM-Assisted Lexicography in Low-Resource Contexts: From Prompt to Dictionary Article

Iztok Kosem, Polona Gantar, Simon Krek & Tjaša Arčon 132
Divide and Define with LLMs: Generating Definitions for the Digital Dictionary Database of Slovene

František Kovařík, Marek Blahuš, Miloš Jakubíček & Vojtěch Kovář 143
Agree to Disagree: Intra-Annotator Agreement in Semi-Automatic Vocabulary Building

Barbara Lewandowska-Tomaszczyk 152
The Integration of British and American Slang in Contemporary Polish and the Role of LLM-Driven Dynamic Lexicography

Tinatin Margalitadze, Giorgi Meladze, Ketevan Mchedlishvili, Tamar Laluashvili & Giorgi Okropiridze 162
Effectiveness of AI for the Lexicography of Low-Resourced Languages: The Dictionary of Georgian Neologisms

Lydia Risberg, Kristina Koppel, Margit Langemets, Hanna Maask, Esta Prangel, Maria Tuulik & Silver Vapper 177
Are Bears More Offensive Than Dogs When Drunk? Using LLMs to Systematize the Labelling of Estonian Synonyms

Ines Röhrer & Manuel Raaf 193
Reflections on the Semantic Understanding of Generative AI: From LLM Experiments to Lexicographical Tools

Antonio San Martín & Catherine Trekker 203
Introducing Human-Centeredness in AI-Assisted Lexicography

Jesús Torres del Rey & M.a Teresa Fuentes Morán 217
A Hybrid Approach to Web Accessibility Evaluation in Digital Lexicography: Comparing and Integrating Rule-Based Validators and Generative AI

Thordis Ulfarsdottir & Ellert Thor Johannsson 233
The Reversal of a Bilingual Dictionary Using AI

Carlos Valcárcel Riveiro, María Teresa Sanmarco Bande, Iván Arias Arias & María José Domínguez Vázquez 245
Can Lexicographic Entries Embed Didactic Modules? A Case Study on LLM-Generated Outputs for a Dictionary Interface

Martina Vokáčová & Anna Marklová 259
Affordance Norms in Czech and English: Human-LLM Misalignment in Embodied Lexical Knowledge

Corpus Linguistics & Corpus-Based Resources

Pol Garriga Martínez 273
The Vocabulary of Mental Health: Exploitation of Specialized Diachronic Corpora with Artificial Intelligence for Qualitative Semantic Analysis

Mariia Kopylova 284
AI-Assisted Detection and Mitigation of Lexicographic Bias: Semantic Reduction in Definitions of Five Socio-Legal Terms

Aleksandra Marković, Cvetana Krstev, Milica Dinić Marinković & Ranka Stanković 295
Towards a Repository of Similes in Contemporary Serbian

Sanni Nimb, Ida Flörke, Agnes Aggergaard Mikkelsen, Nathalie Hau Norman & Nicolai Hartvig Sørensen 310
AI Used to Detect Problematic Content in Corpus Examples

Dušica Terzić, Dejana Mirković-Birtašić & Saša Marjanović 318
Towards a French–Serbian Dictionary of Sports Terms: Corpus-Based or AI-Assisted Methodology?

Ene Vainik, Heete Sahkai, Ahto Kiil, Jelena Kallas & Geda Paulsen 332
Semantic Clustering and Categorisation of Constructional Collexemes Using LLMs

Zongyuan Wang, Valeria Caruso & Anna De Meo 347
Collaborative Lexicography for Heritage Language Learning

Yelena Yerznkyan & Diana Movsisyan 359
Encoding Emotivity in Dictionary Definitions: A Corpus-Based Study of Verbs of Understanding

Digital Lexicography: Design, Data Modelling & Theory

Rukiye Smilla Burkart & Susanne Kabatnik 376
Wie definiert man das Unbekannte? Prompt Engineering als Methode der KI-gestützten Bedeutungsdefinition im Pandemictionary

Lian Chen, Damien Nouvel, Huy-Linh Dao & Alexander Delaporte 390
From Corpus to Dictionary: OntoLex-Lemon-Based Modelling of Neology in Chinese–French and Vietnamese–French Computational Lexicography

Flávia de Oliveira Maia-Pires & Silvio Licurgo Pires 405
User-Oriented Specialized Digital Lexicography and Terminological Accessibility: The LexiC Platform Applied to the Glossary of Brazilian Migration Legislation

Rufus H. Gouws & Theo J.D. Bothma 413
Optimising Interface Design in Online Lexicographic Products

David Lindemann 425
Developing a Digital Version of Gidor Bilbao's Latin-Basque Dictionary

Peter Meyer 437
An AI-Driven Natural Language Interface for Graph Database Queries in a Lexicographical Resource

Krzysztof Nowak & Dorota Mika 449
PoliLex: Modelling Heterogeneous Lexical Resources: Lessons from the Dictionary of Polish Dialects

Vasyl Starko & Andriy Rysin 463
Large-Scale Semiautomatic Detection of New and Unregistered Words in Modern Ukrainian

Pius ten Hacken 471
The Theoretical Foundation of Dictionary Definitions

Lars Trap-Jensen & Henrik Lorentzen 482
Revising the Usage Labelling System of the Danish Dictionary

Historical & Diachronic Lexicography

Petra Bago & Ivana Eterović 493
Compiling a 16th-Century Croatian Dictionary in the Age of AI: Challenges and Perspectives

Harald Bichlmeier 508
Vol. IX of the *Etymological Dictionary of Old High German [Etymologisches Wörterbuch des Althochdeutschen (EWA)]*: History Structure, Contents

Luisa Cimander 517
English Adjectives in German: Processes of Integration Since the 17th Century and Lexicographic Implications

Margarita Freixas, Esther Hernández & Marija Žarković Eriksson 523
e-*DRAE* 1884: Towards a Critical Digital and Hypertextual Edition of a Nineteenth-Century Dictionary

Ellert Thor Johannsson, Jóhannes B. Sigtryggsson & Thordis Ulfarsdottir 530
Unlocking a Nineteenth Century Dictionary: A Closer Look at Konráð Gíslason's *Dönsk orðabók* (1851)

Agnes Kim 537
Digital Metalexicography as a Tool for Historical (Contact) Linguistics

Tibor M. Pintér, Manana Ruseishvili-Cartledge, Dóra Pődör, Katalin P. Márkus & Marine Makhatadze 547
Sublexical Systems of Colour in Biblical Georgian and Hungarian: Findings of the Pentateuch

Ana Mihaljević & Josip Mihaljević 554
Artificial Intelligence as Support for the Digitization and Compilation of the *Dictionary of the Croatian Redaction of Church Slavonic*

Krzysztof Nowak, Iwona Krawczyk & Jagoda Marszałek 568
DiaLexPoL: LLM-assisted Sense Assignment for a Diachronic Dictionary of Latin

Saghar Sharifi & Abolfazl Alamdar 584
Towards a Handbook of Persian Lexicography: A Lexicographical-Functions Analysis of 20th–21st Century Dictionaries

Jóhannes B. Sigtryggsson 598
The Effects of Purism on a Nineteenth-Century Icelandic Bilingual Dictionary

Michael Spagnol, John J. Camilleri & Dwayne Ellul 607
From *Ḋamma* to *Ġabra*: Four Centuries of Maltese Lexicography

Tarrin Wills & Ellert Thor Johannsson 623
Can AI-Assisted Normalisation Improve Access to Historical Lexicographic Data?

Marija Žarković Eriksson 636
Legal Lexicon and the e-*DRAE* 1884: A Diachronic Study of Specialised Vocabulary in a General Dictionary Using a Digital Critical Edition

Learner's Lexicography & Dictionary Use

Ida Dringó-Horváth & Katalin P. Márkus 648
Dictionary Use in a Technology-Driven Era: Longitudinal Evidence from EFL and GFL Graduates

Christine Ganslmayer & Martina Nied Curcio 661
How German Language Textbooks Address the Digital Transformation of Lexicography

Zoe Gavriilidou & Evanthia Konstantinidou 676
Pedagogical Lexicography in Practice: Raising Dictionary Use Strategies through Card-Based Instruction

Laura Rebosio 683
Ostensive Aids in Learner's Dictionaries: A Diachronic, Comparative Analysis

Shigeru Yamada 698
Pedagogical Value and Insights from Comparing EFL Dictionaries: Evidence from *Apparently*, *Be on Speaking Terms*, and *Suffer (from)*

Lexical Semantics, Neology & Phraseology

Elena Berthemet 707
Designing an Idiomatic Dictionary

Sophia Fiedler 717
The Interactional Lexicon of Spoken German: A Functional, Audio-Based Lexicographic Resource

Ivana Filipović Petrović & Slobodan Beliga 729
Modelling Phraseological Semantic Structures in DMLex: Aligning a Croatian Lexonomy-Based Idiom Dictionary with a Thematic Index

Marta Garcia-Casado 743
New Senses, New Methods: Clustering as a Guide for Detecting Semantic Neologisms

Alexander Geyken & Gregor Middell 758
Trend Word Detection for DWDS, a Lexical Information System for Contemporary German

Ondřej Herman 768
Sense Classification for Word Sketches

Dominika Kováříková & Laura A. Janda 776
Emotions in Context: A Semantic Field Viewed through Grammar and Collocations

Gunta Nešpore-Bērzkalne, Mikus Grasmanis & Madara Stāde 788
Possibilities of Representation of Multi-Word Expressions in the Latvian Electronic Dictionary *Tēzaurs*

Emma Josefin Ölander Aadland 798
A Critical Lexicographical Discourse Analysis of New Adjectival Entries in the Standard Dictionaries for Norwegian

Linda Prossliner 813
AI-Supported Phraseography for Young Learners: Designing Definitions and Illustrations through AI – Opportunities and Challenges

Hana Skoumalová 829
Expanding the MWE Database with Variants

Petra Storjohann 836
The Interplay Between Individual New Lexical Items and Multi-Layered Discourse Representations of Neologisms

Silga Sviķe 849
A Lexicographical Analysis of Neologism Dictionaries Based on the Multilingual Dictionary of New Words (2010–2024)

Carole Tiberius, Martin Kroon, Lut Colman & Dalí Chirino 862
Ranking Collocations in *Woordcombinaties*

Ágoston Tóth 876
(UG)TPEX: A (Usage-Guided) Typical Example Score

Ágoston Tóth & Esra Abdelzaher 888
GPT as a Sense-Tagger for Selecting Dictionary Examples

Morphology, Word Formation & Grammatical Lexicography

Ben Black Bear, Jr., Richard Two Dogs & Jan Ullrich 902
Building the Largest Dictionary for an Indigenous Language of North America

Ivana Brač 920
Semantic Class Assignment in a Croatian Verb Lexicon: Challenges of Applying VerbNet's Classes

Katja Brankačkec & Barbora Štěpánková 932
Multifunctional Function Words in Upper Sorbian–Czech Dictionary Mudra 2.0

Polina Iaroshenko & Natalia Loukachevitch 944
Morphological Forms in Emotion Lexicons: Evidence from LLMs

Boris Kern 953
Word-Formation Chains in Slovenian: Evaluating the Effectiveness of AI

Václava Kettnerová, Veronika Kolářová & Jiří Mírovský 966
Where the Boundary between Derivational and Inflectional Morphology Meets Lexicography: Building the DeriVallex Lexicon

Veronika Kolářová, Václava Kettnerová, Michal Olbrich & Jiří Mírovský 978
NomVallex: Valency of Czech Nouns and Adjectives in Networks of Derivationally Related Words

Benedikt Perak, Dragana Špica & Irena Srdanović 991
A Graph-Based Lexicographic Workflow for Japanese Adjectives and Their Lexical Relations in an AI Tutor

Multilingualism, Language Varieties & Contact Lexicography

Luke Omoyemi Akinremi & Elena Martín-Cancela 1006
Bilingual Lexicography Under Resource Asymmetry: Designing Yoruba–German Dictionary Articles in Dialogue with Galician–German

Cormac Breathnach & Pádraig Ó Mianáin 1019
Reversing the New English–Irish Dictionary Database to Create Entry Frameworks for Irish–English and Monolingual Irish Dictionaries

Valeria Caruso & Geraint Paul Rees 1032
From Lexical Theory to AI Translation: Procomplement Verbs and ChatGPT

Zoe Gavriilidou 1043
Spanish Loanwords in Modern Greek: Linguistic Analysis and Lexicographic Treatment in the Dictionary of Standard Modern Greek

Ivana Lovrić Jović & Lana Hudeček 1054
Synonyms That Never Met: On the Compilation of the Dictionary of the Dubrovnik Idiom

Nika Stefan, Johan van der Zwaag & Hindrik Sijens 1065
Beyond the Single Standard: Addressing Variation and User Needs in an Online Frisian Dictionary

Specialised & Terminological Lexicography

Martina Bajčić & Ana Ostroški Anić 1077
Defining Legal Concepts in the AI Age: An Exploratory Study into Influencer Practices

Riccardo Billero 1087
Lexicography in the Age of AI: Training Systems with Human Expertise in Art Terminology

Marek Blahuš, Miloš Jakubíček, Katarína Petreková & Yingtian Wu 1100
Automatic Drafting of Terminological Dictionaries with One-Click Dictionary for 34 Languages

Barbara Heinisch 1107
Dictionaries, Glossaries and Terminology Databases in the Age of Artificial Intelligence: Rethinking Boundaries and Functions

Bruno Nahod & Marijana Tuta Dujmović 1114
From Text to Terms: TermScout, an AI Module for Early Terminology Processing

Deepak Solanki 1127
User-Oriented Specialized Lexicography: Corpus-Driven Models for Targeted Dictionary Design

Anna Vacalopoulou 1142
A Definition Walks into a Dictionary: The Lexicography of Humor

INDEX OF AUTHORS 1149

ACKNOWLEDGEMENTS

EURALEX 2026



ACKNOWLEDGEMENTS

We want to express our sincere gratitude to all sponsors who generously supported the XXII EURALEX International Congress financially. Your contributions were invaluable, and we greatly appreciate your commitment to the success of the event.

Sponsors

CLARIAH-AT

Stadt Wien

Hornby
A. S. Hornby Educational Trust CIO

SKETCH ENGINE

We would like to sincerely thank all those who invested their time and expertise in reviewing the submissions. Your invaluable work has significantly enhanced the quality of our conference, and we greatly appreciate your commitment and support.

Programme Committee

- **Carolina Flinz**, University of Pavia
- **Wolfgang Koppensteiner**, Austrian Academy of Sciences
- **Katharina Korecky-Kröll**, Austrian Academy of Sciences
- **Iztok Kosem**, Faculty of Arts, University of Ljubljana
- **Veronika Lipp**, ELTE Research Centre for Linguistics
- **Geraint Paul Rees**, Universitat Pompeu Fabra
- **Philipp Stöckle**, Austrian Academy of Sciences
- **Kristina Štrkalj Despot**, Institute of Croatian Language
- **Tanja Wissik**, Austrian Academy of Sciences

Local Organising Committee

- **Philipp Stöckle**, Austrian Academy of Sciences
- **Alexandra N. Lenz**, Austrian Academy of Sciences
- **Karlheinz Mörth**, Austrian Academy of Sciences
- **Amelie Dorn**, Austrian Academy of Sciences
- **Fabian Fleißner**, Austrian Academy of Sciences
- **Andreas Gellan**, Austrian Academy of Sciences
- **Wolfgang Koppensteiner**, Austrian Academy of Sciences
- **Katharina Korecky-Kröll**, Austrian Academy of Sciences
- **Maria Luisa Krapp**, Austrian Academy of Sciences
- **Markus Kunzmann**, Austrian Academy of Sciences
- **Sandra Lehecka**, Austrian Academy of Sciences
- **Tanja Wissik**, Austrian Academy of Sciences
- **Anja Wittibschlager**, Austrian Academy of Sciences
- **Anna Woldrich**, Austrian Academy of Sciences

Scientific Committee

- **Andrea Abel**, European Academy Bolzano/Bozen (EURAC)
- **Arleta Adamska-Sałaciak**, Adam Mickiewicz University, Poznań
- **Špela Arhar Holdt**, Centre for Language Resources and Technologies (CJVT), University of Ljubljana
- **Petra Bago**, Faculty of Humanities and Social Sciences, University of Zagreb
- **Martina Bajčić**, Faculty of Law, University of Rijeka
- **Vincent Balnat**, University of Strasbourg
- **Verginica Barbu Mititelu**, Romanian Academy Research Institute for Artificial Intelligence

SCIENTIFIC COMMITTEE

- **Slobodan Beliga**, Department of Informatics, University of Rijeka
- **Vladimír Benko**, Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences
- **Úna Bhreathnach**, Fiontar & Scoil na Gaeilge, Dublin City University
- **Hans Bickel**, Schweizerisches Idiotikon
- **Goranka Blagus Bartolec**, Institute of Croatian Language
- **Petar Božović**, University of Montenegro
- **Ivana Brač**, Institute of Croatian Language
- **Maja Bratanić**, Institute of Croatian Language
- **Thomas Burch**, Trier Center for Digital Humanities
- **Valeria Caruso**, Università degli Studi di Napoli „l'Orientale"
- **Lian Chen**, LLL, University of Orléans
- **Rute Costa**, Universidade NOVA de Lisboa
- **Janet Decesaris**, Universitat Pompeu Fabra
- **Idalete Dias**, University of Minho
- **Amelie Dorn**, Austrian Academy of Sciences
- **María José Domínguez Vázquez**, Universidade de Santiago de Compostela
- **Amelie Dorn**, Austrian Academy of Sciences
- **Anna Dziemianko**, Adam Mickiewicz University, Poznań
- **Pamela Faber**, University of Granada
- **Ilse Feinauer**, University of Stellenbosch
- **Ivana Filipović Petrović**, Croatian Academy of Sciences
- **Edward Finegan**, University of Southern California
- **Fabian Fleißner**, Austrian Academy of Sciences
- **Carolina Flinz**, University of Pavia
- **Thierry Fontenelle**, European Investment Bank
- **Judit Freixa**, Universitat Pompeu Fabra
- **Polona Gantar**, University of Ljubljana
- **Radovan Garabík**, Ľ. Štúr Institute of Linguistics, Slovak Academy of Sciences
- **Zoe Gavriilidou**, Democritus University of Thrace
- **Andreas Gellan**, Austrian Academy of Sciences
- **Alexander Geyken**, Berlin-Brandenburg Academy of Sciences and Humanities
- **Voula Giouli**, Aristotle University of Thessaloniki
- **Volker Harm**, Göttingen Academy of Sciences and Humanities
- **Kris Heylen**, Dutch Language Institute (INT, Leiden) / KU Leuven (University of Leuven), Research Unit QLVL
- **Lana Hudeček**, Institute of Croatian Language
- **Max Ionov**, University of Zaragoza
- **Dubravka Ivšić Majić**, Institute of Croatian Language

SCIENTIFIC COMMITTEE

- **Miloš Jakubíček**, Lexical Computing
- **Besim Kabashi**, University of Tübingen
- **Jelena Kallas**, Institute of the Estonian Language
- **Virna Karlić**, Faculty of Humanities and Social Sciences, University of Zagreb
- **Boris Kern**, Fran Ramovš Institute of the Slovenian Language
- **Ilan Kernerman**, K Dictionaries
- **Fahad Khan**, Istituto di Linguistica Computazionale „Antonio Zampolli“
- **Svetla Koeva**, Institute for Bulgarian Language
- **Veronika Kolářová**, Institute of Formal and Applied Linguistics, Charles University, Faculty of Mathematics and Physics
- **Kristina Koppel**, Institute of the Estonian Language
- **Wolfgang Koppensteiner**, Austrian Academy of Sciences
- **Katharina Korecky-Kröll**, Austrian Academy of Sciences
- **Iztok Kosem**, Faculty of Arts, University of Ljubljana
- **Maria Luisa Krapp**, Austrian Academy of Sciences
- **Simon Krek**, Jožef Stefan Institute
- **Margit Langemets**, Institute of the Estonian Language
- **Alexandra N. Lenz**, Austrian Academy of Sciences
- **Marie-Claude L'Homme**, Observatoire de linguistique Sens-Texte
- **David Lindemann**, University of the Basque Country
- **Veronika Lipp**, ELTE Research Centre for Linguistics
- **Anja Lobenstein-Reichmann**, Göttingen Academy of Sciences and Humanities
- **Irina Lobzhanidze**, Ilia State University
- **Henrik Lorentzen**, Society for Danish Language and Literature
- **Magdalena Majdak**, Institute of Polish Language, Polish Academy of Sciences
- **Tinatin Margalitadze**, Ilia State University
- **Saša Marjanović**, Faculty of Philology, University of Belgrade
- **Ivana Matas Ivanković**, Institute of Croatian Language
- **Peter Meyer**, Leibniz Institute for the German Language (IDS) Mannheim
- **Milica Mihaljević**, Institute of Croatian Language
- **Karlheinz Mörth**, Austrian Academy of Sciences
- **Fabio Mollica**, University of Milan
- **Hilary Nesi**, Coventry University
- **Henrik Nilsson**, European Association for Terminology, C.A.G
- **Pär Nilsson**, Swedish Academy Dictionary
- **Vincent Ooi**, National University of Singapore
- **Christian-Emil Ore**, University of Oslo
- **Petya Osenova**, Sofia University / Institute of Information and Communication Technologies Bulgarian Academy of Sciences

SCIENTIFIC COMMITTEE

- **Ana Ostroški Anić**, Institute of Croatian Language
- **Gokhan Ozkan**, Bogazici University
- **Martina Pavić**, Institute of Croatian Language
- **Sanja Perić Gavrančić**, Institute of Croatian Language
- **Ralf Plate**, Academy of Sciences and Literature Mainz
- **Boris Pritchard**, University of Rijeka
- **Natascia Ralli**, Institute for Applied Research, Eurac Research
- **Geraint Paul Rees**, Universitat Pompeu Fabra
- **Irene Renau**, Universitat Autònoma de Barcelona
- **Ana Salgado**, University of Porto & Academy of Sciences of Lisbon
- **Hindrik Sijens**, Fryske Akademy
- **Ranka Stanković**, University of Belgrade
- **Egon W. Stemle**, Eurac Research
- **Frieda Steurs**, Faculty of Arts, KU Leuven
- **Philipp Stöckle**, Austrian Academy of Sciences
- **Kristina Štrkalj Despot**, Institute of Croatian Language
- **Elsabe Taljard**, University of Pretoria
- **Giovanni Tallarico**, University of Verona
- **Pius ten Hacken**, University of Innsbruck
- **Carole Tiberius**, Dutch Language Institute / Leiden University Centre for Linguistics
- **Lars Trap-Jensen**, Society for Danish Language and Literature
- **Anna Vacalopoulou**, Athena RC (ILSP)
- **Carlos Valcárcel Riveiro**, Universidade de Vigo
- **Federica Vezzani**, University of Padua
- **Craig Volker**, James Cook University
- **Tanja Wissik**, Austrian Academy of Sciences
- **Sascha Wolfer**, Leibniz Institute for the German Language (IDS) Mannheim
- **Mojca Žagar Karer**, Research Centre of the Slovenian Academy of Sciences and Arts (ZRC SAZU)
- **Alexander Ziem**, Heinrich Heine University Düsseldorf
- **Tanara Zingano Kuhn**, University of Coimbra

Special thanks

We also extend our special thanks to Leona Fischer, Hannah Grischany, Magdalena Kirnbauer, Simone Sailer, Sonja Schwaiger, Eleni Steinborn and Anja Wittibschlager for their careful support in ensuring that this book meets all formal requirements.

PART I

CONFERENCE OVERVIEW

EURALEX 2026



CONFERENCE OVERVIEW

Keynotes

- *Robert Lew*
  With Bots Like These, Who Needs Dictionaries?
- *Kate Wild*
  Beyond the Entry: The Evolution of Historical Lexicography at the Oxford English Dictionary
- *Andreas Baumann*
  Polysemy: Operationalization, Emergence, Perception, and the Human in the Loop
- *Ana Salgado*
  Dictionaries as Pipelines: Lexicographic Work at the Lisbon Academy of Sciences
- *Jan Hajič*
  Large Language Models, Symbolic Meaning Representations and Ontologies: Perspectives and Questions

Pre-Conference Workshops

- *Ilan Kernerman & Kris Heylen*
  The 8th Globalex Workshop on Lexicography and Neology (GWLN-8)
- *Voula Giouli, Athina Sioupi, Nina Böbel & Alexander Ziem*
  Constructicography in Practice: Constructions, Challenges, and Opportunities
- *Ondřej Matuška*
  Postediting lexicography with Sketch Engine and Lexonomy
- *Lian Chen & Besim Kabashi*
  Multi-Word Expressions and Phraseology: Corpus-Based and Computer-Processed

Talks

1. *Andrea Abel, Luca Ducceschi & Federico Villa*
   Detecting Semantic Variation between German Standard Varieties through Word Embedding Comparison
2. *Luke Omoyemi Akinremi & Elena Martín-Cancela*
   Bilingual Lexicography Under Resource Asymmetry: Designing Yoruba–German Dictionary Articles in Dialogue with Galician–German
3. *Iván Arias-Arias, María José Domínguez Vázquez, María Teresa Sanmarco Bande & Carlos Valcárcel Riveiro*
   From Content to Code and Back: Evaluating Agentic Coding in LLMs for Building R Shiny Dictionary Interfaces
4. *Petra Bago & Ivana Eterović*
   Compiling a 16th-Century Croatian Dictionary in the Age of AI: Challenges and Perspectives
5. *Martina Bajčić & Ana Ostroški Anić*
   Defining Legal Concepts in the AI Age: An Exploratory Study into Influencer Practices
6. *Michela Bandini, Silvia Piccini, Andrea Bellandi & Emiliano Giovannetti*
   Toward a Narrative-Oriented Dictionary: Computational Modelling of Youth Language in Socially Vulnerable Contexts
7. *Harald Bichlmeier*
   Vol. IX of the *Etymological Dictionary of Old High German [Etymologisches Wörterbuch des Althochdeutschen (EWA)]:* History Structure, Contents

8. *Riccardo Billero*
   Lexicography in the Age of AI: Training Systems with Human Expertise in Art Terminology
9. *Ben Black Bear, Jr., Richard Two Dogs & Jan Ullrich*
   Building the Largest Dictionary for an Indigenous Language of North America
10. *Marek Blahuš, Miloš Jakubíček, Katarína Petreková & Yingtian Wu*
    Automatic Drafting of Terminological Dictionaries with One-Click Dictionary for 34 Languages
11. *Igor Boguslavsky, Vyacheslav Dikonov, Evgeniya Inshakova, Alexandre Lazursky, Svetlana Timoshenko & Tatiana Frolova*
    Lexicographic Resources for Inference-Based Semantic Text Processing System
12. *Biljana Božinovski, Makimi Kano, Peter Gobel & Mihaela Franjić*
    Lexicographic Functions without Dictionaries: GenAI Mediated Lexical Problem Solving Among Non English Major Tertiary EFL Students
13. *Ivana Brač*
    Semantic Class Assignment in a Croatian Verb Lexicon: Challenges of Applying VerbNet's Classes
14. *Cormac Breathnach & Pádraig Ó Mianáin*
    Reversing the New English–Irish Dictionary Database to Create Entry Frameworks for Irish–English and Monolingual Irish Dictionaries
15. *Gabriela Brůhová & Veronika Raušová*
    The "Hell Of A" Construction in Casual Written English Online: A Synchronic Functional Account
16. *Rukiye Smilla Burkart & Susanne Kabatnik*
    How Do You Define the (Known) Unknown? Prompt Engineering as a Method for AI-Assisted Meaning Definition in the Pandemictionary
17. *Edmond Cane*
    Learning Like Humans, Parsing Unlike Machines: A Construction-Based Approach to Lexical Annotation
18. *Valeria Caruso & Geraint Paul Rees*
    From Lexical Theory to AI Translation: Procomplement Verbs and ChatGPT
19. *Elina Chadjipapa, Zoe Gavriilidou, Maria Mitsiaki & Ifigeneia Dosi*
    Chatbots in Pedagogical Lexicography Evaluating AI-Generated Dictionary Entries from Three LLMs
20. *Lian Chen, Damien Nouvel, Huy-Linh Dao & Alexander Delaporte*
    From Corpus to Dictionary: OntoLex-Lemon-Based Modelling of Neology in Chinese–French and Vietnamese–French Computational Lexicography
21. *John J. Chew III, Emma Ferrett & Anastasia Riehl*
    Canadian English Dictionary: A Progress Report
22. *Li Cui, Seol Namkung, Jun Lee, Hae-Yun Jung & Kilim Nam*
    What LLMs Learn and What They Miss: Frequency Effects in the Automated Description of Korean Verbal Inflection
23. *Václav Cvrček, Martin Stluka & Klára Pivoňková*
    Tracking Eight Centuries: The Czech Monitor Corpus as a Resource for Diachronic Lexical Research
24. *Gilles-Maurice de Schryver*
    Chatbot Lexicography: Are GenAI LLMs Simply 'Dictionaries' in their Own Right and 'Lexicographers' Plain and Simple?

25. *Gilles-Maurice de Schryver*
The Ninth Art in Lexicography: The Power of Comic Strips

26. *Veronique De Tier, Katrien Depuydt, Koen Mertens & Mathieu Fannee*
From Pilot to Practice: Two Years of Progress in Developing a Shared Infrastructure for Dutch Dialect Dictionaries

27. *Kristina Š. Despot, Ana Ostroški Anić & Matej Klemen*
Beyond Binary Metaphor Detection: Dataset-Driven Evaluation of Conceptual Metaphor Type Recognition in Croatian and Slovene

28. *Ida Dringó-Horváth & Katalin P. Márkus*
Dictionary Use in a Technology-Driven Era: Longitudinal Evidence from EFL and GFL Graduates

29. *Emma Ferrett, Amir Ghorvei & Anastasia Riehl*
A Dictionary for All Canadians: Establishing an Inclusive Standard for the Canadian English Dictionary

30. *Sophia Fiedler*
The Interactional Lexicon of Spoken German: A Functional, Audio-Based Lexicographic Resource

31. *Ivana Filipović Petrović & Slobodan Beliga*
Modelling Phraseological Semantic Structures in DMLex: Aligning a Croatian Lexonomy-Based Idiom Dictionary with a Thematic Index

32. *Hanna Fischer, Alfred Lameli, Nathalie Mederake & Nico Urbach*
Linking Dictionaries Through Meaning: A Sense-Level Evaluation of Taxonomy-Guided LLM Classification in Lexicography

33. *Emmanuel Sylvain Fomat*
Camfranglais in the Age of Artificial Intelligence: Lexicographic Challenges, Opportunities and a Human-AI Approach to Documenting Linguistic Hybridity in Multilingual Cameroon

34. *Daniele Franceschi*
Gender Representation and Editorial Responsibility in the OED: A case study of *female*, *feminine*, *male* and *masculine*

35. *Marta Garcia-Casado*
New Senses, New Methods: Clustering as a Guide for Detecting Semantic Neologisms

36. *Pol Garriga Martínez*
The Vocabulary of Mental Health: Exploitation of Specialized Diachronic Corpora with Artificial Intelligence for Qualitative Semantic Analysis

37. *Zoe Gavriilidou*
Spanish Loanwords in Modern Greek: Linguistic Analysis and Lexicographic Treatment in the Dictionary of Standard Modern Greek

38. *Alexander Geyken & Gregor Middell*
Trend Word Detection for DWDS, a Lexical Information System for Contemporary German

39. *Rufus H. Gouws & Theo J.D. Bothma*
Optimising Interface Design in Online Lexicographic Products

40. *Wanjing Han & Gilles-Maurice de Schryver*
Corpus-Based and AI-Assisted Design of a Chinese–English Business Dictionary

41. *Ondřej Herman*
Sense Classification for Word Sketches

42. *Kris Heylen, Katrien Depuydt, Carole Tiberius & Jesse de Does*
Building a Sense Inventory for the Central Lexicographic Infrastructure of Dutch

43. *Thordis Ulfarsdottir & Ellert Thor Johannsson*
The Reversal of a Bilingual Dictionary Using AI

44. *Željko Jozić, Marijana Horvat & Martina Horvat*
Digitization of Adam Patačić's Manuscript Dictionary

45. *Boris Kern*
Word-Formation Chains in Slovenian: Evaluating the Effectiveness of AI

46. *Václava Kettnerová, Veronika Kolářová & Jiří Mírovský*
Where the Boundary between Derivational and Inflectional Morphology Meets Lexicography: Building the DeriVallex Lexicon

47. *Agnes Kim*
Digital Metalexicography as a Tool for Historical (Contact) Linguistics

48. *Veronika Kolářová, Václava Kettnerová, Michal Olbrich & Jiří Mírovský*
NomVallex: Valency of Czech Nouns and Adjectives in Networks of Derivationally Related Words

49. *Mariia Kopylova*
AI-Assisted Detection and Mitigation of Lexicographic Bias: Semantic Reduction in Definitions of Five Socio-Legal Terms

50. *Katharina Korecky-Kröll, Philipp Stöckle, Daniel Elsner, Wolfgang Koppensteiner & Elisabeth Eder*
LLM-Assisted Lexicography in Low-Resource Contexts: From Prompt to Dictionary Article

51. *Iztok Kosem, Polona Gantar, Simon Krek & Tjaša Arčon*
Divide and Define with LLMs: Generating Definitions for the Digital Dictionary Database of Slovene

52. *František Kovařík, Marek Blahuš, Miloš Jakubíček & Vojtěch Kovář*
Agree to Disagree: Intra-Annotator Agreement in Semi-Automatic Vocabulary Building

53. *Dominika Kováříková & Laura A. Janda*
Emotions in Context: A Semantic Field Viewed through Grammar and Collocations

54. *Simon Krek, Iztok Kosem & Polona Gantar*
Bridging Lexicography and NLP: The Architecture and Application of the Digital Dictionary Database of Slovene (DDDS)

55. *Nina Ledinek, Mija Michelizza, Špela Petric Žižić, Domen Krvina, Janoš Ježovnik & Andrej Perdih*
Using LLMs in Dictionary Compilation: The Case of Classifying Adjectives in the Dictionary of the Slovenian Standard Language eSSKJ

56. *Seulki Lee*
Reframing Wan (完), Hao (好), and Dao (到) in Chinese–Korean Bilingual Lexicography: A Corpus-Based Cognitive Analysis

57. *Barbara Lewandowska-Tomaszczyk*
The Integration of British and American Slang in Contemporary Polish and the Role of LLM-Driven Dynamic Lexicography

58. *David Lindemann*
Developing a Digital Version of Gidor Bilbao's Latin-Basque Dictionary

59. *Ivana Lovrić Jović & Lana Hudeček*
Synonyms That Never Met: On the Compilation of the Dictionary of the Dubrovnik Idiom

60. *Tinatin Margalitadze, Giorgi Meladze, Ketevan Mchedlishvili, Tamar Laluashvili & Giorgi Okropiridze*
Effectiveness of AI for the Lexicography of Low-Resourced Languages: The Dictionary of Georgian Neologisms

61. *Aleksandra Marković, Cvetana Krstev, Milica Dinić Marinković & Ranka Stanković*
Towards a Repository of Similes in Contemporary Serbian

TALKS

62. *Leonor Martins & Ana Salgado*
Neologisms in Evaristo.ai *vs.* ChatGPT: Implications for LLM-Based Lexicography in European Portuguese

63. *Ondřej Matuška, Miloš Jakubíček & Vojtěch Kovář*
Management of Large Lexicographic Projects with Lexonomy

64. *Cordula Meißner, Anna-Lena Randermann & Janina Deilke*
Indexicality in the Vocabulary of German Communication Verbs: Examining Language Use and Speaker Knowledge as a Basis for Lexicographical Conception

65. *Peter Meyer*
An AI-Driven Natural Language Interface for Graph Database Queries in a Lexicographical Resource

66. *Tomasz Michta & Tomasz Czyż*
ChatGPT or the Dictionary? Resolving Polysemy in L2 Reception

67. *Tomasz Michta & Katarzyna Mroczyńska*
Spoilt for Choice? ChatGPT Versus a Bilingual Dictionary in Selecting L2 Equivalents for Polysemous L1 Words

68. *Tomasz Michta, Geraint Rees & Ana Frankenberg-Garcia*
Corpus-Based versus AI-Generated Dictionary Examples: How They Differ and What Learners Prefer

69. *Ana Mihaljević & Josip Mihaljević*
Artificial Intelligence as Support for the Digitization and Compilation of the *Dictionary of the Croatian Redaction of Church Slavonic*

70. *Julie Moore*
Directions in Lexicography in the Global South

71. *Phoebe Nicholson, Will Rogers & Kate Wild*
Historical Lexicography in the Age of AI: A Progress Report from the OED

72. *Martina Nied Curcio & Christine Ganslmayer*
How German Language Textbooks Address the Digital Transformation of Lexicography

73. *Krzysztof Nowak, Iwona Krawczyk & Jagoda Marszałek*
DiaLexPoL: LLM-Assisted Sense Assignment for a Diachronic Dictionary of Latin

74. *Krzysztof Nowak & Dorota Mika*
PoliLex: Modelling Heterogeneous Lexical Resources: Lessons from the Dictionary of Polish Dialects

75. *Emma Josefin Ölander Aadland*
A Critical Lexicographical Discourse Analysis of New Adjectival Entries in the Standard Dictionaries for Norwegian

76. *Benedikt Perak, Dragana Špica & Irena Srdanović*
A Graph-Based Lexicographic Workflow for Japanese Adjectives and Their Lexical Relations in an AI Tutor

77. *Ana Ostroški Anić, Martina Pavić & Lidija Cvikić*
Types of Definitions in Croatian Primary School Science Textbooks

78. *Andrej Perdih, Janoš Ježovnik & Dejan Gabrovšek*
Improving Dictionary Headword Lists with Word-Prevalence Metrics: An Examination of Low-Prevalence, High-Frequency Words in Slovenian

79. *Silvia Piccini, Andrea Bellandi, Giuliana Elizabeth Vilela Ruiz & Davide Saponaro*
A Semantic Web Model for Diachronic Explanatory Combinatorial Dictionaries: Capturing Conceptual Change and Denominational Variation

80. *Linda Prossliner*
AI-Supported Phraseography for Young Learners: Designing Definitions and Illustrations through AI – Opportunities and Challenges

81. *Laura Rebosio*
Ostensive Aids in Learner's Dictionaries: A Diachronic, Comparative Analysis

82. *Lydia Risberg, Kristina Koppel, Margit Langemets, Hanna Maask, Esta Prangel, Maria Tuulik & Silver Vapper*
Are Bears More Offensive Than Dogs When Drunk? Using LLMs to Systematize the Labelling of Estonian Synonyms

83. *Ines Röhrer & Manuel Raaf*
Reflections on the Semantic Understanding of Generative AI: From LLM Experiments to Lexicographical Tools

84. *Antonio San Martín & Catherine Trekker*
Introducing Human-Centeredness in AI-Assisted Lexicography

85. *Barbara Schmidt-Thieme, Theresa Kruse & Ulrich Heid*
Lexical and Conceptual Tools for Learning Mathematics

86. *Saghar Sharifi & Abolfazl Alamdar*
Towards a Handbook of Persian Lexicography: A Lexicographical-Functions Analysis of 20th–21st Century Dictionaries

87. *Jóhannes B. Sigtryggsson*
The Effects of Purism on a Nineteenth-Century Icelandic Bilingual Dictionary

88. *Deepak Solanki*
User-Oriented Specialized Lexicography: Corpus-Driven Models for Targeted Dictionary Design

89. *Michael Spagnol, John J. Camilleri & Dwayne Ellul*
From *Ḋamma* to *Ġabra*: Four Centuries of Maltese Lexicography

90. *Stefania Spina, Irene Fioravanti & Fabio Zanda*
The Learner Dictionary of Italian Collocations (DICI-A): A New Lexicographic Resource

91. *Vasyl Starko & Andriy Rysin*
Large-Scale Semiautomatic Detection of New and Unregistered Words in Modern Ukrainian

92. *Nika Stefan, Johan van der Zwaag & Hindrik Sijens*
Beyond the Single Standard: Addressing Variation and User Needs in an Online Frisian Dictionary

93. *Petra Storjohann*
The Interplay Between Individual New Lexical Items and Multi-Layered Discourse Representations of Neologisms

94. *Pius ten Hacken*
The Theoretical Foundation of Dictionary Definitions

95. *Dušica Terzić, Dejana Mirković-Birtašić & Saša Marjanović*
Towards a French–Serbian Dictionary of Sports Terms: Corpus-Based or AI-Assisted Methodology?

96. *Carole Tiberius, Martin Kroon, Lut Colman & Dalí Chirino*
Ranking Collocations in *Woordcombinaties*

97. *Jesús Torres del Rey & María Teresa Fuentes Morán*
A Hybrid Approach to Web Accessibility Evaluation in Digital Lexicography: Comparing and Integrating Rule-Based Validators and Generative AI

98. *Ágoston Tóth*
(UG)TPEX: A (Usage-Guided) Typical Example Score

TALKS

99. *Ágoston Tóth & Esra Abdelzaher*
GPT as a Sense-Tagger for Selecting Dictionary Examples

100. *Lars Trap-Jensen & Henrik Lorentzen*
Revising the Usage Labelling System of the Danish Dictionary

101. *Ene Vainik, Heete Sahkai, Ahto Kiil, Jelena Kallas & Geda Paulsen*
Semantic Clustering and Categorisation of Constructional Collexemes Using LLMs

102. *Carlos Valcárcel Riveiro, María Teresa Sanmarco Bande, Iván Arias Arias and María José Domínguez Vázquez*
Can Lexicographic Entries Embed Didactic Modules? A Case Study on LLM-Generated Outputs for a Dictionary Interface

103. *Martina Vokáčová & Anna Marklová*
Affordance Norms in Czech and English: Human-LLM Misalignment in Embodied Lexical Knowledge

104. *Zongyuan Wang, Valeria Caruso & Anna De Meo*
Collaborative Lexicography for Heritage Language Learning

105. *Tarrin Wills & Ellert Thor Johannsson*
Can AI-Assisted Normalisation Improve Access to Historical Lexicographic Data?

106. *Jonas Winnerlöv*
Learner's Dictionaries Based on Etymological Principles: The Example of a German Textbook Dictionary for Swedish Learners

107. *Sylwia Wojciechowska*
Exploring Large Language Models in Word Sense Disambiguation: A Study of Corpus Examples for Perception Verbs in Learners' Dictionaries

108. *Shigeru Yamada*
Pedagogical Value and Insights from Comparing EFL Dictionaries: Evidence from *Apparently*, *Be on Speaking Terms*, and *Suffer (from)*

109. *Yelena Yerznkyan & Diana Movsisyan*
Encoding Emotivity in Dictionary Definitions: A Corpus-Based Study of Verbs of Understanding

110. *Akerke Yessenali*
Phraseological Theoretical Gaps in Digital Lexicography: A Contrastive Study of English–German Phraseologisms with Gender Representations

111. *Manjola Zaçellari*
The Making of Meaning: From the 1980 Communist Albanian Dictionary to the Great Dictionary of 2026

112. *Marija Žarković Eriksson*
Legal Lexicon and the e-*DRAE* 1884: A Diachronic Study of Specialised Vocabulary in a General Dictionary Using a Digital Critical Edition

113. *Tanara Zingano Kuhn, Vojtěch Kovář & Miloš Jakubíček*
Mapping Sociolinguistic Theory into Practice: Headword Candidate Selection in the Dictionary of Pluricentric Portuguese

114. *Valeria Zotti & Rita Gramellini*
From Plurilingual Lexicography and Terminography to Generative AI Translation: Potential, Limits, and the Enduring Role of Dictionaries

Posters

1. *Mikel Alonso*
AI-Assisted Information Extraction from Larramendi's Trilingual Dictionary (1745)

2. *Maria Ammari*
   Pedagogy as the Core of Education: A Domain-Analytic Route to a Corpus-Based Dictionary
3. *Elena Berthemet*
   Designing an Idiomatic Dictionary
4. *Luisa Cimander*
   English Adjectives in German: Processes of Integration Since the 17th Century and Lexicographic Implications
5. *Luisa Cimander & Petra Storjohann*
   Tracing Neologisms in Neo-Discourses: Diachronic and Cross-Domain Perspectives
6. *Margarita Freixas, Esther Hernández & Marija Žarković Eriksson*
   e-*DRAE* 1884: Towards a Critical Digital and Hypertextual Edition of a Nineteenth-Century Dictionary
7. *Anne-Kathrin Gärtig-Bressan & Pia Carmela Lombardi*
   Bilingual Dictionaries with Italian in the 19th and 20th Centuries: The ALON Portal
8. *Zoe Gavriilidou & Evanthia Konstantinidou*
   Pedagogical Lexicography in Practice: Raising Dictionary Use Strategies through Card-Based Instruction
9. *Marius Glebus*
   Choosing Optimal AI Prompt Instructions for Neologism Detection
10. *Barbara Heinisch*
    Dictionaries, Glossaries and Terminology Databases in the Age of Artificial Intelligence: Rethinking Boundaries and Functions
11. *Ellert Thor Johannsson, Jóhannes B. Sigtryggsson & Thordis Ulfarsdottir*
    Unlocking a Nineteenth Century Dictionary: A Closer Look at Konráð Gíslason's *Dönsk orðabók* (1851)
12. *Amila Kugic*
    Analysis of Synthetic Clinical Neologisms: A Corpus-Based Analysis of Large Language Model-Generated Content in the Context of Artificial Intelligence-Assisted Clinical Documentation
13. *Magdalena Majdak, Ewa Rodek, Katarzyna Kryńska & Jagoda Marszałek*
    The Role of LLMs in the Automatic Distribution of Meanings from Knapiusz's Polish–Latin–Greek Dictionary of 1643: A Pilot Study
14. *Olga Menagarishvili*
    Dictionaries of Engineering in the Age of AI
15. *Tibor M. Pintér, Manana Ruseishvili, Dóra Pődör, Katalin P. Márkus & Marine Makhatadze*
    Sublexical Systems of Colour in Biblical Georgian and Hungarian: Findings of the Pentateuch
16. *Perina Vukša Nahod*
    From Field Recordings to an Online Dictionary: A Model for Processing Štokavian Dialect Material
17. *Sanni Nimb, Ida Flörke, Agnes Aggergaard Mikkelsen, Nathalie Hau Norman & Nicolai Hartvig Sørensen*
    AI Used to Detect Problematic Content in Corpus Examples
18. *Iryna Ostapova, Mykyta Yablochkov, Alona Dorozhynska & Iuliia Verbynenko*
    Development of a Web Application for a Multilingual Terminological Dictionary
19. *Ana Ostroški Anić, Federica Vezzani, Kris Heylen & Ilan Kernerman*
    From Detection to Standardization: Practices and Criteria for Terminological Neology
20. *Geraint Paul Rees, Blanca Arias-Badia, Elisenda Bernal, Marta Brescia-Zapata, Clara Anush Del Rey Adanalian, Ana Frankenberg-Garcia, Marta Garcia Casado, Paula Igareda González, Daniel*

Jiménez Casas, Robert Lew, Alba Milà Garcia, Guiomar Salvat, Aurora Troncoso Ruiz, Sergi Torner & Patrick Zabalbeascoa
Accessible Digital Language Resources (ADLaR)

21. *Nicolai Hartvig Sørensen & Andrea Stengaard*
Beyond Word Lists – But Not Beyond Dictionaries

22. *Silga Sviķe*
A Lexicographical Analysis of Neologism Dictionaries Based on the *Multilingual Dictionary of New Words* (2010–2024)

23. *Maria Tuulik, Kristina Koppel, Ene Vainik, Margit Langemets, Lydia Risberg, Hanna Maask, Esta Prangel & Liina Lutsepp*
From User Feedback to Better Definitions: Applying LLMs for Pedagogical Lexicography

24. *Efi Tzelepi & Zoe Gavriilidou*
Digital Lexicography for STEM Education: Designing a Multilingual Digital Glossary of Scientific Terminology in Greek Sign Language

25. *Anna Vacalopoulou*
A Definition Walks into a Dictionary: The Lexicography of Humor

Software demonstrations

1. *Katja Brankačkec & Barbora Štěpánková*
Multifunctional Function Words in Upper Sorbian–Czech Dictionary Mudra 2.0

2. *Elena Castaldi*
Inflected-Form Search in Ancient Greek Dictionaries: A Form-to-Lemma Access Model

3. *Flávia de Oliveira Maia-Pires & Silvio Licurgo Pires*
User-Oriented Specialized Digital Lexicography and Terminological Accessibility: The LexiC Platform Applied to the Glossary of Brazilian Migration Legislation

4. *Abdelouahab Masbah, Andrea Bellandi & Silvia Piccini*
DEC Facile: Semantic-Web Authoring for Combinatorial Explanatory Dictionaries

5. *Bruno Nahod & Marijana Tuta Dujmović*
From Text to Terms: TermScout, an AI Module for Early Terminology Processing

PART II

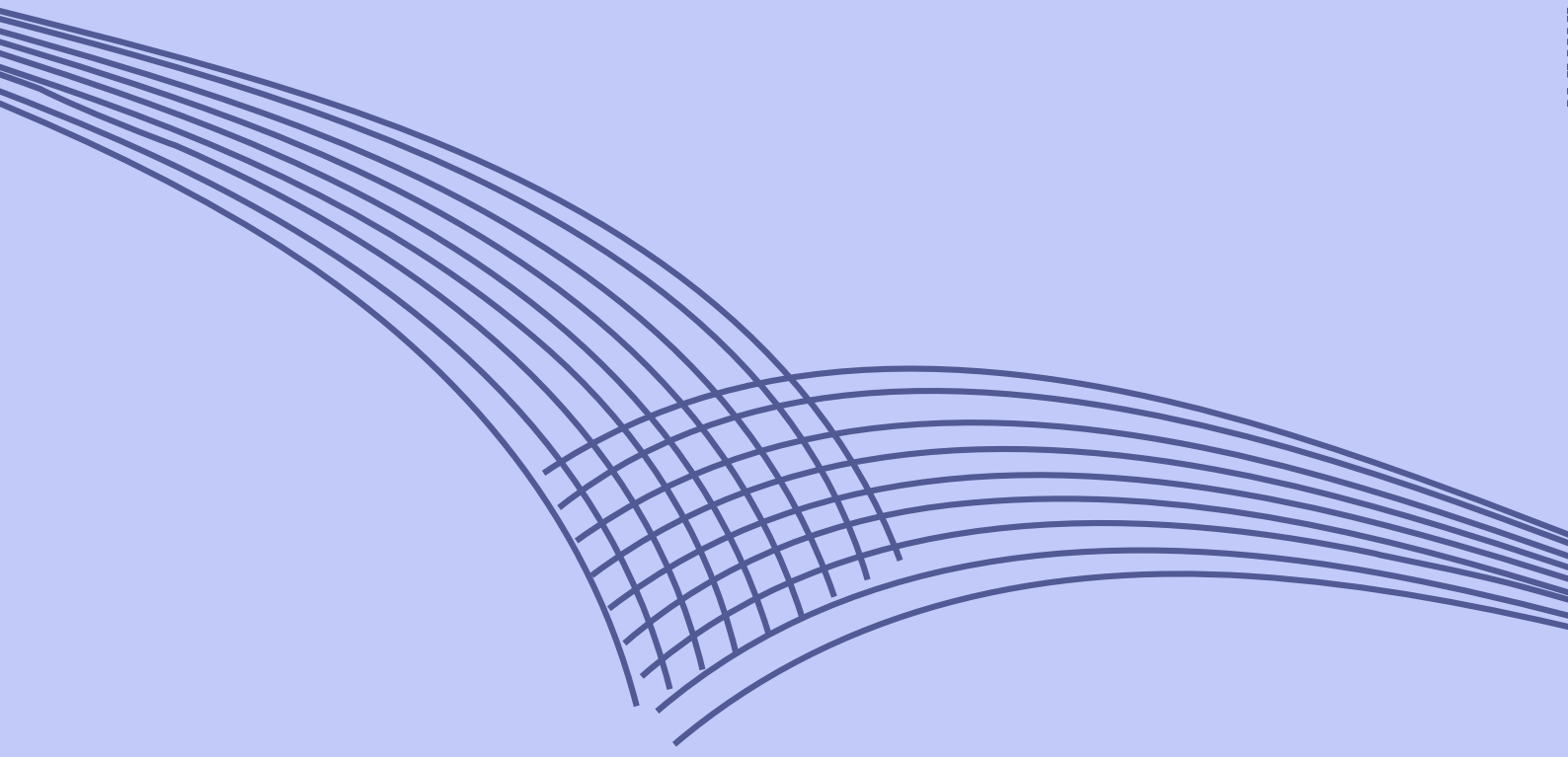
PROCEEDINGS

EURALEX 2026



COMPUTATIONAL & AI-BASED LEXICOGRAPHY

EURALEX 2026



Andrea Abel, Luca Ducceschi & Federico Villa

DETECTING SEMANTIC VARIATION BETWEEN GERMAN STANDARD VARIETIES THROUGH WORD EMBEDDING COMPARISON

Abstract From a pluricentric perspective, German possesses several standard varieties, including the one used in South Tyrol (Italy). While automated lexicography has long focussed on differences of lexical forms across these varieties, meaning differences in formally identical words are more difficult to identify systematically. This study addresses such diatopic semantic variation by comparing usage patterns in corpora from Germany and South Tyrol. Using distributional evidence from word embeddings, it shows how corpus-based comparison can support lexicographic analysis by highlighting shared words whose meanings or typical contexts diverge across varieties. The focus lies on the linguistic interpretation of these patterns and on their relevance for describing variety-specific meanings. The results confirm existing lexicographic descriptions while also identifying additional candidates for future verification.

Keywords Semantic variation; pluricentric German; lexicography; lexical semantics; South Tyrolean German

1 Introduction

From a pluricentric perspective, German possesses various standard varieties (Clyne, 1992; Ammon, 1995) documented in the "Variantenwörterbuch des Deutschen" (VWB). While the first edition (Ammon et al., 2004) relied on printed sources, the recent edition (Ammon et al., 2016) integrated corpus resources. In automated lexicographical recording of pluricentric variation, diatopic distribution of lexical forms has long been the focus (Stemle et al., 2019; Anstein, 2014). However, the automated determination of variety-specific meanings for identical lexical forms remains a significant desideratum for variant lexicography.

1.1 Research questions

The study asks how corpus-based comparison can help identify words whose meanings or typical usage differ between German as used in Germany and in South Tyrol. More specifically, it examines whether distributional NLP methods can offer insights beyond those obtainable from frequency-based statistics alone. The question is addressed from a lexicographic perspective: the aim is to assess to what extent corpus-based evidence can support the identification of region-specific senses, confirm meanings already described in lexicographic sources, and highlight further candidates for manual verification.

1.2 Related work and project context

This study continues earlier work carried out in the "STyrLogism" project (Abel & Stemle, 2018), which developed corpus-based procedures for identifying neologism candidates in the standard German variety used in South Tyrol as a contribution to language observation and lexicographic documentation. "STyrLogism" was primarily concerned with lexical innovation and with the systematic detection of items that are regionally salient, newly emerging,

or potentially relevant for dictionary updating. The present study addresses a different, though closely related, lexicographic problem: instead of focusing on new lexical forms, it examines shared lemmas whose meanings or habitual contexts become locally specialized across varieties. Its innovative contribution therefore lies in extending corpus-based comparison from the level of lexical emergence to that of semantic differentiation within formally identical vocabulary.

This perspective also connects with earlier corpus-based work on the comparison of language varieties, including preliminary contrastive studies with a focus on the standard written German of South Tyrol, carried out by Abel & Anstein (2011) within the Korpus-Südtirol-Initiative and the Vis-À-Vis prototype presented by Anstein (2012, 2014), which demonstrated how computational tools can support the contrastive exploration of variety corpora and help bring potentially relevant differences to the attention of researchers. Whereas such approaches are primarily designed to facilitate corpus comparison at a broader descriptive level, the present study applies distributional semantic modelling more specifically to the lexicographic problem of identifying region-specific senses in shared vocabulary. Its contribution therefore lies less in proposing a new technical architecture than in showing how corpus-based semantic comparison can be used as a practical aid for lexicographic analysis.

The most comprehensive compilation of the varieties of Standard German can be found in the VWB. Whereas the first edition (Ammon et al., 2004) was based on printed sources, the more recent edition (Ammon et al., 2016) incorporates corpus resources. Since we mainly focus on the standard German variety used in South Tyrol, the respective entries are of central importance for the present study. They are essentially based on Abfalterer's study (2007, pp. 167ff.), drawing on her classification of South Tyrolean peculiarities, known as South Tyroleanisms. Primary South Tyroleanisms are lemmas that are either used exclusively in South Tyrol in all their meanings, or those that have only one specific meaning (*Sonderbedeutung*) that is used exclusively in South Tyrol.

2 Corpora and data preparation

To build robust word representations for diatopic variation analysis, we compiled regional web corpora representing two varieties: standard German in Germany (DE-C) and standard Germen in South Tyrol (STYR-C). The material covers a comparable recent timeframe, extending approximately from 2012 to January 2025. The corpora were compiled concurrently in early 2025 in order to minimize the risk of confounding geographical variation with temporal semantic drift.

2.1 Corpus composition & sources selection

DE-C: Represents written standard German as used in journalistic prose in Germany. It comprises scraped articles from two widely read German newspapers: "taz" and "Merkur". The resulting DE-C corpus totals 3,718,648 sentences and 72,163,130 tokens.

STYR-C: Represents written standard German as used in journalistic prose in South Tyrol. It combines regional newspapers with community newspapers: "Die Neue Südtiroler Tageszeitung", "Salto.bz", and local municipal newsletters (*Gemeindeblätter*), scraped from South Tyrolean municipal portals. The resulting STYR-C corpus totals 5,006,396 sentences and 91,932,609 tokens.

TABLE 1 Overview of the corpora used in the study

Corpus	Subcorpus / Source	Sentences	Tokens
DE-C	taz1,837,907	—	
DE-C	Merkur	1,880,741	—
DE-C total	**Germany**	**3,718,648**	**72,163,130**
STYR-C	Tageszeitung	1,454,893	—
STYR-C	Salto	1,688,660	—
STYR-C	Gemeindeblätter	1,862,843	—
STYR-C total	**South Tyrol**	**5,006,396**	**91,932,609**

2.2 Methodological reasoning for sources selection

Journalistic comparability: By matching national newspapers ("taz", "Merkur") with regional news outlets (*Salto.bz*, "Die Neue Südtiroler Tageszeitung"), both corpora share a comparable baseline in register, style, and topical coverage (e.g. general news, politics, culture, and sports), which helps control for register-based semantic discrepancies.

Sociocultural specificity: The inclusion of the "Gemeindeblätter" subcorpus was motivated by considerations of linguistic representativeness. As community-edited local newsletters, the "Gemeindeblätter" provide access to forms of written South Tyrolean German that may remain less visible in supra-regional journalistic sources and are particularly relevant for identifying local semantic specificities. At the same time, this choice introduces an asymmetry in the corpus design, since no directly comparable subcorpus of locally edited municipal or regional publications from Germany was included in the present study. Although resources such as the regional newspapers considered in the "Variantengrammatik" project could provide a useful starting point for constructing such a comparison basis, this was beyond the scope of the current corpus compilation. The resulting asymmetry is therefore treated as a limitation and should be taken into account when interpreting the results, especially in cases where semantic divergence may reflect text type, locality, or register rather than variety-specific meaning alone. Furthermore, treating the German Standard variety in Germany as monolithic by relying solely on "taz" and "Merkur" constitutes a simplification. It does not account for distribution and lexical micro-variation within Germany itself, which could introduce confounding register or dialectal variables. A valuable future extension of this framework would be to utilize a more regionally balanced reference corpus, such as the regional sub-corpora from the German Reference Corpus (DeReKo) or the "Variantengrammatik" project. This would allow the model to control for intra-national variation and confirm whether detected shifts are specific to South Tyrol or reflect wider pluricentric dynamics.

2.3 Preprocessing pipeline

The corpora were assembled and normalized through a preprocessing pipeline designed to ensure comparability across the two varieties while preserving traceability to the original sources. Data collection was carried out using a set of custom Python scripts developed for the present study, which crawled and extracted textual material from the selected web sources and stored it in a uniform intermediate format. Since the *Gemeindeblätter* texts were available only as PDF files, part of the pipeline was specifically designed to extract their textual content

and to reduce the substantial amount of noise introduced by PDF-based conversion. This cleaning stage relied on a set of custom regular expressions and other ad hoc procedures aimed at removing recurrent artefacts and filtering out portions of text that were too fragmentary or corrupted to be linguistically informative, including broken and unusually short sentences. These steps made it possible to harmonize diverse source structures, improve the overall quality of the material, and retain source-level information for subsequent analysis.

Linguistic annotation was then performed with Stanza, a Python toolkit for multilingual natural language processing (Qi et al., 2020). In the present pipeline, the library was used for tokenization, lemmatization, named entity recognition (NER), and dependency parsing. Lemma-level normalization made it possible to group inflectional variants under a common base form and thereby to stabilize lexical comparison across varieties, while the richer annotation layers provided a structured representation of the texts for subsequent analysis. The annotated texts were subsequently saved in CoNLL format, allowing the different levels of linguistic information to be preserved in a consistent and reusable form.

3 Corpus-based comparison of meaning

To compare the two varieties (DE vs. STYR), we examined how words are represented in relation to their surrounding contexts across the two corpora. The idea is the following: if a word regularly appears with different neighbouring words in the two datasets, this may indicate a difference in meaning, usage, or contextual specialization. In lexicographic terms, the procedure serves as a filter that helps identify promising candidates for closer manual analysis.

3.1 Building a comparable basis for analysis

To model and compare lexical meaning across the two standard varieties, each corpus was represented as a semantic vector space, that is, as a system in which words are encoded as dense numerical representations derived from their distributional behaviour. Under the distributional hypothesis, words that occur in similar contexts are expected to display similar semantic profiles. This makes semantic vector spaces particularly suitable for the present study, since they allow shared lemmas from the two corpora to be compared not in isolation, but with reference to the broader contextual environments in which they occur.

For the construction of the semantic vector spaces, we used two widely adopted embedding architectures: Word2Vec (Mikolov et al., 2013) and FastText (Bojanowski et al., 2017). The two models provide complementary perspectives on lexical meaning. Word2Vec treats lexical items as atomic units and is therefore mainly sensitive to contextual co-occurrence patterns, whereas FastText incorporates subword information and can also capture formal similarities related to morphology, orthography, and compounding.

To improve the quality and comparability of the representations, the model hyperparameters, including vector size, context window, and learning rate, were optimized with Optuna (Akiba et al., 2019). Optimization was carried out against the German version of the MEN benchmark dataset, derived from the original MEN dataset (Bruni et al., 2014) and made available through the Word-Embeddings-in-the-Digital-Humanities repository described by Ehrmanntraut et al. (2021). This process selected an optimal vector size of 217 dimensions for the FastText DE model, which achieved a Spearman correlation of 0.7096, and 239 dimensions for the Word2Vec DE model, which achieved a Spearman correlation of 0.6831 with 94.59% coverage of the MEN 3000-word-pair list.

To ensure comparability between the two varieties, the STYR-C models were retrained using the same optimized hyperparameter settings as their DE-C counterparts. This produced two matched model pairs, with 217 dimensions for FastText and 239 dimensions for Word2Vec, and avoided dimensional mismatches during alignment. As an external point of comparison, we also considered embeddings derived from the German Reference Corpus DeReKo (Kupietz et al., 2010), in particular the predictive models available through the DeReKoVecs framework (Fankhauser & Kupietz, 2022). In the cased MEN setting, the DeReKo predictive baseline reaches a Spearman correlation of 0.6330. This comparison supports the view that the preprocessing and optimization pipeline yields semantically informative representations even when trained on smaller, variety-focused corpora.

The resulting models were then aligned in a shared coordinate space so that corresponding lemmas could be compared directly. A distinction must be made here between the frequency thresholds used during training and those applied during comparison. During model training, relatively low minimum-frequency thresholds were retained to maximize vocabulary coverage and preserve less frequent lexical material. For the alignment and comparison stages, by contrast, only lemmas attested with sufficient frequency in both corpora were considered, ensuring that the resulting local semantic neighbourhoods were robust enough for interpretation. Once the two spaces had been aligned, differences in local semantic neighbourhoods and semantic distances could be examined systematically and used as a basis for lexicographic analysis.

3.2 Comparing the two varieties: Alignment and anchor validation

Since the representations for Germany (DE-C) and South Tyrol (STYR-C) were trained on separate corpora, the resulting semantic vector spaces do not initially share the same coordinate system. Before lexical differences can be interpreted, the two spaces must therefore be aligned. In the present study, this was done by means of an orthogonal alignment procedure of the kind commonly used in large-scale embedding comparison (Hamilton et al., 2016), which projects the source variety space onto the target variety space. Crucially, because we retrained the STYR models with the identical dimensions obtained from the DE HPO optimization (217 for FastText and 239 for Word2Vec), the transformation matrix was computed directly without any additional numeric manipulation (e.g. padding).

Since the DE-C and STYR-C embeddings were trained independently, direct comparison requires the two spaces to be brought into a shared coordinate system. In the present study, we did not rely on a single alignment solution. Instead, we compared several alignment strategies derived from the literature and from our own refinements: a Hamilton-style global orthogonal alignment based on the shared vocabulary, a base-anchor alignment relying on manually and frequency-selected stable lemmas, and a combined-anchor alignment that extends this stable set through an empirically derived stability threshold. We adopted this multi-method design because each alignment strategy offers a different compromise between robustness and sensitivity. Global alignment provides broad comparability, but it may partially smooth precisely those local semantic differences that are lexicographically relevant. Anchor-based approaches reduce this risk, but they introduce a stronger dependence on the choice of reference items. Rather than treating one method as universally superior, we use the different alignments as complementary perspectives on diatopic semantic variation.

To establish the anchor-based comparison space, the anchor set was constructed through a hybrid manual and statistical procedure. It began with a manually selected seed list of 73 lemmas, including grammatical function words, basic vocabulary, and culturally stable journalistic

terms. This initial set was then expanded by intersecting the 5,000 most frequent lemmas in both corpora and retaining only those whose relative frequency ranks differed by no more than 500 positions. The underlying assumption was that strongly correlated rank profiles across the two corpora provide a useful proxy for semantic stability. The union of the manually selected seeds and the rank-filtered frequent lemmas yielded a base anchor set of 1,914 items. In a further step, this set was extended through an empirically derived stability threshold based on local-neighbourhood overlap, yielding a combined-anchor configuration that integrates theoretically motivated and statistically promoted stable items. The robustness of these anchor-based solutions was evaluated through 5-fold cross-validation, which showed that the anchor sets provided a stable basis for alignment across different splits of the data. For the purposes of the present paper, the key methodological point is therefore not that one anchor configuration replaces all others, but that the comparison across global, base-anchor, and combined-anchor alignments allows the analysis to balance broad geometric comparability with sensitivity to local semantic differentiation.

3.3 From quantitative signals to lexicographic interpretation

Once the spaces had been aligned, semantic divergence was assessed through complementary measures drawn from lexical semantic change and corpus comparison (Schlechtweg et al., 2019). Cosine distance captures broader positional differences between corresponding lemmas, while local divergence was measured through semantic neighbourhood overlap, operationalized as Jaccard distance (Jaccard, 1901). This measure compares the top nearest neighbours of a lemma in the two varieties and is related to the neighbour-overlap approach of Gonen et al. (2020). Here, we use a normalized distance formulation: higher Jaccard values indicate greater divergence, while lower values indicate greater stability. Candidate extraction was therefore based not on a single ranked list, but on convergence and divergence patterns across several comparison procedures.

$$\text{Jaccard distance} = 1 - |A \cap B| / |A \cup B|$$

For instance, the lemma *Konsortium* ‘consortium’ illustrates the distinction between these distance measures. In Germany, the word is embedded in financial contexts, yielding nearest neighbors like *Investor* ‘investor’ and *Konzern* ‘group of companies’. In South Tyrol, it is specialized in regional administrative and agricultural settings, showing neighbors like *Gemeindenverband* ‘association of local authorities’ and *Apfelkonsortium* ‘apple consortium’. A local comparison measure like Jaccard distance is highly sensitive to this complete local neighborhood restructuring, yielding a distance of 0.9417 (representing nearly disjoint top neighborhood sets), whereas global positional rotations might partially obscure such localized shifts by smoothing them over the entire vocabulary space. Here, A and B denote the sets of the k nearest neighbours of the target lemma in the DE and STYR spaces. Because the measure is based on neighbour sets rather than absolute vector positions, it helps detect localized usage shifts that global positional differences may miss. Values close to 1 indicate almost disjoint semantic neighbourhoods. Lexicographically, such cases are informative but ambiguous: they may reflect variety-specific semantic specialization, register- or genre-driven separation, or referential effects linked to proper names and local institutions.

For each embedding architecture, we retained the top-ranked candidates from Hamilton cosine distance, base-anchor cosine distance, combined-anchor cosine distance, and neighbourhood overlap divergence. This multi-index design maximized coverage: some procedures capture broad geometric displacement, others localized contextual restructuring. Each method may also introduce distortions, through over-normalization, anchor selection, or sparse neighbourhoods. For lexicographic purposes, this is useful: recurring candidates provide stronger

evidence of robust variety-related differentiation, while method-specific candidates may reveal more marginal, formally driven, or context-specific phenomena.

The resulting rankings were treated not as lexicographic conclusions, but as a heuristic filter for closer qualitative inspection. Strongly divergent words were examined through their local semantic neighbourhoods and, where necessary, corpus concordances. Quantitative comparison thus provided a structured entry point for interpretation without replacing manual analysis and contextual verification. In particular, neighbourhood overlap offers an interpretable indication of whether a lemma occurs in substantially different local semantic environments across varieties, in line with corpus-driven approaches in which contextual profiles and collocational patterns support lexicographic description (Kilgarriff & Tugwell, 2002).

4 Results and discussion

The comparison of the shared vocabulary revealed a set of words whose usage differs markedly between the two varieties. These differences cluster above all in domains that are especially salient in South Tyrol, such as administration, politics, local institutions, and everyday cultural life. From a lexicographic perspective, this is significant because it shows that semantic variation is not random: it tends to emerge where regional experience, institutional structures, and recurring discourse traditions create stable local usage patterns.

4.1 Lexicographic relevance of the findings

The main value of the analysis lies in its ability to support lexicographic decision-making. It should help identify words that deserve closer inspection, confirm many already described variety-specific meanings, and point to further candidates that may not yet be fully documented. At the same time, some high-ranking items reflect named entities or locally referential uses whose status must be assessed through manual lexicographic interpretation. In this sense, corpus-based comparison does not replace lexicographic interpretation; rather, it makes the search for relevant evidence more systematic and scalable.

4.2 Baseline comparison: Frequency keyness vs. aligned embeddings

To evaluate how a frequency-only method compares to our aligned semantic embeddings, we ran a baseline keyness analysis using Log-Likelihood (LL) statistics, following established corpus-comparison practice in frequency profiling (Rayson & Garside, 2000). However, cross-referencing the Log-Likelihood rankings with the VWB provides only a partial benchmark, since STYR-labelled VWB items primarily mark forms or meanings treated as specific to South Tyrol and do not systematically capture features that South Tyrolean German shares with one or more other national or regional varieties. Against this restricted reference set, the Top 1,000 LL keywords recovered 28 out of 389 STYR-labelled VWB lemmas (7.2%), and the Top 5,000 recovered 86 (22.1%). These figures should therefore not be interpreted as a complete measure of coverage of South Tyrolean lexical variation. Rather, they show that frequency-based profiling has limited access to VWB-encoded South Tyrolean specificities, especially where regional relevance depends on meaning, usage, or cross-varietal overlap rather than on corpus exclusivity alone.

The critique of the frequency approach highlights two primary issues. First, keyness statistics are highly sensitive to register and genre differences; our LL rankings were dominated by grammatical noise (‘und’ *and*, ‘der’ *the*, ‘werden’ *become*). Second, and more critically,

frequency-based methods are entirely blind to semantic shifts where a word is highly frequent in both varieties but has changed meaning. For instance, 'Bar' *bar* (frequent in both varieties) has a low LL score, but semantic neighbourhood overlap, measured as Jaccard distance, reveals substantial local divergence. The comparison was further strengthened by using two embedding architectures with different inductive biases. FastText, with its subword-based representations, is more sensitive to morphological structure, orthographic variation, and compounding patterns, all of which are potentially relevant in German. Word2Vec, by contrast, treats lexical items as atomic units and therefore provides a more context-driven view of semantic specialization that is less influenced by formal resemblance. The use of both architectures broadens the scope of candidate discovery: convergence across them increases confidence in the robustness of the signal, whereas divergence between them can itself be informative for distinguishing semantically motivated shifts from more form-driven or morphologically patterned variation.

This limitation must also be understood in relation to the structure of the aligned comparison space itself. To be compared across varieties, lemmas had to be attested with sufficient frequency in both corpora. Many of the items documented in the VWB, however, are form regionalisms that are characteristic of only one variety and therefore fall outside the shared high-frequency vocabulary used for alignment. The relatively small overlap between the aligned candidate list and the VWB is therefore not a sign of methodological failure but partly follows from the fact that the present approach is designed primarily to detect semantic specialization in lemmas shared by both corpora. At the same time, the VWB does not provide an exhaustive inventory of all lexical or semantic features relevant to South Tyrolean German, especially where these are shared with other national or regional varieties. The different sources of evidence should therefore be understood as complementary: frequency and exclusivity measures are useful for detecting formally or distributionally distinctive items, while aligned embeddings help identify semantic and contextual differentiation within shared vocabulary.

4.3 Methodological complementarity in candidate detection

One important outcome of the analysis is that no single comparison metric exhausts the range of diatopic variation relevant for lexicography. The different procedures used in this study—Hamilton cosine, base-anchor cosine, combined-anchor cosine, and neighbourhood overlap divergence—do not simply produce redundant rankings. Rather, they highlight partly overlapping but non-identical candidate sets, which suggests that each method is sensitive to different structural properties of the semantic spaces.

This is supported by our empirical analysis of the candidate space. The four methods do not yield identical top rankings. Specifically, when analyzing the pairwise overlaps of the top 200 candidates (an amount that can be increased to any number that the lexicographer desires) generated by each method under the FastText model, we find that the global rotation method (Hamilton) and the local neighborhood overlap method (Jaccard) share only 60 words (a 30% overlap). Similarly, Base Anchor alignment and Jaccard share only 56 words, and Combined Anchor alignment and Jaccard share 60 words. In contrast, the global rotation methods show high alignment among themselves: Hamilton and Base Anchor share 150 words (75% overlap), while Hamilton and Combined Anchor share 176 words (88% overlap). For the Word2Vec model, the divergence is even more pronounced: Hamilton and Jaccard share only 29 words (14.5% overlap), while Base Anchor and Jaccard share 27 words. This low overlap demonstrates that neighborhood-overlap measures capture localized contextual reorganization that global positional rotations miss, whereas rotation-based measures are sensitive to broader coordinate system shifts.

From a methodological point of view, this is expected. Global alignment methods can highlight broad positional displacement but may also reduce local differences by imposing a uniform rotation across the two spaces. Anchor-based methods preserve semantically stable reference points more selectively and are therefore better suited to detecting divergence without excessively flattening contrasts between varieties. Neighbourhood-overlap measures, by contrast, rely less on global vector position than on differences in a lemma's local semantic neighbourhood across corpora. They are therefore especially informative for detecting contextual reorganization even when overall geometric displacement is limited. For lexicographic analysis, this plurality is useful. Words that rank highly across several methods are strong candidates, since their divergence is not an artefact of one specific comparison procedure. At the same time, candidates that emerge only in one ranking are not necessarily noise: they may reflect register specialization, referential concentration, formal regionalism, or morphologically driven analogies. The methodological goal was therefore not to identify a single "best" score, but to construct a richer empirical search space in which different indices jointly support qualitative interpretation.

4.4 From quantitative consensus to qualitative interpretation

To make the quantitative results usable for lexicographic evaluation, the candidate lists were consolidated across methods into a shared inspection space. For each model architecture, candidates appearing among the top-ranked items of at least one comparison procedure were merged into a consolidated list, enriched with contextual neighbour information and ordered by consensus strength. In this way, the quantitative stage does not aim to replace qualitative judgement, but to structure it. High-consensus items provide stronger candidate signals because they recur across several comparison procedures, but they are not automatically lexicographically relevant; their status still depends on qualitative interpretation of neighbourhoods and concordances. Conversely, lower-consensus or method-specific items may reveal more localized phenomena, including edge cases, referential effects, register-specific patterns, or false positives.

The consolidated comparison space covered 13,454 shared lemmas in both models. FastText produced a smaller candidate set than Word2Vec (356 vs. 474 lemmas) and a substantially higher proportion of high-consensus items: 157 FastText candidates were supported by three or four methods, compared with 98 Word2Vec candidates. Conversely, method-specific candidates were more frequent in Word2Vec (291; 61.4%) than in FastText (168; 47.2%), suggesting that FastText yielded a more compact and internally convergent signal, whereas Word2Vec produced a broader but more method-dependent inspection space.

This organization also reflects the interpretive logic of the study. The quantitative procedures identify loci of possible divergence, but only qualitative inspection can determine whether a candidate corresponds to a region-specific sense, a form-based regional pattern, a proper-name effect, or a statistical artefact. In addition, we used examples identified in Abel & Anstein (2011) as additional guidance when selecting items for an illustrative qualitative analysis.

The examples below were selected to illustrate different interpretive outcomes within the candidate space: confirmation of VWB entries, cases in which corpus evidence points to special meanings already described lexicographically, candidates not yet recorded in the VWB, proper-name effects, and morphologically induced false positives. In the discussion, FT and W2V refer to FastText and Word2Vec, respectively.

Abgeordnetenkammer 'Chamber of Deputies' is a lemma used to refer to the Chamber of Deputies of the Italian Parliament and as such a primary South Tyroleanism in the VWB without any other uses in the German-speaking area. The identified neighbours in DE-C unexpectedly overlap with those of STYR-C, for example *Parlamentskammer* 'chamber of parliament' in both FT and W2V. However, closer inspection of the DE-C concordances shows that these matches are mainly attributable to reports on Italian politics.

Further examples concern primary South Tyroleanisms with special meanings. For example, *innerhalb* 'within' in DE-C is associated with temporal durations, such as *binnen* 'within' and *Gesamtdauer* 'total duration', while in STYR-C references to points in time and deadlines predominate, such as *spätestens* 'at the latest' and Frist 'deadline', consistently across FT and W2V and in line with the VWB. Similarly, for *Konsortium* 'consortium', DE-C identifies banking and economic contexts, such as *Konzern* 'group of companies', *Rüstungskonzern* 'defence industry group', and *Investor* 'investor', whereas STYR-C results highlight agricultural and administrative contexts, such as *Apfelkonsortium* 'apple consortium' and *Gemeindenverband* 'association of local authorities', both in FT and W2V. This also aligns with Abel & Anstein (2011), who identified additional South Tyrolean nuances of meaning beyond the VWB entry. *Autonom* 'autonomous' is another lemma recorded in the VWB with South Tyrolean special meanings, particularly in the context of housing and in relation to South Tyrol's legally enshrined powers of self-determination within the Italian state, alongside the more general meaning 'administratively independent' shared across the German-speaking area. In DE-C, the nearest neighbours predominantly refer to independently moving objects, such as *Drohne* 'drone', *ferngesteuert* 'remote-controlled', and *fahrend* 'moving' in W2V, and partly also to military contexts, such as *entmilitarisiert* 'demilitarised' and *Revolutionsarmee* 'revolutionary army' in FT. In STYR-C, by contrast, the FT and W2V neighbours are strongly connected to South Tyrol's legal status, including *Sonderstatut* 'special statute', *Provinz* 'province', and *Zuständigkeit* 'competence'.

Other cases, such as *Bar* 'bar', appeared relatively similar across both varieties in parts of their semantic neighbourhoods, sharing items such as *Restaurant* 'restaurant', *Lokal* 'restaurant, pub', and *Hotel* 'hotel' among the top nearest neighbours. However, the VWB lists *Café* 'café' as a South Tyrolean special meaning, while other meanings such as *Nachtclub* 'nightclub' and *Theke* 'bar counter' are shared across the German-speaking area. This case therefore illustrates a broader limitation of lemma-level embedding comparison: since the models represent each lemma through an aggregated distributional profile, they cannot directly separate variety-specific senses from shared senses within the same form. Polysemous items that combine an exclusive regional meaning with more widespread meanings therefore require concordance-based verification and, ideally, sense-level modelling.

The analysis also brings into view candidates whose regional semantic profiles have not yet been systematically recorded. One example is *wandern* 'to hike', which is not included in the VWB. "Duden online" distinguishes several uses of the verb, including both hiking in nature and more general forms of walking. In our data, these uses appear to be distributed differently across the two corpora. In DE-C, the neighbourhood of "wandern" points more clearly to leisurely walking, with neighbours such as *Wattwanderung* 'mudflat walk' and *entlangspazieren* 'stroll along'. In STYR-C, by contrast, the word is more strongly associated with nature hiking, sporting activity, and mountainous terrain, as shown consistently across FT and W2V by neighbours such as *Hangquerung* 'slope crossing', *hinauf* 'upwards', and *Waldsteig* 'forest trail'. Such cases are promising for lexicographic follow-up, but they also require further investigation to distinguish genuine semantic variation from geographical, topical, or journalistic biases.

A different pattern is illustrated by *Dekadenz*. In DE-C, the lemma primarily evokes the general abstract meaning of 'decadence'. In STYR-C, however, its semantic neighbourhood is strongly shaped by the cabaret venue *Die Dekadenz* in Brixen/Bressanone, with neighbours such as *Kabarettabend* 'cabaret evening', *Theaterpremiere* 'theatre premiere', and *Carambolage*, the name of another South Tyrolean cabaret venue, in both FT and W2V. The example shows how proper-name usage and local cultural salience can affect the distributional profile of an otherwise common lemma, and why such cases require qualitative interpretation before they can be treated as lexicographically relevant.

The item *Sprengel* shows clearly distinct neighbour sets, with a Jaccard distance of 1 for both FT and W2V, but its interpretation requires caution. The lemma is also recorded in the VWB and refers to the remit of an office or institution in Austria and South Tyrol, whereas in Germany it is used only in ecclesiastical contexts and is otherwise obsolete. However, FT mainly lists items connected to *sprengen* 'to blow up', which share surface features with Sprengel but have completely different meanings.

5 Limitations and future work

The present approach has two main limitations. First, it operates at lemma level and therefore compares aggregated distributional profiles across corpora. This is useful for detecting candidates whose contextual behaviour differs across varieties, but it cannot directly separate South Tyrolean-specific meanings from meanings shared across the German-speaking area. Polysemous items may therefore show mixed neighbourhoods, and relevant regional specializations can be diluted by more frequent shared uses.

Second, evaluation against the *VWB* should be understood as conservative. STYR-labelled entries provide an important reference point, but they do not exhaustively cover all lexical or semantic features relevant to South Tyrolean German, especially when these are shared with other national or regional varieties. Future work should therefore combine lemma-level candidate detection with sense-level validation, for example through concordance analysis, clustering of contextualized occurrences, or contextual embedding models. Extending the comparison to Austrian and Swiss German corpora would also help determine whether a feature is specific to South Tyrol or shared across several standard varieties.

6 Conclusion

This paper has examined how corpus-based semantic comparison can support the lexicographic analysis of diatopic variation in shared German vocabulary. Focusing on German as used in Germany and in South Tyrol, we showed that aligned word embeddings can help identify lemmas whose contextual profiles differ across varieties and can therefore serve as candidates for closer lexicographic inspection. The results confirm the usefulness of this approach as a heuristic rather than as an automatic decision procedure. Several high-ranking candidates correspond to meanings already described in the VWB, while others point to potentially relevant regional usages that require further verification. At the same time, cases such as proper-name effects, register-driven patterns, and polysemous lemmas show that quantitative divergence must always be interpreted against neighbourhoods, concordances, and existing lexicographic evidence. Overall, the study suggests that semantic variation in pluricentric German is best approached through the convergence of complementary corpus-based signals. Frequency-based comparison remains useful for detecting formally or distributionally distinctive items, whereas aligned embeddings are particularly valuable for identifying semantic and contextual differentiation within shared lemmas. Used in combination, these methods provide

a scalable way of structuring candidate discovery while preserving the central role of expert lexicographic judgement.

References

Abel, A. & Anstein, S. (2011). Korpus Südtirol – Varietätenlinguistische Untersuchungen. In A. Abel & R. Zanin (Eds.), *Korpora in Lehre und Forschung* (pp. 29–54). Bozen-Bolzano University Press.

Abel, A. & Stemle, E. W. (2018). On the detection of neologism candidates as basis for language observation and lexicographic endeavours: The STyrLogism project. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (Eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts* (pp. 535–544).

Abfalterer, H. (2007). *Der Südtiroler Sonderwortschatz aus plurizentrischer Sicht*. Innsbruck University Press.

Akiba, T., Sano, S., Yanase, T., Ohta, T. and Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2623–2631).

Ammon, U. (1995). *Die deutsche Sprache in Deutschland, Österreich und der Schweiz: Das Problem der nationalen Varietäten*. De Gruyter.

Ammon, U., Bickel, H., Ebner, J., Esterhammer, R., Gasser, M., Hofer, L., Kellermeier-Rehbein, B., Löffler, H., Mangott, D., Moser, H., Schläpfer, R., Schloßmacher, M., Schmidlin, R. & Vallaster, G. (2004). *Variantenwörterbuch des Deutschen: Die Standardsprache in Österreich, der Schweiz und Deutschland sowie in Liechtenstein, Luxemburg, Ostbelgien und Südtirol*. De Gruyter.

Ammon, U., Bickel, H. & Lenz, A. N. (Eds.) (2016). *Variantenwörterbuch des Deutschen: Die Standardsprache in Österreich, der Schweiz, Deutschland, Liechtenstein, Luxemburg, Ostbelgien und Südtirol sowie Rumänien, Namibia und Mennonitensiedlungen*. 2nd ed. De Gruyter.

Anstein, S. (2012). Comparing variety corpora with Vis-À-Vis – a prototype system presentation. In *Proceedings of KONVENS 2012* (pp. 243–247).

Anstein, S. (2014). *Computational approaches to the comparison of regional variety corpora* [PhD thesis, Stuttgart].

Bojanowski, P., Grave, E., Joulin, A. & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5, 135–146. https://doi.org/10.1162/tacl_a_00051

Bruni, E., Tran, N. K. and Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, 49, 1–47. https://doi.org/10.1613/jair.4135

Clyne, M. (1992). *Pluricentric languages: Differing norms in different nations*. De Gruyter.

Ehrmanntraut, A., Hagen, T., Konle, L. & Jannidis, F. (2021). Type- and token-based word embeddings in the digital humanities. In M. Ehrmann, F. Karsdorp, M. Wevers, T. L. Andrews, M. Burghardt, M. Kestemont, E. Manjavacas, M. Piotrowski & J. van Zundert (Eds.), *Proceedings of the Second Conference on Computational Humanities Research (CHR 2021), Amsterdam* (pp. 16–38).

Fankhauser, P. & Kupietz, M. (2022). Count-based and predictive language models for exploring DeReKo. In P. Bański, A. Barbaresi, S. Clematide, M. Kupietz & H. Lüngen (Eds.), *Proceedings of the 10th Workshop on Challenges in the Management of Large Corpora (CMLC-10 2022)* (pp. 27–31).

Gonen, H., Jawahar, G., Seddah, D. & Goldberg, Y. (2020). Simple, interpretable and stable method for detecting words with usage change across corpora. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 538–555).

Hamilton, W. L., Leskovec, J. & Jurafsky, D. (2016). Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics* (pp. 1489–1501).

Jaccard, P. (1901). Étude comparative de la distribution florale dans une portion des Alpes et des Jura. *Bulletin de la Société Vaudoise des Sciences Naturelles*, 37, 547–579. https://doi.org/10.5169/seals-266450

Kilgarriff, A. & Tugwell, D. (2002). Sketching words. In M.-H. Corréard (Ed.), *Lexicography and Natural Language Processing: A Festschrift in Honour of B.T.S. Atkins* (pp. 125–137).

Kupietz, M., Belica, C., Keibel, H. & Witt, A. (2010). The German Reference Corpus DeReKo: A primordial sample for linguistic research. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner & D. Tapias (Eds.), *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)* (pp. 1848–1854).

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv preprint. arXiv:1301.3781

Qi, P., Zhang, Y., Zhang, Y., Bolton, J. & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 101–108).

Rayson, P. & Garside, R. (2000). Comparing corpora using frequency profiling. In *Proceedings of the Workshop on Comparing Corpora* (pp. 1–6).

Schlechtweg, D., Hätty, A., Del Tredici, M. & Schulte im Walde, S. (2019). A wind of change: Detecting and evaluating lexical semantic change across times and domains. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 732–746).

Stemle, E. W., Abel, A. & Lyding, V. (2019). Language varieties meet one-click dictionary. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (Eds.), *Electronic lexicography in the 21st century: Proceedings of the eLex 2019 conference, 1–3 October 2019, Sintra, Portugal* (pp. 537–546). Lexical Computing CZ.

Contact information

Andrea Abel
Free University of Bolzano/Bozen & Institute for Applied Linguistics, Eurac Research, Bozen/Bolzano, Italy
andrea.abel@unibz.it, andrea.abel@eurac.edu

Luca Ducceschi
Institute for Applied Linguistics, Eurac Research, Bozen/Bolzano, Italy
luca.ducceschi@eurac.edu

Federico Villa
Independent Researcher
federico.villa03@icatt.it

Iván Arias-Arias, María José Domínguez Vázquez,
María Teresa Sanmarco Bande & Carlos Valcárcel Riveiro

FROM CONTENT TO CODE AND BACK

Evaluating Agentic Coding in LLMs for Building R Shiny Dictionary Interfaces

Abstract While recent scholarship highlights the generative capabilities of LLMs for drafting lexicographic entries, their "black box" nature poses challenges regarding training corpora, linguistic representativity, and semantic hallucination. We propose that this opacity is less problematic when LLMs are deployed not as linguistic experts, but as software engineers. By leveraging LLMs as agentic coding tools, lexicographers can bridge the gap between linguistic data and digital publication, democratizing access to custom lexicographic tools. This study investigates how LLMs conceptualize "user-friendliness" when tasked with designing lexicographic interfaces. We evaluate three leading models in agentic coding – Claude Opus 4.6, ChatGPT-5.5, and Gemini 3.1 Pro – using a zero-shot prompting strategy within a case study on noun valency. Each model generates an R Shiny application from curated CSV data, accompanied by explicit justifications for its design choices. The generated interfaces are assessed against criteria fundamental to e-lexicography: access structures, visual design, information architecture, and alignment with user-centred principles for language learners. Results reveal substantial differences across models in how they conceptualize and operationalize usability, with each model emphasizing fundamentally different design aspects. Crucially, not all models prove equally reliable as agentic coders for generating lexicographic interfaces that meet professional standards, highlighting the need for critical evaluation before adoption.

Keywords LLMs; agentic coding; e-lexicography; R Shiny; user-friendliness

1 Introduction

The integration of large language models (LLMs) into lexicographic practice has generated considerable scholarly attention in recent years. Much of this discourse has focused on the generative capabilities of these models – their potential to draft dictionary entries, produce definitions, or suggest usage examples (de Schryver & Joffe, 2023; Fuertes-Olivera, 2024). Yet reliance on LLMs for lexical content generation raises well-documented concerns. As these models operate as "black boxes", lexicographers face fundamental questions regarding training data composition, linguistic representativity, bias, and the persistent risk of hallucination (Rundell, 2023). While human expertise remains indispensable for content revision – particularly for under-resourced languages and varieties (Lew, 2024; Stöckle et al., 2025) – these limitations have arguably overshadowed a different, potentially more productive avenue for LLM adoption in the field.

We argue that there is an under-explored domain where the opacity of LLMs is substantially less problematic: the automation of software infrastructure for lexicographic tools. When deployed not as linguistic experts but as software engineers, LLMs are tasked with generating code whose correctness can be verified, tested, and iteratively refined. This starting point represents a different epistemic situation from content generation, where linguistic errors may go undetected without specialist knowledge. It should be stressed, however, that this advantage is only partial: security-relevant flaws or logic errors in generated code can likewise remain invisible to a non-specialist reader, even when the application appears to function as intended. What code offers is a path to systematic testing and debugging that is less readily available for purely linguistic content.

By leveraging LLMs as agentic coding tools – that is, LLMs operating within a natural-language-driven development loop in which they autonomously plan, generate, execute, and iteratively refine functional software artifacts – lexicographers can bridge the gap between curated linguistic data and its digital publication, thereby reducing dependence on external IT support and democratizing access to custom end-user lexicographic tools.[1] This shift in perspective – from LLMs as content producers to LLMs as development tools – aligns with the broader phenomenon informally termed "vibe coding" (Karpathy, 2025), a label that, despite its at times disparaging connotations, has come to designate the emerging practice of software creation through natural language interaction with LLMs rather than manual code writing.[2] In the context of lexicography, this means that an expert with curated data and a clear understanding of user needs can prompt an LLM to generate a functional dictionary application.

The potential of this approach is illustrated by recent developments in digital lexicography, where R Shiny has emerged as a framework that enables dictionary curators to build interactive applications using only R (Lugli, 2024). Lugli (2024) demonstrates how Shiny-based prototyping enabled lexicographers to build dictionary interfaces directly from annotated data, though still requiring them to write R code.[3] The present study extends this logic by asking a question that has received little systematic attention: whether interfaces produced through natural language prompting meet the standards of user-friendliness established by e-lexicographic research, or default instead to generic software design conventions.

We address this question through a comparative evaluation of three leading LLMs in agentic coding – Claude Opus 4.6, ChatGPT-5.5, and Gemini 3.1 Pro – using a case study on German noun valency, following the selection procedure described in Section 3.2. Each model is prompted to generate an R Shiny application from a curated CSV dataset describing the argument structures of German nouns, accompanied by explicit instructions targeting language learners as end users. The generated interfaces are then assessed against a purpose-built evaluation framework grounded in e-lexicographic research on user-centred design.

Our central research question is: Can LLMs reliably generate the end-user interface layer of lexicographic software – i.e., the graphical interface through which language learners access dictionary content – that meets professional standards for access structures and usability, provided that linguistic accuracy is assured by linguist-curated data? To this end, each model is additionally asked to justify its design choices, identifying the aspects of user-friendliness it considers most important. This enables a systematic comparison between what the models say about their own designs and the evaluators' independent, empirical assessments of those designs.

[1] We use "lexicographic software" throughout this paper in this narrow, end-user-facing sense. The broader ecosystem of software used to produce, curate, and manage lexicographic data – e.g., corpus-annotation tools, database back-ends – is not the object of the present study, even though it might also reasonably be called "lexicographic software" in a wider sense.

[2] Recent large-scale studies report that over 25% of active GitHub projects already show traces of coding agent adoption (Robbes et al., 2026), and agentic coding frameworks are increasingly recognized as a distinct paradigm in software engineering research (Rasheed et al., 2025; Dong et al., 2026). Empirical observations further confirm that LLMs can expedite the prototyping of intuitive interfaces and aesthetically coherent designs, even when the resulting code requires subsequent optimization by domain experts (Smutny & Bojko, 2024).

[3] It should be noted, however, that even when agentic coding removes the need to write R code by hand, transferring a prototype interface to a production-grade, maintainable dictionary application remains a substantial challenge in itself; the present study accordingly treats the three LLM-generated applications as prototypes, not as production-ready dictionaries, and the discussion below should be read with that distinction in mind.

The article is organized as follows: Section 2 establishes the theoretical framework; Section 3 describes the methodology; Section 4 presents the results; and Sections 5 and 6 discuss implications and conclusions.

2 User-friendliness and LLMs

User-friendliness constitutes a central principle in user-oriented lexicography (Huete-García, 2023), encompassing the accessibility of information, the clarity of metalanguage, the functional relevance of data, and cognitive economy. Studies consistently indicate that typical lexicographic consultations tend to be brief, focused, and oriented towards the immediate resolution of specific problems, translating into a clear preference for concise definitions, transparent access structures, and content directly aligned with the task at hand (Kemmer, 2010; Lew, 2011; Müller-Spitzer et al., 2018; Wolfer et al., 2021; Lew & Wolfer, 2024). Conversely, information overload, lack of data hierarchization, and the use of unexplained metalanguage increase cognitive load and generate friction in the interaction, reducing the effectiveness of digital lexicographic resources (Müller-Spitzer, 2022; Bothma & Fourie, 2024). Research in e-lexicography therefore underscores the need for an information architecture oriented towards the context of use, capable of filtering, prioritizing, and adapting data according to the situated needs of the user, rather than offering undifferentiated exhaustive repertoires (Klosa-Kückelhaus & Michaelis, 2024; Bothma & Fourie, 2024). Taken together, empirical evaluation of electronic dictionaries converges in showing that users value these resources primarily in terms of clarity, cognitive economy, functional adequacy, and perceived reliability, prioritizing the speed with which they can identify and apply the information needed to resolve a specific linguistic problem (Lew, 2011; Wolfer et al., 2021).

In this landscape, LLMs have increasingly been adopted as general-purpose linguistic consultation tools for resolving queries related to meaning, usage, and text formulation – a trend that is well documented qualitatively, though not yet at the scale of large empirical usage studies (cf. Arias-Arias, Domínguez-Vázquez & Valcárcel-Riveiro, 2024, on ChatGPT's capacity to emulate lexicographic text). From a functional perspective, these systems can be understood as linguistic problem-solving devices comparable to electronic dictionaries, albeit operating in a conversational and generative environment that lacks a stable lexicographic structure. Human-centred evaluation studies have shown that their value depends not solely on the correctness of the generated content, but on factors related to usability, cognitive load, transparency, and perceived satisfaction (Alabduljabbar, 2024; Blodgett et al., 2024; Kim et al., 2024; Lew, 2024). Nevertheless, response variability, opacity of data sources, and difficulty in evaluating reliability introduce specific challenges, highlighting the need to integrate dimensions such as transparency, perceived reliability, and system predictability into their evaluation (Liao & Vaughan, 2024; Xu et al., 2024). Consequently, user-friendliness in the context of LLMs can be defined as the system's capacity to generate responses that enable users to effectively resolve their linguistic needs through clear, functionally relevant, and cognitively accessible information, within a conversational environment that lacks a stable lexicographic structure.

The considerations outlined above apply primarily to scenarios in which LLMs function as content generators – producing definitions, examples, or explanations in direct interaction with end users. It is worth distinguishing here between two possible levels of LLM involvement in lexicographic content that are sometimes conflated: (i) the LLM as the dictionary itself, consulted directly by the end user – the scenario just described; and (ii) the LLM as a tool used by the lexicographer to draft or generate dictionary content prior to publication, a distinct use case (addressed by, e.g., de Schryver & Joffe, 2023; Fuertes-Olivera, 2024) that also involves LLMs as content creators, but not in direct interaction with the end user. The present study investigates a third, largely unexplored role that cuts across this distinction: the LLM as a builder of the

interface through which lexicographic content – however it was produced, and by whichever of the two levels above – is ultimately delivered to the user. When an LLM is tasked with generating a lexicographic application in this third role, the relevant dimensions of user-friendliness shift from the quality of linguistic content – which, in our design, is assured by the curated input data – to the quality of the interface through which that content is presented. The question is no longer whether the LLM can produce an accurate definition, but whether it can design an access structure that enables a language learner to efficiently locate, interpret, and apply the information contained in a linguist-curated dataset. In this scenario, the LLM must operationalize principles of e-lexicographic design – information architecture, visual hierarchy, cognitive load management, access structure transparency – without any confirmed exposure to lexicographic theory.[4]

This creates a productive tension. On the one hand, LLMs bring to the task a broad familiarity with interface design patterns, responsive layouts, and interaction conventions from mainstream web development. On the other hand, lexicographic interfaces have domain-specific requirements that generic UX principles do not necessarily address in full: the need for multiple, complementary access routes to the same data (Müller-Spitzer & Koplenig, 2014); the presentation of complex relational information such as argument structures, collocational patterns, or semantic networks; and the adaptation of information density to specific user profiles and consultation scenarios (Fuertes-Olivera, 2025; Klosa-Kückelhaus & Michaelis, 2024).[5] The evaluation framework developed for this study (Section 3) is therefore designed to capture both dimensions: the generic usability of the generated interfaces and their alignment with domain-specific lexicographic principles. While several evaluation dimensions – such as cognitive load management or task alignment – are not exclusive to lexicography, their operationalization in this study is anchored in lexicographic theory, where key concepts such as metalanguage accessibility, explicitness of argument structure, or corpus-based example presentation acquire domain-specific meanings related to dictionary use (see Table 1). The contribution is therefore not generic UI design, but the operationalization of lexicographic principles in AI-generated interfaces.

3 Methodology

3.1 Data preparation

The linguistic data used in this study consists of a curated CSV dataset describing the valency behaviour of three German nouns – *Diskussion* 'discussion', *Frage* 'question', and *Flucht* 'escape' – drawn from the Portlex dictionary (Domínguez Vázquez & Valcárcel Riveiro 2020). The dataset contains 307 records, each representing one argument slot for a given noun, characterized by its semantic role, syntactic realization type, semantic annotation of the filler class, the full argument structure pattern, an authentic corpus example, and a list of German nouns

[4] This being said, since the training data of the models evaluated are not publicly disclosed, it is not possible to determine whether or to what extent they have been exposed to e-lexicography literature during pre-training. The claim in the main text should therefore be read narrowly: none of the models received any lexicography-specific fine-tuning, instruction-tuning, or in-prompt training for this study; this is not the same as certainty about what, if anything, they encountered on this topic during pretraining.

[5] Some of these requirements – for instance, providing multiple access routes to the same content – overlap with general guidelines for accessible web design (cf. the Web Content Accessibility Guidelines, WCAG); what we characterize as domain-specific here is not the abstract principle in isolation, which lexicography shares with accessibility-oriented web design more broadly, but its particular instantiation and prioritization in a lexicographic context, informed by e-lexicographic scholarship on dictionary consultation behaviour.

attested as fillers in that slot. The data encompasses 24 distinct actant codes, various morphosyntactic realizations (e.g., prepositional phrase, genitive phrase, compound, and attributive adjective), and semantic annotations spanning domains such as human agents, institutions, events, or abstract entities.

TABLE 1 Illustrative record from the dataset (noun: *Diskussion*)

Field	Value
Noun	*Diskussion*
Actant code	A1N
Semantic role	Agent (that which performs an action)
Syntactic realization	Adjective (Nominative)
Semantic annotation (filler class)	{animate, human, religious belief}
Argument structure pattern	determiner + [adjective] + actant + nucleus
Language	de (German)
Corpus example	*die* [*theologische*] *anglikanische Diskussion*
N (attested fillers)	17
Attested fillers (excerpt)	*anglikanisch*, *asketisch*, *atheistisch*, *buddhistisch*, *calvinistisch*, *christlich*, *gnostisch*, *heidnisch*, *jüdisch*, *kalvinistisch*, *katholisch*, *lutherisch*, *moslemisch*, *muslimisch*, *orthodox*, *pantheistisch*, *theistisch*

This dataset was prepared by linguists with rigorous annotation schemes and machine-readable formatting, ensuring that the LLMs were tasked exclusively with structuring verified data rather than generating linguistic content.

3.2 Model selection

Three LLMs were selected for evaluation based on their ranking as leading performers in agentic coding tasks – as defined in Section 1 – according to the Artificial Analysis leaderboard as of April 2026: Claude Opus 4.6 (Anthropic), ChatGPT-5.5 (OpenAI), and Gemini 3.1 Pro (Google). The selection criterion was oriented specifically towards this agentic coding capability rather than general language performance, as the task under evaluation is fundamentally a software engineering task.

3.3 Prompt design

A single prompt was used for all three models, following a zero-shot strategy in which no prior examples of lexicographic Shiny applications were provided. The prompt was structured in several components:

(a) A role specification instructing the model to act as an expert software developer (R/Shiny) and UI/UX designer specializing in educational technology – expertise framed here in software-engineering and instructional-design terms, not in linguistics or lexicography, which remained the researchers' domain throughout;

(b) An explicit instruction protecting the integrity of the linguistic content, stating that "The data has been curated by linguists and is accurate – your task is exclusively to build the software infrastructure around it, not to interpret or modify the linguistic content";

(c) The complete CSV dataset embedded directly in the prompt;

(d) A detailed column description explaining the meaning of each field in the dataset;

(e) A full actant code lookup table mapping each notation code to its human-readable semantic role and syntactic realization;

(f) Technical build instructions specifying the R/Shiny stack, allowed packages, and the requirement to produce a single self-contained app.R file;

(g) A target user profile describing intermediate to advanced learners of German as a foreign language (B1–C1) with some linguistic training;

(h) A request for a structured design rationale addressing aspects such as layout and navigation, handling of technical notation, lemma list presentation, colour and visual encoding, and the model's own operationalization of user-friendliness.

This last component is central to the study: by asking each model to justify its design choices and to articulate what it understood by user-friendliness, we obtain both the generated interface and the model's explicit reasoning about its design, enabling a systematic comparison of how different LLMs conceptualize lexicographic usability. Models were not explicitly instructed to rank or prioritize these features by importance; component (h) asked them to identify and justify design choices, not to produce an ordered list.

Component (b) explicitly protected the integrity of the linguistic content, instructing each model that the data had been curated by linguists and was accurate, and that its task was exclusively to build the software infrastructure around it, not to interpret or modify the linguistic content. Manual inspection of the three generated app.R files after generation confirmed compliance with this instruction. All discrepancies identified concerned interface logic and design (see Section 4.1), not the linguistic content itself.

3.4 Application generation and deployment

Each model produced a single app.R file that was tested locally and subsequently deployed to shinyapps.io for evaluation. Minor adjustments were necessary to ensure deployment compatibility, but no substantive changes were made to the interface design or functionality generated by the models. The three deployed applications are publicly accessible[6] and were used as the basis for the evaluation described below.

3.5 Evaluation framework

The evaluation framework operationalizes the concept of user-friendliness from the functional, user-oriented perspective developed in Section 2. The analysis centres on the outcome of the interaction from the user's perspective rather than on the internal technical properties of the

[6] In order of mention: https://ivanarias98.shinyapps.io/shiny_app_chatgpt/ (ChatGPT), https://ivanarias98.shinyapps.io/shiny_app_gemini/ (Gemini), and https://ivanarias98.shinyapps.io/shiny_app_claude/ (Claude). All URLs last accessed 30 April 2026.

system, attending to criteria such as task adequacy, cognitive economy, and perceived reliability, in line with human-centred evaluation approaches (Alabduljabbar, 2024; Blodgett et al., 2024) and with the principles of effectiveness, efficiency, and satisfaction formulated in ISO 9241-11:2018 (ISO, 2018). It should be stressed from the outset that this evaluation reflects expert judgement rather than direct observation of end users interacting with the applications. The rubric comprises nine interdependent dimensions, each rated on a five-point Likert scale (1 = strongly disagree, 5 = strongly agree):

TABLE 2 Dimensions of the rubric to evaluate user-friendliness by LLMs

Dimension	Description
D1. User-oriented definitional clarity	This dimension evaluates whether the interface enables immediate and unambiguous comprehension of the core linguistic content – whether information is clear, non-circular, and finite, and whether it provides an effective point of entry into the data (Alabduljabbar, 2024; Blodgett et al., 2024; Müller-Spitzer, 2014; Rundell, 1998; Tarp, 2008). In this case, this includes whether the interface conveys the semantic profile of each noun and anticipates its possible argument structures.
D2. Task alignment	This facet assesses the correspondence between the information offered by the interface and the communicative or learning task it is designed to support. Informational relevance depends on its fit with the intended use rather than on its descriptive value in isolation, and empirical evidence shows that irrelevant information increases cognitive load and reduces consultation effectiveness (Blodgett et al., 2024; Kim, Maathuis & Sent, 2024; Lew, 2011; Tarp, 2008). Here, task alignment is evaluated with respect to the description of the noun's valency and combinatorial behaviour.
D3. Cognitive load management	This aspect focuses on whether the presentation facilitates the identification and processing of the linguistic information presented. In digital environments, information architecture and visual structuring directly influence ease of use, and empirical evidence confirms that hierarchically organized, scannable entries with proportional extension and no redundancies reduce consultation time and errors (Alabduljabbar, 2024; Kim et al., 2024; Müller-Spitzer, 2014; Wolfer et al., 2021). In the present study, this is assessed in relation to the presentation of argument structures and their associated information.
D4. Usage grounding	This dimension evaluates whether the examples provided effectively help the user understand the use and combinatorial behaviour of the linguistic unit described. Examples contribute to user-friendliness only when they are prototypical and linguistically natural, explicitly illustrate the relevant patterns, and make visible the relevant combinatorial properties of the unit (Dziemianko, 2006; Müller-Spitzer & Wolfer, 2018; Rundell, 1998). In the scope of this research, the examples are drawn from the Portlex corpus and evaluated in terms of how effectively the interface presents them in relation to the argument structures described.
D5. Metalanguage accessibility	This criterion concerns whether the metalanguage employed is adequate for the target user profile. Empirical evidence confirms that metalanguage constitutes one of the main barriers to accessing lexicographic information and that its management influences perceived user-friendliness more decisively than descriptive precision itself (Dziemianko, 2006; Müller-Spitzer, 2014). This is evaluated with respect to the terminology used to describe argument structure and valency.

TABLE 2 (*CONTINUED*)

Dimension	Description
D6. Transparency and perceived reliability	This area assesses whether the interface conveys the reliability and scope of the information presented. It considers whether the system makes its sources and limitations visible, avoids presenting uncertain or partial information as established fact, and provides explicit cues that enable users to gauge the trustworthiness of the content. In the context of this study, this includes whether the interface distinguishes between well-attested combinatorial patterns and marginal or context-dependent uses. Unfounded assertions or the absence of explicit uncertainty markers rapidly erode user trust (Liao & Vaughan, 2024; Müller-Spitzer, 2014; Rundell, 1998; Tarp, 2008; Xu et al., 2024).
D7. Internal consistency	This aspect focuses on whether the information presented across the interface is coherent and free from contradictions. Internal inconsistencies – such as mismatches between different sections, unstable descriptive approaches, or conflicting information across views – increase cognitive load, lead to consultation abandonment, and negatively affect perceived reliability (Blodgett et al., 2024; Müller-Spitzer, 2014; Xu et al., 2024). In the present study, this includes coherence between definitions, examples, and the combinatorial patterns displayed for each noun.
D8. Actionability	This dimension evaluates whether, after consulting the interface, the user would plausibly be able to apply the linguistic information in productive language use. Because no end users interacted directly with the applications, actionability was assessed by expert evaluators as the user's hypothesized ability, on the basis of the interface, to construct well-formed phrases, select appropriate complements and prepositions, and avoid unacceptable combinations, given the valency information provided (Alabduljabbar, 2024; Kim et al., 2024; Lew, 2011; Tarp, 2008); it therefore reflects expert-inferred rather than empirically observed actionability.
D9. Explicitness of argument structure and valency	This criterion examines whether the interface makes the noun's valency explicit in a usable manner. It considers whether the argument participants involved are identified, whether categorical roles are made visible, whether syntactic realization is indicated, and whether the number and type of required or habitual arguments can be recognized.

Four lexicographers with expertise in German linguistics and lexicography independently rated the three LLM-generated applications across the nine dimensions, resulting in a total of 108 individual ratings. Inter-rater reliability was assessed using Gwet's AC2, which is robust against the prevalence paradox that affects Cohen's and Fleiss' kappa when distributions are skewed (Gwet, 2014). The evaluation instrument was implemented as an HTML form that allows evaluators to select the LLM under assessment, consult the corresponding Shiny application in an embedded frame, rate each dimension on the Likert scale, and download the results as a CSV file for subsequent analysis.

4 Evaluation and discussion of results

This section presents the results of the comparative evaluation in three stages. Section 4.1 describes the interfaces generated by each model and analyses how each LLM conceptualizes user-friendliness in its design rationale. Section 4.2 reports the quantitative ratings assigned by the four evaluators across the nine dimensions, including inter-rater agreement. Section 4.3 integrates both perspectives, contrasting the models' stated design priorities with the evaluators' empirical assessments.

4.1 Interface design and self-reported user-friendliness

All three applications allow users to explore the argument slots of German nouns through human-readable labels, semantic annotation, and authentic corpus examples. However, the interfaces differ substantially in structural complexity, interaction design, and visual elaboration.

TABLE 3 Overview of the interface designs generated by the three LLMs (see Section 3.4 for the corresponding application URLs)

ChatGPT	It produced the structurally simplest solution: a sidebar with filtering controls (noun selection, argument type, lemma search) and a code legend, paired with a main panel containing two tabs. The Overview tab displays a widget intended to show the valency data, while Details statically lists four categories (Example, Syntactic Pattern, Semantic Role, Lemma List) as separate output fields. The sidebar controls do not update the table content and the noun selector contains duplicates, constituting a critical implementation failure.[a]
Gemini	It adopted a three-tab navbar layout. Valency Overview presents a paginated table with five columns showing the semantic role, argument realization, the semantic class, the full syntactic structure, and the example. Lemma Browser implements a master-detail[b] pattern: the left-hand table lists the available argument slots, and upon selecting one, the right-hand panel dynamically displays the nouns compatible with that slot. Technical Legend maps annotation codes to semantic roles and syntactic realizations from the linguist-curated classification.
Claude	It produced the most visually elaborate interface, also organized in three tabs. Word Explorer displays individual cards per argument slot, each headed by a colour-coded badge with semantic role, syntactic pattern, and example; selecting a card reveals compatible nouns with a total count. Full Data presents a unified filterable table across all slots and nouns, with a detail panel on row selection. Notation Guide breaks down the code structure with examples and a searchable reference table, based on the linguistic classification provided as part of the input data.

[a]A plausible explanation is that the source CSV contains one row per argument slot rather than one row per noun; if the model populated the noun-selector list directly from the raw noun column without deduplicating it, each noun would appear once per associated argument slot, producing the duplicate entries observed. This remains a post hoc hypothesis, as the model's underlying reasoning is not directly observable; we flag it as a hypothesis rather than an established explanation.
[b]Cf. the master-detail interface pattern as documented at https://web.archive.org/web/20060526234555/http://www.oracle.com/technology/tech/blaf/specs/masterDetail_template.html (last accessed 30 April 2026).

When asked to justify their design choices, the three models articulated substantially different priorities (Table 3). All three named progressive disclosure as their first listed principle, though with different scope: ChatGPT formulated it generically as complexity management ("show complexity only when needed"); Gemini linked it specifically to one concrete interaction – hiding compatible nouns until the user selects a particular argument slot; Claude extended it to every layer of the interface (noun lists, code explanations, filters), making it a systemic architectural principle rather than a local interaction pattern.

Models were not explicitly instructed to rank these criteria by importance; component (g) of the prompt (Section 3.3) asked them to justify their design choices, not to produce an ordered list. The sequence shown in Table 3 reflects the order in which each model listed its criteria in its free-text design rationale and is reported here for descriptive completeness; it should not be over-interpreted as a validated priority ranking, even though it is suggestive of relative emphasis.

These design priorities – which map onto distinct subsets of the evaluation rubric (Table 2) – can be summarized as follows: for ChatGPT, user-friendliness is primarily about managing

TABLE 4 User-friendliness criteria as listed by each model (order of mention, not an explicit ranking)

ChatGPT	Gemini	Claude
Progressive disclosure	Progressive disclosure	Progressive disclosure
Semantic transparency	Descriptive labelling	Consistent visual grammar
Cognitive economy	Master-detail interaction	Zero raw code
Explorability	Contextual reinforcement	Flat interaction depth
Consistency	Searchable lexemes	Corpus grounding
Linguistic fidelity	Visual hierarchy	Learnable structure
Scannability		Minimal jargon
Learnability		

cognitive load to facilitate understanding (D2 and D3); for Gemini, it centres on enabling active and productive use of lexical information (D4, D8); for Claude, it resides in providing a clear, consistent, and cognitively stable architecture (D5, D6, D7). Whether these stated priorities are reflected in the evaluators' ratings is examined in Section 4.3.

4.2 *Quantitative evaluation*

4.2.1 Overall ratings

Figure 1 presents the overall mean ratings for each LLM across all nine dimensions (see Section 3.5) and four evaluators (n = 36 ratings per model). Claude obtained the highest overall mean (M = 4.06, SD = 0.86, Md = 4), followed by Gemini (M = 3.03, SD = 1.46, Md = 3) and ChatGPT (M = 2.08, SD = 1.05, Md = 2). The global distribution of ratings underscores the magnitude of these differences: 77% of Claude's ratings fell at 4 or 5, while 69% of ChatGPT's fell at 1 or 2. ChatGPT received no rating of 5 on any dimension from any evaluator. Gemini's distribution was more dispersed, reflecting a heterogeneous profile with both strong and weak dimensions.

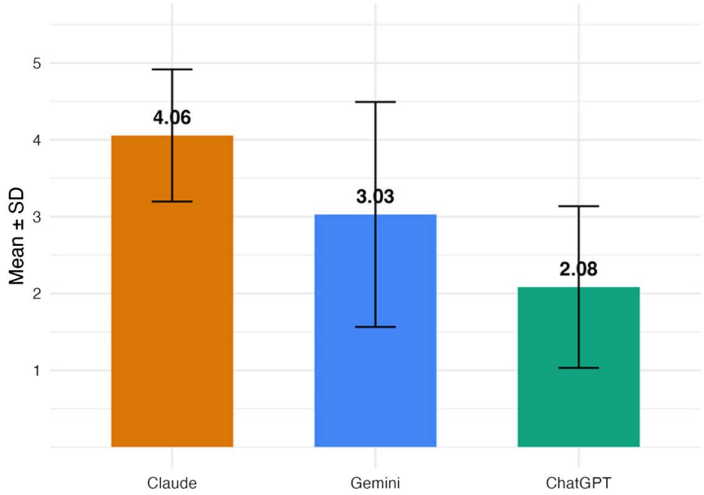


FIGURE 1 Overall mean user-friendliness scores (±SD) per LLM across all nine evaluation measures (n = 36 ratings per model)

4.2.2 Dimensional profiles

The dimensional profiles reveal that the three models differ not only in their overall performance but in the specific aspects of user-friendliness they implement most effectively (cf. Figure 2 and Figure 3).

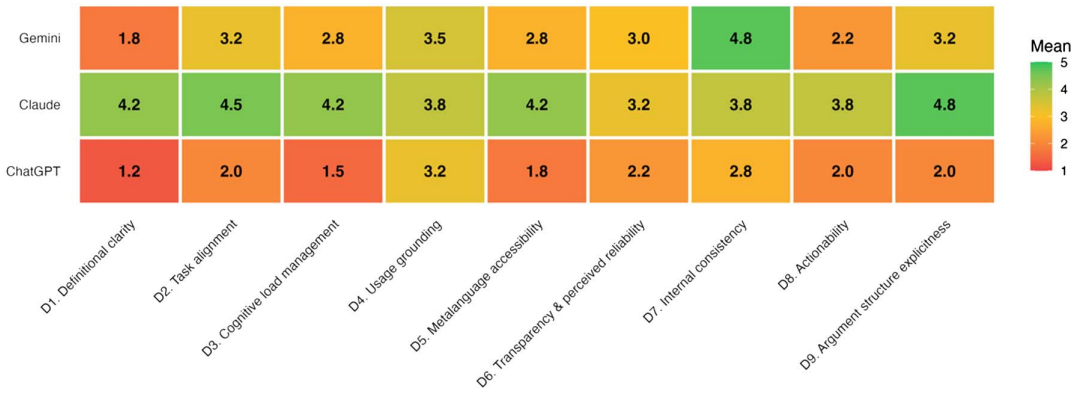


FIGURE 2 Heatmap of mean scores per LLM and dimension

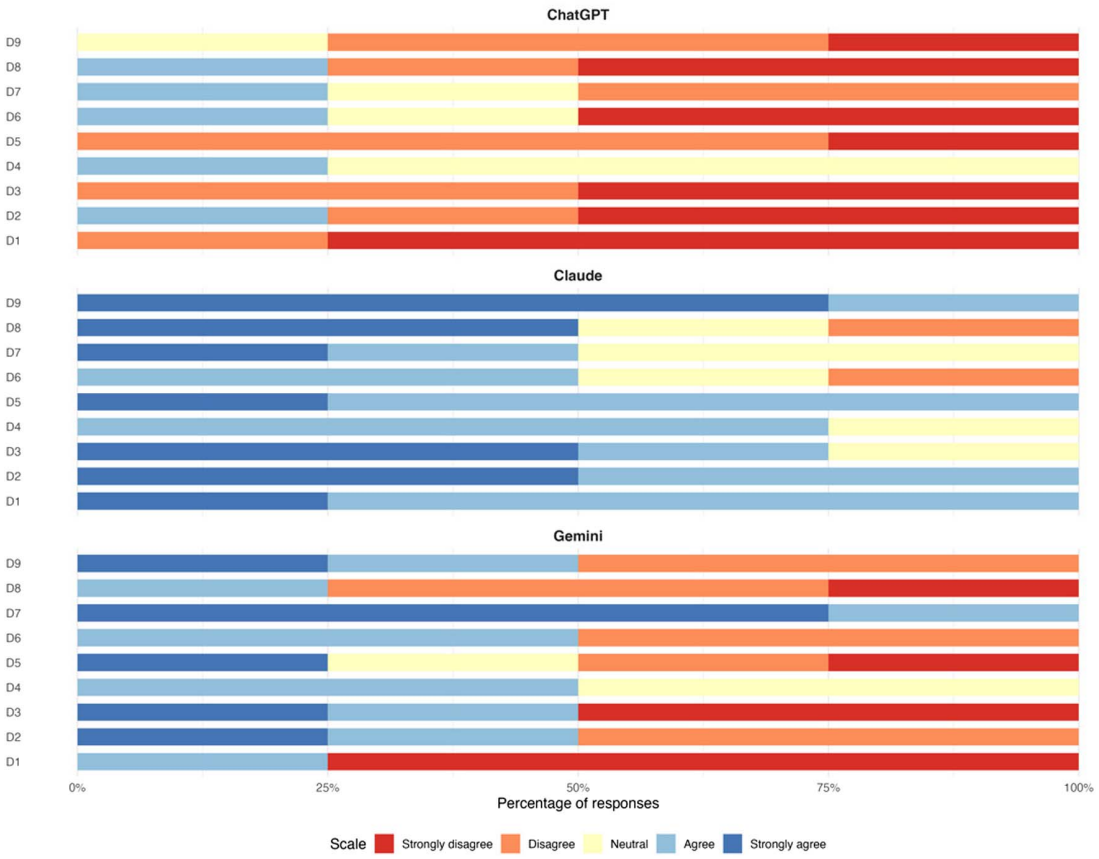


FIGURE 3 Stacked Likert bar chart per LLM

Claude received the highest mean rating in seven of the nine dimensions, with particularly strong scores in D9 (M = 4.75, SD = 0.50), D2 (M = 4.50, SD = 0.58), D1 (M = 4.25, SD = 0.50), D3 (M = 4.25, SD = 0.96), and D5 (M = 4.25, SD = 0.50). Its lowest-rated dimension was D6 (M = 3.25, SD = 0.96); at 1.00, its range across models is also comparatively narrow – though, as noted below, not the narrowest overall – suggesting that transparency and perceived reliability remain a challenge for all three systems, irrespective of their overall performance differences. Gemini's most distinctive result was D7 (M = 4.75, SD = 0.50), where it achieved the highest score of any model on any dimension. Three of the four evaluators assigned a score of 5, reflecting a strong consensus that the Gemini-generated interface maintains a coherent descriptive approach throughout. However, Gemini scored notably lower on D1 (M = 1.75, SD = 1.50) and D8 (M = 2.25, SD = 1.26), indicating that internal coherence did not translate into clarity of presentation or practical usability for the end user. ChatGPT scored below 2.00 on five of the nine dimensions (D1, D3, D5, D8, D9), consistent with the functional disconnection between filter controls and data display observed in the interface description. Its strongest dimension was D4 (M = 3.25, SD = 0.50), which – notably – is also the dimension with the smallest range across all three models: 0.60, spanning from 3.2 (ChatGPT) to 3.8 (Claude), rather than the previously narrower dimension mentioned for D6. This convergence in D4 is attributable to the fact that usage examples are drawn directly from the linguist-curated corpus rather than generated by the model, making this the dimension least dependent on the LLM's design decisions.

While Figure 2 summarizes central tendencies, Figure 3 additionally shows the full distribution of ratings within each dimension. This makes visible, for instance, the extent to which ChatGPT's low means on D1, D3, D5, D8 and D9 stem from consistently low ratings across all four evaluators rather than from a small number of outlying scores – information that the means and standard deviations reported above do not, on their own, convey.

4.2.3 Inter-rater agreement

Inter-rater reliability was assessed using Gwet's AC2 with ordinal weights, as recommended for ordinal scales where prevalence effects may inflate or deflate traditional kappa statistics (Gwet, 2014). Gwet's AC2 is bounded between –1 (systematic disagreement beyond what chance would predict) and 1 (perfect agreement), with 0 indicating agreement no better than chance; values above roughly 0.60 are conventionally described as substantial, though thresholds vary across the literature (Gwet, 2014). The evaluators showed notable differences in overall severity: Evaluators 1 and 2 assigned higher mean ratings (M = 3.67 and 3.63, respectively), while Evaluators 3 and 4 were more conservative (M = 2.33 and 2.59)[7]. Despite this difference in calibration, the evaluators showed substantial agreement on the relative ranking of the three models: all four consistently rated Claude highest and ChatGPT lowest across the majority of dimensions.

For six of the nine dimensions (D1, D2, D3, D5, D6, D9), the upper bound of the 95% CI reaches the coefficient's theoretical ceiling of 1.000. This reflects a standard property of the Wald-type interval used here (coefficient value ± 1.96 × SE, bounded within the [−1, 1] range in which AC2 is defined): whenever the unbounded interval would exceed this range, the reported bound is capped accordingly rather than reported as a value outside the coefficient's possible range.

[7] This pattern reflects the well-known distinction in annotation and classification research between splitters – evaluators who tend to draw finer distinctions and assign more differentiated (often more severe) ratings – and lumpers, who favour broader, more lenient categorisations (O'Connor & Joffe, 2020).

TABLE 5 Inter-rater agreement (Gwet's AC2, ordinal weights) by scope. Coefficients were calculated using ordinal weighting (linear). 95% CI was estimated based on standard error distribution

Scope	AC2	95% CI
Global	0.3849	[0.19, 0.58]
Gemini	0.0988	[−0.318, 0.516]
ChatGPT	0.5779	[0.306, 0.849]
Claude	0.7169	[0.486, 0.948]
D1. Definitional clarity	0.6952	[0.069, 1.000]
D2. Task alignment	0.3416	[−0.861, 1.000]
D3. Cognitive load management	0.1822	[−1.632, 1.000]
D4. Usage grounding	0.8990	[0.856, 0.942]
D5. Metalanguage accessibility	0.4667	[−1.476, 1.000]
D6. Transparency & perceived reliability	0.3870	[−0.587, 1.000]
D7. Internal consistency	0.6543	[0.020, 1.000]
D8. Actionability	0.1111	[−0.758, 0.980]
D9. Argument structure explicitness	0.5226	[−0.459, 1.000]

Beyond calibration differences in severity, the pattern of statistical significance across scopes is informative in its own right. Overall agreement across the full set of 36 ratings is fair-to-moderate and significantly above chance (AC2 = 0.3849, 95% CI [0.19, 0.58], $p < .001$). Disaggregating by LLM reveals a sharp split concealed by this global figure: agreement among evaluators is substantial for Claude (AC2 = 0.7169, 95% CI [0.486, 0.948], $p < .0001$) and moderate-to-substantial for ChatGPT (AC2 = 0.5779, 95% CI [0.306, 0.849], $p = .001$), but statistically indistinguishable from chance for Gemini (AC2 = 0.0988, 95% CI [−0.318, 0.516], $p = .60$). Evaluators therefore converged far more consistently when judging Claude's and ChatGPT's interfaces than when judging Gemini's, echoing the more heterogeneous rating profile already noted for Gemini in Section 4.2.1.

At the level of individual dimensions, only three of the nine reach conventional significance: D4 (Usage grounding; AC2 = 0.8990, 95% CI [0.856, 0.942], $p < .001$), D1 (Definitional clarity; AC2 = 0.6952, $p = .041$), and D7 (Internal consistency; AC2 = 0.6543, $p = .047$). The remaining six dimensions yield wide, non-significant intervals. This pattern is best read as a sample-size limitation specific to the dimension-level breakdown: each dimension-level coefficient is computed over only three subjects (the three LLM interfaces) rated by four evaluators, whereas the global and per-model estimates pool over more subjects and correspondingly return tighter, significant intervals. D4's markedly narrow interval illustrates this directly: because its worked examples are drawn verbatim from the curated corpus rather than generated by the model (Section 4.2.2), evaluators had almost no room to disagree, yielding a very small standard error despite the same $n = 3$ subjects. Taken together, these results indicate that inter-rater agreement is fair-to-substantial at the level that matters most for the study's central claims – the relative ranking of the three models – while the finer-grained, per-dimension breakdown would benefit from a larger set of stimuli in future replications to achieve adequate statistical power.

The severity difference affects the absolute level of agreement coefficients, but it does not fully offset the interpretability of the results for the purposes of this study, as the evaluation is comparative rather than norm-referenced: the relevant, and more robust, finding is the consistent ordering across models, not the absolute score assigned to any single dimension.

4.3 Integrating design rationales and evaluator assessments

The juxtaposition of the models' self-reported design priorities (Section 4.1) with the evaluators' ratings (Section 4.2) reveals a consistent pattern: the dimensions that each model claims to prioritize are not necessarily the dimensions on which it performs best according to expert assessment.

ChatGPT's stated emphasis on cognitive load management and progressive disclosure (D2, D3) is contradicted by its lowest scores precisely on those dimensions (M = 2.00 and 1.50). The discrepancy is explained by the functional failure detected in the implementation: the filters in the sidebar do not update the data display, rendering the stated principle of progressive disclosure inoperative in practice. This finding illustrates a fundamental risk of agentic coding: a model may articulate sound design principles while failing to implement them in functional code.

Gemini's focus on actionability and contextual reinforcement through examples aligns with D4 (usage grounding) and D8 (actionability). While D4 received a moderately favourable score (M = 3.50), D8 was rated notably lower (M = 2.25), suggesting that interaction in Lemma Browser does not sufficiently enable users to construct well-formed phrases or select appropriate complements – the core of actionability in a valency dictionary. Conversely, Gemini scored highest on D7 (internal consistency; M = 4.75), a dimension it did not explicitly identify as a priority – a case where the interface's implicit design quality exceeds what the model articulated about its own design.

Claude's stated priorities – visual consistency, flat interaction depth, minimal jargon exposure, and zero unexplained codes – can be read as claims to prioritize D5 (metalanguage accessibility), D6 (transparency), and D7 (internal consistency). The evaluators confirmed Claude's strength in metalanguage management (D5; M = 4.25), but rated D6 (transparency; M = 3.25) and D7 (consistency; M = 3.75) lower than might be expected given the model's explicit emphasis on these properties. Notably, Claude's highest-rated dimensions – D9 (explicitness of argument structure; M = 4.75) and D2 (task alignment; M = 4.50) – are aspects it did not foreground in its design rationale. This suggests that Claude's design strengths derive at least in part from latent UX competencies absorbed during training, rather than from the deliberate operationalization of the stated principles.

A cross-cutting observation concerns D4 (usage grounding), the dimension with the smallest inter-model range (0.60, spanning from 3.2 to 3.8). Since all examples are drawn from the linguist-curated Portlex dictionary and displayed as received, D4 reflects the quality of the input data rather than the model's design decisions. This convergence provides indirect validation of the study's central premise: when linguistic accuracy is assured by curated data, the relevant variable is the quality of the interface, not the quality of the content.

Finally, the analysis reveals a hierarchy of agentic coding reliability across the three models. Claude demonstrates the most consistent capacity to translate design principles into functional implementations, though its strongest dimensions are not always those it explicitly identified as priorities. Gemini produces a functionally sound application with particular strength in internal consistency, but limited actionability. ChatGPT articulates the most comprehensive set of design principles, yet produces the weakest implementation – underscoring that the quality of a model's design rationale is not a reliable predictor of the quality of its generated code.

5 Final remarks

The results of this study have several implications for the use of LLMs as agentic coding tools in lexicography. First, the fact that the same prompt leads to markedly different conceptualizations and implementations of user-friendliness depending on the LLM used highlights that usability is not a stable or model-independent outcome. Rather, user-friendliness emerges as a negotiated property, shaped by each model's implicit design heuristics, training-related biases, and effective implementation capacity, even when linguistic data and task specifications are strictly controlled.

Based on the elements recurrently identified by all three LLMs, it is possible to formulate a minimal shared definition of user-friendliness as the property of an interface that enables users to access, explore, and understand complex linguistic information without prior technical decoding effort, through a content organization that is progressive, semantically transparent, and cognitively efficient. This convergence suggests that the selected LLMs share an abstract and general understanding of usability that aligns, at least partially, with established principles of user-oriented lexicography. However, as the analysis shows, this shared conceptual baseline does not guarantee coherence in design quality or functional adequacy across the generated interfaces.

From an applied perspective, the misalignments between the priorities declared by the LLMs and expert evaluations reveal a limitation inherent in using these models as autonomous design agents: the ability to articulate plausible usability principles does not necessarily entail their correct implementation in code. LLMs can serve as valuable tools for accelerating prototyping and lowering technical barriers, particularly when high-quality, expertly curated linguistic data are available, but systematic expert evaluation grounded in established lexicographic criteria remains indispensable.

More broadly, the study shows that in agentic coding scenarios, the primary risk does not lie in linguistic hallucination, but in usability misalignment. Two limitations should temper this conclusion. First, each model produced a single application from a single prompt execution; since LLMs are stochastic systems, output variability across multiple runs remains a relevant avenue for future replication work, alongside the inclusion of additional models, user profiles, or lexicographic tasks, and the combination of expert assessment with empirical user studies. Second, this limitation is compounded by the small effective sample size underlying the dimension-level reliability estimates documented in Section 4.2.3 (n = 3 subjects per dimension), which leaves six of the nine dimensions statistically underpowered; the overall and per-model reliability estimates are comparatively robust and significant, but dimension-level conclusions should be treated as provisional pending a larger set of stimuli.

Ultimately, the question is not whether LLMs can generate lexicographic interfaces, but whether they can do so in ways that are genuinely user-friendly by lexicographic standards. The comparative evidence presented here suggests that this remains model-dependent, requires expert verification rather than trust in a model's self-reported design rationale, and – as with agentic coding more generally – shifts the locus of quality control from linguistic accuracy to interface reliability and usability.

References

Alabduljabbar, R. (2024). User-centric AI: Evaluating the usability of generative AI applications through user reviews on app stores. *PeerJ Computer Science*, 10, e2421. https://doi.org/10.7717/peerj-cs.2421

Arias-Arias, I., Domínguez Vázquez, M. J., Valcárcel Riveiro, C. (2024). Der Effizienz- und Intelligenzbegriff in der Lexikographie und künstlichen Intelligenz: Kann ChatGPT die lexikographische Textsorte nachbilden? *Lexikos*, 34(1), 51–76. https://doi.org/10.5788/34-1-1879

Blodgett, S. L., Cheung, J. C. K., Liao, Q. V. & Xiao, Z. (2024). Human-centered evaluation of language technologies. In J. Jessy Li & F. Liu (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts* (pp. 39–43). Association for Computational Linguistics.

Bothma, T. J., Fourie, I. (2024). Enhancing conceptualisations of information behaviour contexts through insights from research on e-dictionaries and e-lexicography. *Information Research an International Electronic Journal*, 29(2), 179–197. https://doi.org/10.47989/ir292821

de Schryver, G.-M., Joffe, D. (2023). The dictionary of the future is here: Generative AI as a lexicographic tool. *International Journal of Lexicography*, 36(4), 389–412.

Domínguez Vázquez, M. J. & Valcárcel Riveiro, C. (2020). PORTLEX as a multilingual and cross-lingual online dictionary. In M. J. Domínguez Vázquez, M. Mirazo Balsa, & C. Valcárcel Riveiro (Eds.), *Studies on multilingual lexicography* (pp. 135–158). De Gruyter. https://doi.org/10.1515/9783110607659-008

Dong, Y., Jiang, X., Qian, J., Wang, T., Zhang, K., Jin, Z., Li, G. (2026). A survey on code generation with LLM-based agents. *Journal of Software*, 37(8), 1–27. https://doi.org/10.13328/j.cnki.jos.007593

Dziemianko, A. (2006). *User-friendliness of verb syntax in pedagogical dictionaries of English*. Niemeyer.

Fuertes-Olivera, P. A. (2024). Making lexicography sustainable: Using ChatGPT and reusing data for lexicographic purposes. *Lexikos*, 34, 123–140. https://doi.org/10.5788/34-1-1883

Fuertes-Olivera, P. A. (2025). *A Guide to Practical Online Lexicography*. Routledge.

Gwet, K. L. (2014). *Handbook of Inter-Rater Reliability: The Definitive Guide to Measuring the Extent of Agreement among Raters*. Advanced Analytics, LLC.

Huete-García, A. (2023). *Introducción a la Teoría de las Funciones Lexicográficas*. Arco/Libros – La Muralla.

International Organization for Standardization. (2018). *ISO 9241-11:2018: Ergonomics of human–system interaction — Part 11: Usability: Definitions and concepts*. ISO.

Karpathy, A. (2025). *2025 LLM year in review*. Bear Blog. Retrieved April 30, 2026, from https://karpathy.bearblog.dev/year-in-review-2025/

Kemmer, K. (2010). *Onlinewörterbücher in der Wörterbuchkritik: Ein Evaluationsraster mit 39 Beurteilungskriterien*. Institut für Deutsche Sprache. (OPAL – Online publizierte Arbeiten zur Linguistik, 2010/2).

Kim, J., Maathuis, H., Sent, D. (2024). Human-centered evaluation of explainable AI applications: A systematic review. *Frontiers in Artificial Intelligence*, 7, 1456486. https://doi.org/10.3389/frai.2024.1456486

Klosa-Kückelhaus, A. & Michaelis, F. (2024). The design of Internet dictionaries. In A. Klosa-Kückelhaus (Ed.), *Internet Lexicography* (pp. 171–190). De Gruyter.

Lew, R. (2011). User-centred design of electronic dictionaries. In P. Fuertes-Olivera & H. Bergenholtz (Eds.), *e-Lexicography: The Internet, Digital Initiatives and Lexicography* (pp. 343–361). Continuum.

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities & Social Sciences Communications*, 11. https://doi.org/10.1057/s41599-024-02889-7

Lew, R., Wolfer, S. (2024). What lexical factors drive look-ups in the English Wiktionary? *SAGE Open*, 14(1). https://doi.org/10.1177/21582440231219101

Liao, Q. V., Vaughan, J. W. (2024). AI transparency in the age of LLMs: A human-centered research roadmap. *Harvard Data Science Review*. https://doi.org/10.1162/99608f92.8036d03b

Lugli, L. (2024). Agile lexicography: Rapid dictionary prototyping with R Shiny, with examples from projects on Sanskrit and Tibetan. In J. Martín Arista & A. E. Ojanguren López (Eds.), *Structuring Lexical Data and Digitising Dictionaries* (pp. 121–149). Brill.

Müller-Spitzer, C. (2022). User-adapted views in online dictionaries. In A. Klosa-Kückelhaus & F. Michaelis (Eds.), *Internet lexicography: An introduction* (pp. 191–214). De Gruyter. https://doi.org/10.1515/9783110712599-010

Müller-Spitzer, C. (Ed.). (2014). *Using online dictionaries*. De Gruyter.

Müller-Spitzer, C. & Koplenig, A. (2014). Online dictionaries: Expectations and demands. In C. Müller-Spitzer (Ed.), *Using online dictionaries* (pp. 143–188). De Gruyter.

Müller-Spitzer, C. & Wolfer, S. (2018). Nutzerstudien zur lexikographischen Beispieldarstellung. In I. Kosem et al. (Eds.), *Electronic lexicography in the 21st century* (pp. 159–176). Lexical Computing CZ.

Müller-Spitzer, C., Domínguez Vázquez, M. J., Nied Curcio, M., Silva Dias, I. M., Wolfer, S. (2018). Correct Hypotheses and Careful Reading Are Essential: Results of an Observational Study on Learners Using Online Language Resources. *Lexikos*, 28(1), 287–315. https://doi.org/10.5788/28-1-1466

O'Connor, C. & Joffe, H. (2020). Intercoder reliability in qualitative research: Debates and practical guidelines. International Journal of Qualitative Methods, 19. https://doi.org/10.1177/1609406919899220

Rasheed, Z., Waseem, M., Kemell, K. K., Ahmad, A., Sami, M. A., Rasku, J., Systä, K., Abrahamsson, P. (2025). Large language models for code generation: The practitioners' perspective. arXiv. https://arxiv.org/abs/2501.16998

Robbes, R., Matricon, T., Degueule, T., Hora, A. & Zacchiroli, S. (2026). Agentic much? Adoption of coding agents on GitHub. arXiv. https://doi.org/10.48550/arXiv.2601.18341

Rundell, M. (1998). Recent trends in English pedagogical lexicography. *International Journal of Lexicography*, 11(4), 315–342.

Rundell, M. (2023). Automating the creation of dictionaries: Are we nearly there? In *Proceedings of ASIALEX 2023: Lexicography, Artificial Intelligence and Dictionary Users* (pp. 9–17). The Asian Association for Lexicography. https://asialex.org/pdf/Asialex-Proceedings-2023.pdf

Smutny, P., Bojko, M. (2024). Comparative analysis of chatbots using large language models for web development tasks. *Applied Sciences*, 14(21). https://doi.org/10.3390/app142110048

Stöckle, P., Elsner, D., Koppensteiner, W. & Korecky-Kröll, K. (2025). LLM-assisted dialect lexicography: Challenges and opportunities. In I. Kosem, M. Jakubíček, M. Medved, K. Zgaga, Š. Arhar Holdt, T. Munda, & A. Salgado (Eds.), *Electronic Lexicography in the 21st Century (eLex 2025): Intelligent Lexicography* (pp. 453–475). Lexical Computing CZ.

Tarp, S. (2008). *Lexicography in the borderland between knowledge and non-knowledge*. Niemeyer.

Wolfer, S., Michaelis, F., Müller-Spitzer, C. (2021). Nutzungsstatistiken von linguistischen Online-Ressourcen: Potentiale und Limitationen. *Zeitschrift für Angewandte Linguistik*, 75, 395–430. https://doi.org/10.1515/zfal-2021-2070

Xu, Y., Gao, W., Wang, Y., Shan, X., Lin, Y.-S. (2024). Enhancing user experience and trust in advanced LLM-based conversational agents. *Computing and Artificial Intelligence*, 2(2), 1467.

Contact information

Iván Arias-Arias
Instituto da Lingua Galega, Universidade de Santiago de Compostela
ivanarias.arias@usc.gal

María José Domínguez Vázquez
Instituto da Lingua Galega, Universidade de Santiago de Compostela
majo.dominguez@usc.es

Carlos Valcárcel Riveiro
Instituto de Investigación Lingua, Universidade de Vigo
carlos.valcarcel@uvigo.gal

María Teresa Sanmarco Bande
Instituto da Lingua Galega, Universidade de Santiago de Compostela
mt.sanmarco@usc.es

Acknowledgements

This contribution was developed within the framework of the research project ESMAS-ES+ (PID2022-137170OB-I00), funded by MCIU/AEI/10.13039/501100011033 and by the European Regional Development Fund (ERDF/EU). Iván Arias-Arias also acknowledges support from grant FPU21/00188, funded by the *Formación de Profesorado Universitario* (FPU) programme of the *Ministerio de Ciencia, Innovación y Universidades* (MCIU).

Igor Boguslavsky, Vyacheslav Dikonov, Evgeniya Inshakova,
Alexandre Lazursky, Svetlana Timoshenko & Tatiana Frolova

LEXICOGRAPHIC RESOURCES FOR INFERENCE-BASED SEMANTIC TEXT PROCESSING SYSTEM

Abstract This paper describes the lexicographic resources developed for the ETAP text analysis and generation system, with particular emphasis on its semantic module, SemETAP. In our approach, semantic analysis is viewed not merely as the construction of a semantic representation, but as the derivation of inferences licensed by linguistic and background knowledge. The semantic component operates in two stages. First, it constructs a Basic Semantic Structure (BSemS), which represents the meaning explicitly conveyed by the text. This representation is then enriched through ontological knowledge and inference rules, yielding an Extended Semantic Structure (EnSemS) that incorporates both explicit and implicit information. To support this process, ETAP relies on a set of large-scale, knowledge-rich lexicographic resources characterized by a high degree of reusability. Linguistic knowledge is distributed across a morphological dictionary, a combinatorial dictionary, and rule-based grammars, whereas background knowledge is represented in an ontology, including a repository of individuals, and a collection of inference rules. The paper describes the information encoded in these resources, shows formal languages used to represent it, and the interaction between linguistic and extralinguistic knowledge during semantic analysis.

Keywords Morphological dictionary; combinatorial dictionary; ontology; repository of individuals; inference; digital linguistic resources

1 Introduction

Knowledge-based natural language processing systems rely on a variety of linguistic resources. What kind of resources are specifically required depends on the goals pursued by the system. ETAP, and in particular its semantic module SemETAP, is designed for the analysis and generation of unrestricted natural-language texts. Its theoretical foundation is the Meaning–Text Theory (MTT) developed by Igor Mel'čuk (Mel'čuk, 1974).

The Meaning–Text framework is of particular relevance to lexicography because it assigns a central role to the lexicon in linguistic description. As Mel'čuk (2012, p.109) observes, "the Meaning–Text approach firmly places the lexicon at the center [of linguistic description]; grammar is viewed as a derivative component of language—a system of rules representing a generalization of the observable properties of lexical units". Within this framework, language is viewed as a system of correspondences between meanings and their means of expression, and these correspondences are encoded primarily in lexical rather than grammatical resources. Consequently, a substantial portion of the information required for text analysis and generation must be represented in lexicographic form.

The ETAP project has been developed consistently within the knowledge-based paradigm. This immediately raises the question of how such an approach relates to the recent success of Large Language Models (LLMs). Before turning to the main topic of the paper, a brief comparison is therefore in order.

The relationship between expert-compiled dictionaries and LLMs has become an important topic in contemporary lexicography (Zhao, 2023; de Schryver, 2023; Rundell, 2024; Lew, 2024). The impressive linguistic performance of LLMs sometimes creates the impression that

they can serve as universal dictionaries capable of replacing both traditional lexicographic resources and lexical databases developed for NLP applications. The prevailing view in current lexicographic research, however, is that LLMs do not eliminate the need for dictionaries. Rather than competitors, the two should be regarded as complementary tools serving different purposes.

LLMs excel at flexible, context-sensitive language processing. They can explain words in context, suggest paraphrases, assist in translation, and provide information in a conversational format. Dictionaries, in contrast, are designed to provide explicit, structured, verifiable, and stable linguistic knowledge. Several differences are particularly important for semantic text processing systems whose goal is to construct semantic representations and derive inferences from them.

First, lexicographic resources are transparent and explainable. Dictionaries record linguistic information explicitly, whereas knowledge in LLMs is distributed across model parameters and cannot normally be inspected directly. For systems that perform reasoning, it is essential to know which pieces of knowledge support a particular inference and how reliable that knowledge is.

Second, dictionaries provide stability. A dictionary entry remains unchanged regardless of the user's query. By contrast, an LLM may produce different descriptions of the same lexical item depending on the prompt, the conversational context, or stochastic factors involved in generation.

Third, lexicography places special emphasis on the systematic analysis of lexical meaning. Dictionaries identify and distinguish word senses, establish semantic relations among them, describe regular polysemy, and trace metaphorical extensions. LLMs often exhibit what may be termed *semantic blending*: they can discuss a word plausibly without necessarily maintaining clear boundaries between its distinct senses.

Finally, lexicographic descriptions are verifiable and open to scholarly scrutiny. Dictionary entries can be examined, challenged, revised, and improved. By contrast, LLMs may conflate meanings, generate unattested usages, or fail to distinguish between frequent and normatively accepted forms. Such limitations are particularly problematic in applications that require reliable semantic analysis.

Let us now turn to the main subject of the paper. The primary goal of SemETAP is to construct semantic representations enriched with information that can be inferred from the text on the basis of linguistic and background knowledge. SemETAP adopts the working assumption that the depth of text understanding is reflected in the range of inferences derivable from a semantic representation. The richer the set of derivable inferences, the deeper the level of understanding achieved by the system. Many of the inferences humans draw from texts depend on knowledge that extends beyond language itself. Understanding routinely relies on commonsense assumptions, knowledge of the world, information about the discourse situation, and assumptions concerning the interlocutors. This places substantial demands on both the quantity and the organization of knowledge available to the system.

ETAP employs several types of lexicographic and knowledge resources, each serving a distinct function while maintaining a high degree of reusability. Linguistic knowledge is distributed across the Morphological Dictionary, the Combinatorial Dictionary, and several sets of grammatical rules. Background knowledge is represented in the Ontology, and a collection of inference rules.

This paper addresses the following research questions:

1. What types of information must be represented in lexicographic resources to support semantic analysis based on both linguistic and background knowledge?
2. What formal languages are required to represent this information?
3. How do linguistic and background knowledge interact during semantic analysis?

Section 2 briefly introduces SemETAP. Section 3 describes the principal lexicographic and semantic resources employed by the system and illustrates their interaction. Section 4 concludes the paper.

2 SemETAP: A knowledge-rich semantic processing system

SemETAP is the semantic component of ETAP, a knowledge-based linguistic processor designed for the analysis and generation of natural-language texts. ETAP was developed within the general linguistic framework of the Meaning–Text Theory (MTT) proposed by Igor Mel'čuk (Mel'čuk, 1974). An important extension of this framework is provided by Jurij Apresjan's theory of systematic lexicography and integrated language description (Apresjan, 2000).

MTT aims to model two fundamental linguistic abilities: the production and understanding of natural-language texts. Accordingly, ETAP supports two principal modes of operation—text generation and text analysis. At present, SemETAP is primarily oriented toward semantic analysis. The morphological and syntactic components of ETAP, together with their associated linguistic resources, currently support Russian and English as their principal working languages. Prototype implementations are available for several other languages. The semantic component, however, currently operates only on Russian texts.

Text analysis proceeds in successive stages: morphological analysis, syntactic analysis, normalization, and semantic analysis. The morphological module processes each word form in the sentence and generates all possible morphological analyses, each consisting of a lemma, a part of speech, and a set of morphological features. The syntactic module takes this representation as input and builds a dependency structure. Its nodes correspond to lexical units annotated with morphological and syntactic information, while its arcs represent specific syntactic relations linking heads and dependents. A distinctive feature of the ETAP syntactic model is its fine-grained inventory of construction types: the system distinguishes 68 syntactic relations. The complete list of syntactic relations supplied with comments and examples can be found at https://ruscorpora.ru/file/annotation_syntagrus/.

Normalization eliminates certain lexico-syntactic peculiarities of the sentence. It removes syntactically motivated semantically empty prepositions and conjunctions, simplifies analytic constructions, identifies the values of lexical functions and the antecedents of anaphoric pronouns, and performs several other operations. The semantic module takes the normalized syntactic structure as input and proceeds in two stages. The first stage constructs **the Basic Semantic Structure (BSemS)**, which captures the information explicitly conveyed by the sentence. During this process, omitted arguments are restored, certain syntactic constructions are further normalized, semantically empty collocates are eliminated, and lexical units are replaced by their ontological correlates. In many cases, lexical meanings are decomposed into more elementary semantic components.

The second stage enriches the BSemS through ontological knowledge and inference rules, yielding **the Extended Semantic Structure (EnSemS)**. Unlike the BSemS, the EnSemS incorporates not

only information explicitly expressed in the text but also information inferred from linguistic and background knowledge encoded in the system's lexicographic and knowledge resources.

The following section describes the principal resources employed by ETAP and illustrates how they interact during semantic analysis.

3 Lexicographic and knowledge resources of ETAP

ETAP relies on a set of large-scale lexicographic and knowledge resources that support different stages of text processing. Linguistic knowledge is represented in the Morphological Dictionary, the Combinatorial Dictionary, and the grammatical rule sets. Background knowledge is represented in the Ontology, including its repository of individuals, and in the inference rules. The following subsections briefly characterize these resources and illustrate their role in semantic analysis.

3.1 Morphological Dictionary

The Morphological Dictionary (MD) is a large-scale lexicographic resource providing a comprehensive formal description of Russian inflectional morphology. It covers both nominal and verbal inflection and accounts for a variety of morphologically complex phenomena, including:

1. **Stem alternations**, e.g. *stanok* 'machine', nominative singular vs. *stank-a* 'machine', genitive singular;
2. **Suppletion**, e.g. *čelovek* 'human', singular vs. *ljudi* 'people', plural; *idu* 'go', present tense vs. *šel* 'go', past tense;
3. **Aspectual pairing**, including prefixal (*čitat'* 'read', imperfective – *pro-čitat'*, perfective), suffixal (*bros-a-t'* 'throw', imperfective – *bros-i-t'*, perfective), and stem-alternation (*sobr-at'*, 'gather', perfective – *sobir-at'*, imperfective) patterns;
4. **Additional inflectional forms**, including the Russian partitive case (*saxaru* 'sugar', partitive) and locative forms such as *v lesu* 'in the forest';
5. **Form synonymy**, e.g. *skorej* and *skoree* 'faster'; *streloj* and *streloju* 'with an arrow';
6. **Form homonymy**, e.g. the word form *steklo*, which admits several morphological analyses: *STEKLO* 'glass', nominative singular; *STEKLO* 'glass', accusative singular; and *STEKAT'* 'to flow down', past tense, neuter.

As of May 2026, the MD covers more than 137,000 lexemes.

An example of an MD entry is given below. The entry is reproduced in its original form, without transliteration of Cyrillic symbols.

```
ENTRY:ИДТИ1
[НЕСОВ!] acct:vn_bb
   хар:V,несов,осн:(шё`|шед|ш|ид)
   чер:изъяв,прош,ед,муж|прич,прош/деепр,прош|изъяв,прош,ед,жен/    *
   изъяв,прош,ед,сред/изъяв,прош,мн
   сок:102,110,ок:'ти'инф
   ф:16  ф:32  ф:42 ф:41 ф:19
|| acct:осн:{2}=ПОВ+МН;  хар:V,несов,1-л,мн,пов,осн:идё`мте \\(откуда куда)
   trs:go
```

FIGURE 1 MD entry for IDTI 'go'

3.2 Combinatorial Dictionary

The Combinatorial Dictionary (CD) occupies a central position among the lexicographic resources of ETAP. It serves as the principal interface between lexical, syntactic, semantic, ontological, and multilingual information used by the system. The CD is a reduced but fully formalized version of the explanatory combinatorial dictionaries developed within the Meaning–Text Theory. In particular, lexical definitions are omitted, while the combinatorial component is retained and represented in a form suitable for automatic processing. At present, it contains 113,739 entries and is used by both the syntactic and semantic components of ETAP.

The CD provides a wide range of lexicographic information describing the behaviour of lexical items at different linguistic levels. This information is organized into a number of dedicated zones within each lexical entry, each corresponding to a particular type of linguistic or processing information.

A distinctive feature of the CD is the division of lexical information into two major categories: classificatory information and operational information.

Classificatory information specifies the properties of lexical items that may be referenced by processing rules. These properties include, among other things, part of speech, syntactic and semantic features, subcategorization frames, lexical functions, and ontological correlates.

Operational information consists of processing rules associated with individual lexical items. The presence of such rules within lexical entries is one of the distinctive characteristics of the CD. ETAP distinguishes between two classes of rules. General rules, which apply to broad classes of lexical items and constitute part of the grammar, are stored separately from the dictionary. Lexically restricted rules, whose scope may be limited to a small group of words or even to a single lexical item, are stored directly in lexical entries.

This distinction reflects a more general principle of the Meaning–Text approach, according to which grammar can be viewed as a system of generalizations over lexical behaviour. Whenever a rule attains sufficiently broad coverage, it can be moved from the dictionary into the grammar. As a result, for each input sentence the system automatically constructs a task-specific rule set consisting of all general rules together with the lexicalized rules contained in the entries of the words occurring in that sentence.

The principal types of information represented in CD entries are briefly characterized below.

- **Comments and examples** illustrate the corresponding lexical sense and are used in interactive disambiguation. When the system cannot reliably determine the intended sense of a polysemous word, it may request assistance from the user. Comments and examples help the user select the appropriate interpretation.
- **Syntactic features** specify a lexical item's ability or inability to participate in particular syntactic constructions. The system currently employs approximately 200 such features. Examples include:

  RDOP — transitive verbs whose direct object may appear in the genitive rather than the accusative under negation: *Vse čitali etu knigu* ('book', accusative) 'Everybody read this book' vs. *Nikto ne čital etoj knigi* ('book', genitive) 'Nobody read this book'.
  SOGLAKT-2-1 — verbs requiring agreement between the first and second actants in number and gender, such as *byt'* 'be', *kazat'sja* 'seem', and *stanovit'sja* 'become'; *Plan* ('plan', masculine singular) *kažetsja nereal'nym* ('unrealistic', masculine singular) 'The plan seems unrealistic'.

- **Semantic features** are used throughout the processing pipeline, particularly in rules that check semantic compatibility between predicates and their actants. These features are represented directly in terms of ontology concepts. Examples include *Human*, *Animal*, *PhysicalObject*, *Event*, and *ContentBearingObject*.
- **Ontological correlates** specify the ontology concept used to represent the meaning of a lexical item in semantic structures. When no single ontology concept corresponds precisely to the lexical meaning, the ontological correlate may take the form of a semantic fragment composed of several ontology concepts and relations. For example:

  KOŠKA 'cat'
  Ontological correlate: *DomesticCat*
  KOTENOK 'kitten'
  Ontological correlate: *DomesticCat hasAttribute NotFullyFormed*

- **Government patterns** (subcategorization frames) describe the realization of syntactic actants in terms of lexical, morphosyntactic, and semantic constraints.
- **Lexical functions** are represented in accordance with the Meaning–Text Theory and describe a broad range of paradigmatic and syntagmatic lexical relations[1].
- **Lexicalized processing rules** make it possible to describe idiosyncratic properties of individual lexical items and systematic exceptions to more general rules in a uniform formalism.

As noted above, an important design objective of ETAP is the reusability of its lexicographic resources. The CD was therefore designed to support multiple applications and processing tasks. In addition to zones describing language-internal properties of lexical items, lexical entries contain dedicated multilingual zones specifying correspondences between lexical units in different languages.

For example, the EN (= English) zone of a Russian lexical entry contains information about its English equivalents. One field specifies the default single-word translation (or translations), while other fields contain context-dependent translation rules or references to such rules. Similar zones are available for several other languages incorporated into ETAP, including English, German, and Spanish.

This multilingual architecture allows a single lexicographic resource to be reused across different bilingual applications and tasks. Importantly, this principle extends beyond concrete applications. Monolingual semantic analysis itself may be viewed as a special type of translation—from a natural language into the language of semantic representations. Consequently, the same CD may support both multilingual processing and single-language semantic interpretation. Switching from one task to another merely requires changing the target language: English, German, Spanish, or the semantic representation language.

This architecture enables a single lexicographic resource to support multiple processing tasks while maintaining consistency across applications.

An example of a CD entry is given below. The entry represents the first sense of the Russian noun *UGROZA* and corresponds to the first sense of the English noun *THREAT* (as in *The letter was full of threats*). Some comments are given under double slash (//) below the line they are related to.

[1] The complete list of lexical functions with comments and examples (in Russian) can be found at https://ruscorpora.ru/page/instruction-lexical-functional/

```
ENTRY:УГРОЗА1
    COMMENT:"ЗАПУГИВАНИЕ"
    EXAMPLE:"УГРОЗА УВОЛЬНЕНИЕМ"
// comment and example to help the user to chose the correct sense of a polysemous word in
the interactive mode
  POR:S
// part of speech
  SYNT:ЖЕНСК,ИСЧИСЛ
// syntactic features
  DES:FCP-Action,Event,AbstractObject
// semantic features used for semantic checks
  SEM-TRANS:Threatening
// this field offers information underlying the match between the Russian word and its
counterparts in the Ontology
  GPATTERN
// government pattern
    1.1 ОТ;Agent
    1.2 СО$СТОРОНЫ1;Agent
    1.3 РОД,Agent
    2.1 ИНФ
    2.2 ЧТО1
    2.3 ТВОР
    3.1 ДАТ
  LEXFUNCS
// lexical functions
    _V0:УГРОЖАТЬ1
// threaten
    _ANTI:ОБЕЩАНИЕ
// promise
    _VER:РЕАЛЬНЫЙ/НЕПРИКРЫТЫЙ
// real/genuine/credible
    _ANTIVER:ПУСТОЙ
// empty/idle
    _REAL1-М:ВЫПОЛНЯТЬ
// carry out
// below are syntactic rules that are specific for the given word
  TRAF:КВАЗИАГЕНТ.11
  TRAF:1-КОМПЛ.20
  TRAF:2-КОМПЛ.21
  TRAF:1-КОМПЛ.14
  TRAF:1-КОМПЛ.15
  REG:1-КОМПЛ.D0
    N:1 //угроза[X] уничтожения[Y] городов со стороны[Z] Китая[Z1]
      CHECK
// applicability conditions of the rule
        1.1 R-EQU(X,Y,2,S,РОД,Event)&#(Y,Agent)
        1.2 R-LEXR(Y,Z,5,СО$СТОРОНЫ1,ОТ)&R-EQU(Z,Z1,1,РОД,Agent)
        1.3 COR(1,Y,Z1)
      DO
// operations performed when the conditions are satisfied
        1 LINK-NODES(X,Y,1-КОМПЛ)
  REG:INTERSYNT.D0
    N:1 //угроза[X] уничтожения городов со стороны[Z] Китая
      CHECK
        1.1 R-DOM-LEXR(X,Z,5,КВАЗИАГЕНТ,СО$СТОРОНЫ1,ОТ)
        1.2 DOM-EQU(Z,*,ПРЕДЛ,Agent)
        1.3 ^PININT(X,Z,PUN)
      DO
        1 SET-HYP-PRIOR(X,Z,КВАЗИАГЕНТ,STRONGHYP)

    ********************
  ZONE:EN
// English zone. Contains information needed for translating from Russian into English and
from English into Russian
    TRANS:THREAT1
    TRAF:EXPANS.92
      LR:OT
    TRAF:EXPANS.98
      LR:ЧТО1
    TRAF:EXPANS.97
      LR:ЧТО1
    ********************
  ZONE:SEM
// Semantic zone. Contains information needed for constructing the semantic structure. The
rules presented below specify how syntactic relations are mapped onto semantic relations in
the semantic structure.
    TRAF:SEM-CONV2.10
      T1:КВАЗИАГЕНТ,T2:hasAgent
    TRAF:SEM-CONV2.10
      T1:1-КОМПЛ,T2:hasTopic
    TRAF:SEM-CONV2.10
      T1:2-КОМПЛ,T2:hasRecipient
```

FIGURE 2 CD entry for UGROZA ‘threat’

3.3 Semantic resources

As noted above, the background knowledge used by ETAP is represented in two closely related knowledge resources: the Ontology and the inference rules. Together, these resources provide the information necessary for deriving inferences from semantic representations.

The Ontology plays a dual role in the system. On the one hand, it serves as a repository of structured knowledge about concepts representing classes of entities and their properties. On the other hand, it functions as the metalanguage of semantic description. Semantic structures produced by the analyser are built from ontological elements, and all meaningful linguistic units —lexical items, morphological markers, and syntactic constructions—must therefore be interpreted in terms of ontology concepts and relations.

To position our approach within the broader landscape, it is expedient to briefly compare it to one of the most well-known and influential semantic resources – FrameNet. It is a lexical database and annotated corpus based on Fillmore's frame semantics (Ruppenhofer et al., 2016). It represents meaning in terms of conceptual structures linking events and participants through semantic roles. In many respects, our ontology concepts resemble FrameNet frames. However, there are two important differences.

First, our concept descriptions contain substantially more information than a specification of semantic roles alone. In addition to participant structure, they encode a broad range of linguistic and background knowledge relevant to semantic interpretation.

Second, the Ontology is intended to support automatic inference. Concept descriptions are represented in a formal language suitable for computational processing and inference generation. By contrast, FrameNet descriptions are primarily intended for human interpretation and are not designed as a basis for large-scale automated reasoning. There have been some attempts to use FrameNet for reasoning (Dolbey et al., 2006; Scheffczyk et al., 2006; Ovchinnikova et al., 2010), but none of them aims to develop a reasoning system for general language.

The Ontology used in ETAP is based on the publicly available SUMO ontology (https://www.ontologyportal.org/), which has been substantially extended and adapted to the needs of semantic text processing. Since SUMO is a high-level ontology, many domain-specific concepts required for text understanding are not represented in it. Our modifications therefore include additional concepts, individuals, object properties, and data properties. As of May 2026, the Ontology contains 9,290 classes, 425 individuals, 317 object properties, and 42 data properties.

Concepts denote classes of entities. The description of a concept therefore characterizes all members of the corresponding class. Such members may include both subclasses and individual entities. For example, the concept *City* describes cities in general, whereas the individual *Moscow* inherits the properties of *City* and additionally contains information specific to this particular city. An important component of the Ontology is its repository of individuals, which contains descriptions of unique entities instantiating ontological classes. For example, an individual may instantiate the class Human while specifying values for such properties as name, nationality, profession, or date of birth.

Describing an ontology concept means specifying the relations that connect it to other ontological entities. The most fundamental relation is the class–subclass (*isA*) relation, which organizes concepts into a hierarchical network. Examples include *Human* → *Woman*, *SocialRole* → *Position*, and *Urging* → *Request*. At the top of the hierarchy stands the concept *Thing*, which is divided into two major subclasses: *Object* and *Event*. All remaining ontology concepts belong, directly or indirectly, to one of these two categories.

In addition to taxonomic relations, concept descriptions contain numerous properties represented by ontology slots. Some of these correspond to familiar semantic roles. For example, the concept *Giving* includes the relations *hasAgent*, *hasObject*, and *hasRecipient*, whereas *Gratitude* includes *hasExperiencer*, *hasObject*, and *hasStimulus*.

Other slots describe properties that are not directly related to argument structure. For instance, the concept *Human* includes such properties as *hasFirstName*, *hasFamilyName*, *hasBirthDate*, *hasBirthPlace*, and *hasNationality*. For each slot, the Ontology specifies the type of entity that may fill it. Thus, *hasBirthDate* requires a date, whereas *hasBirthPlace* requires a geographical entity.

A useful analogy can be drawn between ontology slots and semantic valencies. If a lexical item corresponds to a particular ontology concept, each semantic valency of that lexical item is associated with a slot of the corresponding concept. The converse, however, does not hold. Ontology slots need not correspond to semantic valencies. Thus, the semantic valencies of *give* correspond naturally to the slots of *Giving*, whereas properties such as *hasBirthDate* or *hasFirstName* do not correspond to semantic valencies of the noun *human*. Nevertheless, the analogy between semantic valencies and ontology slots is useful and helps clarify the relationship between lexical and ontological descriptions.

Some properties are represented by variable slot values, whereas others have fixed values. For example, different instances of *Human* may have different birth dates or nationalities, but all instances share the same value of certain properties. If a property such as *hasNumberOfLegs* were introduced for *Human*, its value would normally be fixed at two.

Concept descriptions may also contain more complex knowledge structures. For example, the concept *Helping* specifies that the beneficiary seeks to achieve a goal but experiences difficulty in doing so; the helper performs actions that facilitate the achievement of that goal; and these actions are evaluated as beneficial to the beneficiary.

An important feature of Ontology design is inheritance. Properties assigned to a concept are automatically inherited by its subclasses. Thus, concepts such as *Man*, *Woman*, *Boy*, and *Girl* inherit the relevant properties of *Human*. This makes it possible to construct compact concept descriptions containing only information that distinguishes a concept from its parent.

The range of information associated with ontology concepts is broad. For physical objects, concept descriptions may include information about constituent parts, function or purpose, material, typical location, origin, and participation in characteristic events or situations. The purpose of such descriptions is not to reproduce encyclopaedic knowledge in full, but to provide the background knowledge required for semantic interpretation and inference.

The Ontology by itself merely stores knowledge. The derivation of implicit information is performed by inference rules. These rules are written in the formal language Etalog (Rygaev, 2018), although for presentational purposes we describe them here informally.

From the viewpoint of reliability, ETAP distinguishes between two classes of inferred information: **strict inferences** and **plausible expectations**. Strict inferences are licensed by knowledge that guarantees the truth of the conclusion. Plausible expectations represent default assumptions about the most likely course of events and may subsequently prove to be false.

For example, if Peter forgot a telephone number, one can infer that he no longer knows that number, although he knew it previously. This is a strict inference. By contrast, if Peter was promised a bicycle, one may plausibly expect that the bicycle will eventually be bought. However, this expectation may fail to materialize. This is a plausible expectation.

Inference rules play a crucial role in everyday communication because much of the information extracted from a text is not stated explicitly. Texts typically provide only a limited set of explicit

facts, while listeners reconstruct additional information through inference. The following example illustrates the practical impact of the proposed resources on semantic analysis. The sentence is taken from a Russian football match report and is presented here in English translation:

> A corner at the home team's goal ended with a close-range shot by Necid. Dikan proved equal to the occasion.

The system must determine whether a goal event took place, although this is not explicitly stated. To do so, it consults several resources. The Combinatorial Dictionary links the expression *be equal to the occasion* to the ontology concept *EqualToOccasion*. The repository of individuals identifies Andrey Dikan as a goalkeeper. Inference rules provide two additional pieces of information. One defines *EqualToOccasion* as 'successfully doing what one is expected to do'. Another specifies that the function of a goalkeeper is to prevent the ball from entering their team's goal, whereas the ball entering the opponent's goal constitutes a goal event.

Combining these pieces of knowledge, the system concludes that Dikan successfully performed his role and prevented the ball from entering the goal. It therefore infers that no goal event occurred. Had the repository of individuals identified Dikan as a striker rather than a goalkeeper, the conclusion would have been the opposite: successful performance of a striker's role would imply that a goal was scored.

4 Conclusion

This paper has described the principal lexicographic and knowledge resources underlying the ETAP text processing system, with particular emphasis on its semantic component, SemETAP. The architecture of the system is based on the assumption that semantic analysis should not be limited to representing information explicitly expressed in a text. Rather, semantic processing should support the derivation of inferences based on both linguistic and background knowledge.

The resources employed by ETAP can be divided into two major groups. Linguistic knowledge is represented in the Morphological Dictionary, the Combinatorial Dictionary, and the grammatical rule sets. Background knowledge is represented in the Ontology, including its repository of individuals, and in the inference rules. Together, these resources support the construction of a Basic Semantic Structure (BSemS) and its subsequent enrichment, resulting in an Extended Semantic Structure (EnSemS) that incorporates inferred information.

The paper addressed three research questions. First, we examined the types of information that must be represented to support inference-based semantic analysis. The discussion showed that such processing requires the integration of diverse kinds of knowledge, including morphological, syntactic, semantic, ontological, and encyclopaedic information. Second, we outlined the formal languages used to represent these types of knowledge in different resources. Third, we demonstrated how linguistic and background knowledge interact during semantic analysis and jointly contribute to the derivation of inferred information.

Particular attention was paid to the Combinatorial Dictionary, which occupies a central position among the resources of ETAP. The dictionary serves not only as a repository of lexical information but also as an interface linking lexical, syntactic, semantic, ontological, multilingual, and procedural knowledge. This architecture makes it possible to reuse a single lexicographic resource across multiple processing tasks, including machine translation and semantic analysis.

The ETAP experience suggests that large-scale lexicographic and knowledge resources remain an essential component of knowledge-rich natural language processing. While contemporary language models provide powerful tools for many linguistic tasks, explicit lexical and ontological knowledge continues to offer important advantages in transparency, controllability, explainability, and inference generation. These properties make such resources particularly valuable for applications requiring transparent semantic interpretation and knowledge-based reasoning.

References

Apresjan, Ju. (2000). *Systematic Lexicography*. Oxford University Press.

Dolbey, A., Ellsworth, M., and Scheffczyk, J. (2006). *BioFrameNet: A Domain-specific FrameNet Extension with Links to Biomedical Ontologies*. KR-MED 2006 "Biomedical Ontology in Action" November 8, 2006, Baltimore, Maryland, USA.

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, 11. https://doi.org/10.1057/s41599-024-02889-7

Mel'čuk, I. (1974). *Opyt teorii lingvisticheskix modelej «Smysl ↔ Text»* [The theory of linguistic models of the Meaning ↔ Text type]. Nauka.

Mel'čuk, I. (2012). *Jazyk: ot smysla k tekstu* [Language: from Meaning to Text]. Jazyki slavjanskoj kultury.

Ovchinnikova, E., Vieu, L., Oltramari, A., Borgo, S., Alexandrov, T. (2010). Data-Driven and Ontological Analysis of FrameNet for Natural Language Reasoning. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

Rundell, M. (2004). Automating The Creation of Dictionaries: Are We Nearly there? *Humanising Language Teaching Magazine*, 26(1).

Ruppenhofer, J., Ellsworth, M., Petruck, M., Johnson, Ch., Baker, C. & Scheffczyk, J. (2016). *FrameNet II: Extended Theory and Practice*. International Computer Science Institute (ICSI).

Rygaev, I. (2018). Etalog — a natural-looking knowledge representation formalism. In *ITiS 2018: Sbornik trudov 42-oi mezhdistsiplinarnoi shkoly-konferentsii IPPI RAN «Informatsionnye tekhnologii i sistemy 2018»* [ITiS 2018: Proc. of the 42nd Conf. "Information technologies and systems 2018"] (pp. 473–482). A.A.Kharkevich Institute for Information Transmission Problems.

Scheffczyk, J., Baker, C., Narayanan, S. (2006). Ontology-based Reasoning about Lexical Resources. In Proc. of OntoLex 2006: Interfacing Ontologies and Lexical Resources for Semantic Web Technologies (pp. 1-8).

de Schryver, G.-M. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography*, 36(4), 355–387. https://doi.org/10.1093/ijl/ecad021

Zhao, H., Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, Mengnan Du. (2023). *Explainability for Large Language Models: A Survey*. Retrieved May 25, 2026, from arXiv:2309.01029v3.

Contact information

Igor Boguslavsky
A.A.Kharkevich Institute for Information Transmission Problems; Universidad Politécnica de Madrid
Igor.M.Boguslavsky@gmail.com

Vyacheslav Dikonov
A.A.Kharkevich Institute for Information Transmission Problems
sdiconov@mail.ru

Tatyana I. Frolova
A. A. Kharkevich Institute for Information Transmission Problems
tfrolova@gmail.com

Evgeniya Inshakova
A.A.Kharkevich Institute for Information Transmission Problems
e.s.inshakova@gmail.com

Alexandre V. Lazursky
A. A. Kharkevich Institute for Information Transmission Problems
lazursky@mail.ru

Svetlana P. Timoshenko
A. A. Kharkevich Institute for Information Transmission Problems
timoshenko@iitp.ru

Acknowledgements

This work was supported by the RSF grant No. 24-18-00988, which is gratefully acknowledged.

Li Cui, Seol Namkung, Jun Lee, Hae-Yun Jung & Kilim Nam

WHAT LLMS LEARN AND WHAT THEY MISS

Frequency Effects in the Automated Description of Korean Verbal Inflection

Abstract The present study aims to empirically examine the possibilities and limitations of LLMs as a lexicographic tool for Korean, by automatically generating inflected forms of Korean verb and adjective headwords using Large Language Models and contrasting the results with corpus frequency data. Focusing on the description of inflected forms of Korean verbs and adjectives, the study analyses in detail the extent to which LLMs reproduce corpus-based frequency rankings and the accuracy of inflected form generation, divided into the categories of regular, irregular, and defective predicates. The results show that LLMs consistently produce various errors in low-frequency inflected forms and reveal clear limitations in precisely reproducing frequency information. This empirical examination points to the possibilities and risks inherent in LLMs as lexicographic assistance tools from multiple angles. While LLMs hold meaningful value as an assistive means for drafting dictionary descriptions, lexicographic reliability can only be guaranteed when this is complemented by systematic corpus-based verification and the distinctive expertise of lexicographers. These findings reaffirm the fundamental role that lexicographers must play in the automation of dictionary compilation.

Keywords LLM-assisted lexicography; low-resource languages; Korean verbal inflection; lexicographic automation; long-tail problem; low-frequency words; agglutinative language

1 Introduction

From a typological perspective, Korean is an agglutinative language in which functional morphemes such as postpositions, endings, and affixes are sequentially attached to the stems of content words to realize a range of semantic and grammatical functions, including tense, aspect, modality, and clause connection. In Korean, adjectives are characterised as *verb-like* adjectives (Dixon & Aikhenvald, 2004) or *verby* adjectives (Stassen, 1997), and both verbs and adjectives function as independent predicates and share the same inflectional system, forming the part-of-speech category of *yongen*[1] 'predicates'. Therefore, the way inflectional patterns of individual predicates (i.e., verbs and adjectives, whether regular and irregular) are presented in Korean dictionaries is crucial.[2] However, the lexicographic coverage of inflected forms remains limited. Major dictionaries tend to present only a small number of fixed ending combinations,[3] reflecting the legacy of print lexicography, where space constraints limited the presentation of inflected forms to around three to five per entry. This practice,

1 This paper follows the Yale Romanization system.

2 Although Korean verbs and adjectives are morphologically similar, they show different inflectional patterns in the present tense and with certain verb endings. While verbs and adjectives are semantically distinguished as categories expressing action/process and state/property respectively, the differences in their allomorphic patterns constitute a frequent source of errors for learners of Korean.

3 The *Korean Basic Dictionary*, which is designed for learners of Korean, presents four fixed ending combinations (adnominal ending, sequential connective ending, causal connective ending, and honorific sentence-final ending). The *Standard Korean Language Dictionary* does not appear to provide explicit guidelines for the selection and presentation of inflected forms and presents only two types of connective endings. A similar approach is adopted by the *Korea University Korean Language Dictionary*. As for the *Yonsei Korean Dictionary*, it presents three forms: adnominal ending, connective ending, and honorific sentence-final ending.

however, does not adequately reflect the range of endings with which individual predicates combine in actual language use, nor the frequency of such combinations. For example, the high-frequency Korean verb 되다 *toyta* 'become' can theoretically produce over 30,000 inflected forms,[4] although only 21 forms are needed to cover the top 80% of cumulative frequency in actual usage. Example (1) illustrates the discrepancy between actual usage patterns (1a) and their lexicographic representation (1b).

(1) a. 되다 *toyta* 'become': 2,104 attested inflected forms; e.g. 된다 *toynta* [present declarative finite form] (10.9%); 될 *toyl* [future/prospective adnominal form] (10.3%), 되는 *toynun* [present adnominal] (9.8%), 되어 *toye* [connective form] (6.6%), 된 *toyn* [past/perfective adnominal form] (6.6%) etc.
b. Entry for 되다 *toyta* 'become' in the *Standard Korean Language Dictionary*: 되어 (돼) *toye (toay)*, 되니 *toyni*

From a corpus-based lexicographic perspective, the limitations of existing dictionaries become even more evident. Korean predicates exhibit a long-tail distribution, consisting of a small number of high-frequency lexical items and a large number of low-frequency items, and the distribution of endings that combine with individual predicates is highly uneven. Therefore, the lexicographic description of inflected forms should move beyond the presentation of a fixed set of representative endings and reflect the actual usage frequency and combinatorial constraints of individual predicates.

Recent advances in LLM performance have led a growing interest in the automatic generation of various elements of the dictionary microstructure (de Schryver, 2023; Rundell, 2023; Lew, 2023, 2024). Such studies demonstrate that LLMs can be assist dictionary compilation in many ways, from drafting definitions to generating examples and describing senses. However, research on LLM-based lexicography has consistently emphasized the necessity of expert human review alongside the usefulness of automatically generated content (Jakubíček et al., 2023; Klosa-Kückelhaus & Tiberius, 2025; Widmann, 2025). Furthermore, because existing discussions have focused primarily on English, the extent to which LLMs can function reliably as a lexicographic assistance tool in languages with developed agglutinative morphological structure, such as Korean, has not been sufficiently examined.

This issue is also related to the way LLMs process language. Subword-based tokenization may not adequately reflect language-specific morpheme boundaries, and LLMs have been reported to show weaknesses in learning and reproducing long-tail knowledge compared to high-frequency knowledge (Rust et al., 2021; Kandpal et al., 2023). It is therefore necessary to separately evaluate LLM performance in the description of Korean inflected forms, which include low-frequency predicates, irregular inflection, and defective predicates. In that sense, the present study analyses the extent to which LLMs can accurately generate and rank Korean predicate inflected forms according to corpus frequency ranges. In doing so, the study aims to shed light on both the potential contribution of LLM-assisted lexicographic compilation to the description of Korean predicate inflected forms and the limitations that emerge in the treatment of low-frequency, irregular, and defective predicates.[5]

[4] According to Choi (2012), there are eight realisations of prefinal endings expressing tense, aspect, and honorifics in Korean, and over 4,000 types of final endings functioning as connective, sentence-final, transformative, and -honorific endings; as such, over 30,000 inflected predicate forms can theoretically be generated.

[5] Similar limitations in the morphological processing of LLMs have also been reported in Korean-related research. Park et al. (2024) pointed out that ChatGPT-based models achieved a certain level of performance in Korean grammatical error correction but remained less effective in handling fine-grained morphological and grammatical errors, and Shin et al. (2025) reported relatively low performance in the domain of morphology when assessing advanced Korean language proficiency.

The study addresses the following research questions.

1) Can LLMs accurately generate inflected forms of Korean verbs and adjectives by combining a given list of endings with headwords, and rank them in an order close to their corpus frequencies?

2) For defective predicates that only combine with a restricted set of endings, can LLMs independently generate permissible endings and inflected forms, and rank them in an order close to their corpus frequencies?

3) From the perspective of automating Korean dictionary description, what are the potential strengths and limitations of LLMs in complementing corpus data and native speaker intuition? What tendencies does the adequacy of LLM outputs show in relation to predicate frequency, regularity, and defectiveness?

To answer these questions, the study evaluates the ability of LLMs to combine regular and irregular predicate headwords with a given set of endings and rank the resulting forms by frequency. As for defective predicates, it assesses the ability to generate permissible endings and inflected forms without being given a set of endings.

2 Research data and methodology

2.1 Data

For a precise analysis of lexical and inflectional patterns across a wide range of registers, the present study uses the *Spoken Corpus* (approximately 100 million *ecel* 'Korean word units') and the *Written Corpus* (approximately 300 million *ecel*) compiled by the National Institute of Korean Language. The *Spoken Corpus* consists of 4,130 texts representing spoken and quasi-spoken registers, including TV broadcast dialogues and drama scripts, whereas the *Written Corpus* comprises 7,077 texts drawn from written registers such as newspapers, magazines, fiction, and non-fiction books.

After performing a morphological analysis of these corpora,[6] 77,493 verbs and adjectives were initially extracted from approximately 2.12 million identified morphemes. To distinguish homonymous predicates whose inflected forms differ according to meaning, such as 묻다 *mwutta* 'ask' or 'bury', the extracted forms were matched against the 25,138 predicates listed in *Urimalsaem*, a user-participatory dictionary and the largest Korean dictionary, to disambiguate headwords. Of the 77,493 predicates identified, only 33 account for the top 60% of cumulative frequency, while 72,279 are concentrated in the 95-100% frequency range. Furthermore, as the frequency range decreases, the proportion of -여 *-ye* irregular predicates, including the -하다 *-hata* derivational predicates, increases. To prevent overrepresentation in the sample, the proportion of *-hata* derivational predicates was therefore capped at 20%. Headwords were selected at a ratio of 1:2:4 for the high-, medium-, and low-frequency ranges respectively, taking into account the lexical distribution; however, as the high-frequency range contained fewer items than the target sample size, all predicates in that range were included in the analysis. Defective predicates were defined as predicates with fewer than 10 inflected form types, and a total of 113 headwords were selected based on this criterion. The distribution of the selected headwords by predicate type is presented in Table 1.

[6] Morphological analysis was conducted using the Kiwi morphological analyser.

TABLE 1 Number of selected headwords by predicate type

Range	Cumulative range	Regular predicates	Irregular predicates	Defective predicates
High-frequency (top 60%)	~ top 60%	22	11	-
Medium-frequency (60-95%)	top 70% range	40	40	21
	top 80% range	40	40	
Low-frequency (95-100%)	top 96% range	80	80	92
	top 98% range	80	80	
Total		262	251	113

2.2 Methodology

This study uses GPT-5.4-mini and CLAUDE-sonnet-4.6 to evaluate the ability of LLMs to process Korean predicate inflection across two dimensions: corpus-based frequency reproduction and inflected form generation accuracy. For regular and irregular predicates, the top 10 endings attested in the corpus for each headword were presented to the models in random order, and the models were instructed to rank them according to their expected usage frequency and generate the inflected form resulting by combining each ending with the headword. For defective predicates, no list of endings was provided, and the models were instructed to independently generate the permissible endings and inflected forms for each defective predicate.[7]

Corpus-based frequency reproduction was evaluated using rank deviation, which was calculated as the difference between the ranking produced by the model and the corpus-based ranking. Deviations were categorised into three ranges: ±0-1, ±2-4, and ±5-9. This measure was intended to assess how closely the model approximated corpus-based rankings, rather than whether the rankings matched exactly.

Inflected form generation accuracy was evaluated based on the morphological adequacy of the generated forms. To this end, inflectional and lexical the error rates were calculated, and error instances were classified by type. The presence and type of errors were determined manually by the researchers, and cross-checking was also conducted to ensure consistency in the evaluation.

3 Frequency reproduction and generation of inflected forms for regular and irregular predicates

3.1 Results of frequency ranking reproduction

This section presents the ranking results for regular and irregular predicates and examines the extent to which LLMs can reproduce corpus-based inflected form frequency. Table 2 summarises the distribution of ranking errors for regular and irregular predicates across frequency range and models.

[7] The prompts are available at the following link: https://drive.google.com/drive/folders/1-J0Z0Zao7wpxzkYjlqhEpOFz9HR5XFbW?usp=drive_link

TABLE 2 Rank error rate by interval (%) (regular and irregular predicates)

Category	Range	±0-1		±2-4		±5-9	
		GPT	Claude	GPT	Claude	GPT	Claude
Regular predicates	High-frequency	33.6	37.7	45	45.5	20.9	16.8
	Medium-frequency	34.8	38	44.2	42.4	21	19.4
	Low-frequency	32.7	36.6	46.4	46.6	20.9	16.2
Irregular predicates	High-frequency	35.5	30	36.4	49.1	27.3	17.3
	Medium-frequency	36.4	38.8	44.4	44.5	18.8	16.5
	Low-frequency	37.9	38	45.7	45.3	16.2	15.8

As shown in Table 2, the distribution patterns of rank deviations were similar across all frequency ranges, with no marked differences between ranges. For both regular and irregular predicates, the highest proportion of cases fell within the ±2-4 range, suggesting that the rankings produced by the LLMs are not entirely random. However, the proportion of highly accurate rankings within the ±0-1 range remained at approximately 30-38%, while substantial errors within the ±5-9 range were also consistently observed. These results indicate that while LLMs are able to identify inflected form candidates to some extent, they show limitations in precisely reproducing the actual frequency distribution of individual predicates.

3.2 *Results of inflected form generation*

This section examines the inflected form generation results for regular and irregular predicates, to assess whether LLMs achieve a level of morphological accuracy sufficient for use in the lexicographic description of Korean inflection. The results for regular predicates are presented in Table 3 and for and irregular predicates in Table 4.

TABLE 3 Analysis of inflected form generation errors by frequency range for regular predicates

Category	High-frequency		Medium-frequency		Low-frequency	
	GPT	Claude	GPT	Claude	GPT	Claude
Number of predicates (n)	22		80		160	
Number of inflected forms	220		800		160	
Number of inflected form errors and rate (%)	2 (0.9%)	1 (0.5%)	5 (0.6%)	1 (0.1%)	72 (4.5%)	22 (1.4%)
Number of lexical errors and rate (%)	2 (9.1%)	1 (4.5%)	4 (5.0%)	1 (1.2%)	29 (18.1%)	13 (8.1%)

Overall, the error rate for regular predicates was lower than that for irregular predicates. In the high- and medium-frequency ranges, inflected form error rates were as low as 0.1-0.9%; in the low-frequency range, however, the error rate increased markedly to 4.5% for GPT and 1.4% for Claude. At the lexical level, one or more errors were found in 29 out of 160 headwords (18.1%) for GPT and 13 headwords (8.1%) for Claude. This suggests that even in the case of regular inflection, inflected form generation is not entirely stable for low-frequency lexical items.

The errors observed in regular predicates fall into four main types, as shown in (2).

(2) **a. Non-application of identical vowel deletion rule:** (GPT) 특출나-+-아 → *특출나아 *thukchwulnaa* [correct: 특출나 *thukchwulna* 'be/being outstanding']

b. Confusion with irregular inflection: (GPT) 간드러지-+-ㄴ → 간드러운 *kantulewun* [correct: 간드러진 *kantulecin* 'graceful']; (GPT) 휘몰-+-으라 → 휘몰라 *hwimolla* [correct: 휘몰아 *hwimola* 'sweep along']

c. Stem/ending recognition errors:

- **Stem recognition errors:** (Claude) 쭐-+-으니까 → 쫙니까 *ccyoknikka* [correct: 쭐으니까 *ccwulunikka* 'because (it) shrinks']; (GPT) 잡숫-+-다 → 잡숟다 *capswutta* [correct: 잡숫다 *capswusta* 'to eat'
- **Ending recognition errors:** (Claude) 비탈지-+-ㄴ대요 → 비탈지대요 *pithalcitayyo* [correct: 비탈진대요 *pithalcintayyo* '(they) say it is steep']

d. Allomorph selection errors: (GPT) 메기-+-ㅁ → 메기음 meki.um [correct: 메김 mekim 'assigning; assignment']; (GPT) 가냘프-+-아 → 가냘퍼 *kanyalphe* [correct form: 가냘파 *kanyalpha* 'is slender and delicate']

The identical vowel deletion rule is a phonological process whereby one vowel is deleted when the ending -아/어 *-a/e* is attached to a vowel-final stem, a rule LLMs did not consistently apply, as illustrated in (2a). Example (2b) shows instances of the overapplication of irregular inflection patterns to regular predicates. In *휘몰라 *hwimolla*, the model incorrectly inserted an additional ㄹ (romanized as *l*), producing a pattern associated with the irregular inflection of predicates whose stems end in 르 (*lu*), whereas in *간드러운 *kantulewun*, it incorrectly introduced the surface form 우 (*wu*), which is characteristic of predicates with stem-final ㅂ (*p*). Example (2c) illustrates errors in the recognition of stem or ending boundaries:[8] In *잡숟다 *capswutta*, the stem 잡숫- *capswus-* was not correctly produced, while in *비탈지대요 *pithalcitayyo*, the ending provided in the prompt does not match the ending actually produced by the model. Example (2d) reflects unstable allomorph selection, including the choice between the nominalising ending allomorphs -ㅁ *-m* or -음 *-um* conditioned by stem-final phonology, and the choice between the ending allomorphs -아 *-a* or -어 *-e* conditioned by vowel harmony.

TABLE 4 Analysis of inflected form generation errors by frequency range for irregular predicates

Category	High-frequency		Medium-frequency		Low-frequency	
	GPT	Claude	GPT	Claude	GPT	Claude
Number of predicates (n)	11		80		160	
Number of inflected forms	110		800		1,600	
Number of inflected form errors and rates (%)	0 (0.0%)	0 (0.0%)	14 (1.8%)	3 (0.4%)	129 (8.1%)	52 (3.2%)
Number of lexical errors and rates (%)	0 (0.0%)	0 (0.0%)	7 (8.8%)	2 (2.5%)	65 (40.6%)	44 (27.5%)

[8] However, in a small number of cases, the surface form did not provide sufficient evidence to determine whether the error originated in stem recognition or ending recognition.

In the high-frequency range, both GPT and Claude achieved an inflected form error rate of 0.0%. Errors began to appear in the medium-frequency range and increased sharply in the low-frequency range, reaching 8.1% for GPT and 3.2% for Claude. This tendency is even more pronounced in the lexical error rate: in the low-frequency range, one or more errors were found in 65 out of 160 headwords (40.6%) generated by GPT and 44 headwords (27.5%) generated by Claude. Such results indicate that the models experience considerable difficulty in producing correct surface forms for irregular predicates, even when the ending to be combined is provided as input.

The errors observed in irregular predicates can be grouped into four main categories, as shown in (3).

(3) **a. Confusion with regular inflection**: (GPT) 너르- + -어 → *너르어 *nelue* [correct: 널러 *nelle* 'be/being spacious]; (GPT) 부자연스럽- + -은 → *부자연스럽은 *pucayensulepun* [correct: 부자연스러운 *pucayensulewun* 'unnatural']

b. Confusion with a different irregular inflection pattern: (GPT) 똥그랗- + -아요 → *똥그라워요 *ttongkulaweyo* [correct: 똥그래요 *ttongkulayyo* 'is round']; (Claude) 누르- + -러 → *눌러 *nwulle* [correct: 누르러 *nwulule* 'be/being yellowish']

c. Stem/ending recognition errors:

- **Stem recognition errors:** (GPT) 분지르-+-는 → *분지는 *puncinun* [correct: 분지르는 *puncilunun* 'breaking']
- **Ending recognition errors:** (GPT) 천진스럽- + -고 → *천진스러워서 *chencinsulewese* [correct: 천진스럽고 *chencinsulepko* 'and (is) innocent']; (Claude) 경망스럽-+-(으)ㄴ → *경망스러움 *kyengmangsulewum* [correct: 경망스러운 *kyengmangsulewun* 'frivolous']

d. Allomorph selection errors: (Claude) 퍼렇-+-었다 → *퍼랬다 *phelayssta* [correct: 퍼렜다 *pheleysta* 'was blue']; (GPT) 한숨짓-+-을 → *한숨질 *hanswumcil* [correct: 한숨지을 *hanswumciul* 'will sigh']

In Example (3a), the regular stem-ending combination was applied in contexts where irregular inflection should have occurred. Example (3b) involves confusion between irregular types. In *똥그랗다 *ttongkulahta*, the model treated the irregular predicate with stem-final ㅎ (*h*) as a predicate with stem-final ㅂ (*p*). As for *눌러 *nwulle*, the model confused two homonymous irregular predicates, generating the irregular inflected form 눌러 *nwulle* of the predicate 누르다 *nwuleta* 'to press' instead of the irregular inflected form 누르러 *nwulule* of the homonym 누르다 *nwuleta* 'to be yellowish'. Example (3c) illustrates incorrect recognition of stem, such as the omission of the stem-final syllable 르 in *분지는 *pwuncinun*, as well as ending recognition errors, as in *천진스러워서 *chencinsulewoese* where the model produced the wrong ending (-어서 *-ese* 'because' instead of -고 *-ko* 'and'), and *경망스러움 *kyengmangsulewum* where the nominalizing ending -ㅁ *-m* was produced instead of the adnominal ending -ㄴ *-n*. Example (3d) reflects unstable stem alternation and allomorph selection in environment involving the irregular stem-final consonants ㅎ (*h*) and ㅅ (*s*).[9]

[9] In 한숨질 *hanswumcil*, the model correctly applied the stem-final 〈ㅅ〉 required by this type of irregular inflection. However, instead of generating the expected form 한숨지을 *hamswumciul*, it selected the adnominal ending -ㄹ *-l* in place of the allomorph -을 *-ul*, and was therefore classified as an allomorph-selection error.

While LLMs demonstrate a certain potential for automating the description of inflected forms of regular and irregular Korean predicates, they still show clear limitations when dealing with low-frequency and irregular predicates. In particular, the generation of inflected forms was relatively stable for high-frequency regular predicates, whereas errors increased in the low-frequency range, and the models showed limited ability to accurately reproduce frequency order of inflected forms. These findings suggest that while LLM outputs can be useful for drafting dictionary descriptions, corpus-based verification and the morphological expertise and judgement of lexicographers remain essential for the outputs to be applicable to actual dictionary compilation.

4 Frequency reproduction and generation of inflected forms for defective predicates

4.1. *Results of frequency ranking reproduction*

Defective predicates have a limited number of permissible endings, and a particular ending often accounts for a substantial proportion of overall usage. Therefore, even when a model generates endings attested in the corpus, it is also necessary to examine whether those endings occupy positions in the inflected form distribution similar to their corpus frequency rankings. Accordingly, the study analysed the extent to which the endings generated by the models reproduce the corpus-based distribution of ending frequencies. Ranking errors for defective predicates are presented in Table 5.

TABLE 5 Rank error rate by frequency range for defective predicates (%)

Category	Range	±0-1		±2-4		±5-9	
		GPT	Claude	GPT	Claude	GPT	Claude
Defective predicates	Medium-frequency	31.5	48.4	38.9	42.5	29.6	9
	Low-frequency	44.4	58	38.5	32.5	17.1	9.5

For both models, a substantial proportion of generated endings fell within the ±0-1 range, indicating that the models (especially Claude) were able to approximate the corpus-based rankings of endings for defective predicates to some extent. However, the proportion observed in the ±2-4 range is also relatively high, making it difficult to conclude that the rankings produced by the LLMs consistently align with the corpus-based rankings. This tendency was particularly pronounced for GPT, as 29.6% of endings it generated for medium-frequency predicates fell within the ±5–9 range, indicating significant deviations from actual usage rankings in a considerable number of cases. These results show that, while LLMs can generate some of the endings actually attested for defective predicates and approximate their corpus-based rankings, they cannot be said to reliably reproduce corpus-based frequency distributions.

4.2 *Results of inflected form generation*

This section examines the generation results for defective predicates, which are summarized in Table 6.

TABLE 6 Inflected form generation results for defective predicates

Category	Medium-frequency		Low-frequency	
	GPT	Claude	GPT	Claude
Number of predicates (n)	21		92	
Number of inflected forms	199		918	
Inflected form errors	40	31	238	145
Inflected form error rate (%)	20.1	15.6	25.9	15.8
Number of lexical errors	12	12	59	40
Lexical error rate (%)	57.1	57.1	64.1	43.5

The generation of inflected forms for defective predicates consistently showed high error rates regardless of frequency. Compared with the rates of below 10% observed for regular and irregular predicates, those for defective predicates were up to approximately ten times higher. This contrast becomes even more apparent when lexical error rates are considered. The lexical error rate refers to the proportion of predicates for which at least one generation error occurred in the inflected forms presented. Even in the medium-frequency range, both GPT and Claude recorded lexical error rates of 57.1%. This indicates that generation errors in defective predicates are not confined to a small number of inflected forms, but rather recur across the paradigm of a given headword. As such, defective predicates constitute a substantially more challenging generation task for LLMs than regular or irregular predicates, suggesting that the models are not yet able to handle the combination patterns of Korean predicates in a reliable manner.

The error types for defective predicates are classified in Table 7.

TABLE 7 Classification of error types for defective predicates

Error type	GPT (proportion)	Claude (proportion)	Examples
MWE generation	84 (30.2%)	138 (78.4%)	(GPT) 걷잡을 수 없다 *kethcapul swu epta* 'to be out of control' (Claude) 귀담지 않다 *kwitamci anhta* 'to not pay attention'
Stem/ending recognition	109(39.2%)	22(12.5%)	(GPT) 노르스레해도 *nolusuleyhayto* [target: 노릇하- + -여도 → 노릇하여도 *nolusleyhayto* 'even if one acts as']
Allomorph selection	51(18.3%)	16(9.1%)	(GPT) 열띠아 *yelttia* [target: 열띠- + -어 → 열띠어 *yelttie* 'be enthusiastic/fervent']
Generation of agrammatical inflected forms	34(12.2%)	0(0.0%)	(GPT) *부둥켰- *putwungkhyess* (부둥키- *putongkhi* 'hug' + -었- *-ess-* (past tense marker)) (GPT) *머나멀어라 *menamelela* (머나멀- *menamele* 'be very far' + -어라 *-ela* (imperative ending))
Total	278(100%)	176(100%)	

The error analysis revealed that defective predicates often gave rise to outputs falling outside the scope of simple morphological errors. From a lexicographic perspective, the most significant error type was MWE generation. 84 instances were identified for GPT and 138 for Claude,

which demonstrates that the models tend to generate conventionalised multi-word constructions (e.g., predicate + auxiliary predicate) that frequently occur in the corpus, rather than treating the restricted inflectional patterns of defective predicates as properties of the predicate itself. Examples such as 겉잡을 수 없다 *kethcapul swu epta* and 귀담지 않다 *kwitamci anhta* are not simple stem-ending combinations but phrasal constructions in which the predicate is combined with a dependent noun (수 *swu* 'ability') and an auxiliary predicate (않다 *anhta* 'do not') respectively. Although the LLMs could identify high-frequency expression units in actual usage, they often fail to distinguish and organize inflected forms, collocations, and phrasal constructions within the microstructure of a dictionary entry.

Another characteristic error type was the generation of agrammatical inflected forms. For example, *부둥켰- *pwutungkhyes-* combines the predicate stem with only the prefinal ending -었- *-ess-* but no sentence-final ending, and *머나멀어라 *menameleula* combines the adjective stem with an imperative ending, which is not permitted in Korean grammar. These errors suggest that the limitations of the models are deeper than the mere generation of individual infected forms and extend to the grammatical conditions governing the completion of Korean predicate inflections. In other words, while LLMs exhibits a high degree of sensitivity to formal patterns in their subword-based processing, they struggle to evaluate semantic completeness, and this leads to errors in low-frequency defective predicates.

The errors in defective predicates can be seen as occurring on two levels. The first consists of morphological errors, such as deletion, contraction, and allomorph-selection errors, in which the model fails to generate the correct surface form; the second consists of syntactic errors, including auxiliary predicate combinations and MWEs, which are beyond the scope of inflected forms. Meanwhile, the phrasal constructions generated by LLMs suggest that these predicates might be used in actual language not as individual lexical units but as part of larger meaning units, which is in line with Sinclair (1996), who argued that meaning emerges from recurrent patterns of co-occurrence rather than being fixed in individual items. As such, these outputs, which constitute errors from the perspective of inflectional description, may nevertheless provide supplementary evidence for lexicographic analysis and compilation.

5 Discussion: How viable is the LLM-assisted lexicographic description of Korean predicate inflection?

This study examined the extent to which LLMs can reproduce corpus-based frequency rankings and accurately generate inflected forms of Korean verbs and adjectives. To this end, regular, irregular, and defective predicates were analysed separately. The results show that LLMs performed relatively reliably in generating inflected forms of high-frequency regular predicates. However, errors involving stem/ending recognition, allomorph selection, and confusion between regular and irregular inflection increased for low-frequency and irregular predicates. For defective predicates, the models generated auxiliary predicate constructions or phrasal expressions in place of simple inflected forms, as well as agrammatical forms. Furthermore, the models show limitations in accurately reproducing corpus-based frequency rankings across all three predicate types.

When the morphological description that lies at the core of agglutinative language lexicography is automated through LLMs, a separate and rigorous verification procedure proves necessary. Although individual headwords differ in the endings they most frequently combine, dictionaries typically present inflectional information according to a standardised framework, thereby failing to fully reflect actual language use patterns. The present study has shown that

LLMs also struggle to capture dynamic frequency patterns; as such, future dictionary description should move beyond the simple listing of inflected forms and instead selectively incorporates inflectional information that is meaningfully attested in actual corpora.

From the perspective of lexicographic automation, LLMs show potential as an assistive tool for generating and organising inflected form candidates. However, in order to ensure the accuracy of inflected form generation and the authenticity of frequency ranking reproduction, the principles proposed by Kilgarriff & Rychlý (2010) for semi-automatic lexicographic compilation remains valid for LLM-assisted lexicography today. Their model, which combines computer-assisted processing with final human verification, also applies to the task of describing the morphological and syntactic constraints and actual usage of Korean with the assistance of LLMs.

While LLMs demonstrated a certain potential as a draft generation tool in the description of Korean predicate inflected forms, they revealed structural weaknesses in the long-tail domain. This points to the need for a new lexicographic methodology capable of simultaneously overcoming both the fixed descriptive approach of existing dictionaries and the generative limitations of LLMs. Ultimately, the core challenge of LLM-assisted lexicography lies not so much in automation itself, but in how to organically combine LLM outputs, corpus data, and the specialised linguistic intuition of lexicographers in order to accurately represent actual language use in the dictionary.

References

Choi, S.-J. (2012). The study of combined endings' condition. *Journal of Korean Linguistics*, 63, 275–311. https://doi.org/

De Schryver, G.-M. (2023). Generative AI and lexicography: The current state of the art using ChatGPT. *International Journal of Lexicography*, 36(4), 355–387. https://doi.org/10.1093/ijl/ecad021

Dixon, R. M. W. & Aikhenvald, A. Y. (Eds.) (2004). *Adjective Classes: A Cross-Linguistic Typology*. Oxford University Press.

Jakubíček, M., Rundell, M., Medveď, M., Měchura, M., Tiberius, C., Kosem, I. & Kallas, J. (2023). The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography? In M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček & S. Krek (Eds.), *Electronic Lexicography in the 21st Century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference*. (pp. 518–533). Brno: Lexical Computing CZ s.r.o.

Kandpal, N., Deng, H., Roberts, A., Wallace, E. & Raffel, C. (2023). Large language models struggle to learn long-tail knowledge. In A. Krause, E. Brunskill, K. Cho, B. Engelhardt, S. Sabato & J. Scarlett (Eds.), *Proceedings of the 40th International Conference on Machine Learning* (pp. 15696–15707). PMLR.

Kilgarriff, A. & Rychlý, P. (2010). Semi-automatic dictionary drafting. In G.-M. de Schryver (Ed.), *A Way with Words: Recent Advances in Lexical Theory and Analysis. A Festschrift for Patrick Hanks* (pp. 299–312). Menha Publishers.

Klosa-Kückelhaus, A. & Tiberius, C. (2025). The lexicographic process revisited. *International Journal of Lexicography, 38*(1), 1–12. https://doi.org/10.1093/ijl/ecae016

Lew, R. (2023). ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences Communications*, 10(1), 704. https://doi.org/10.1057/s41599-023-02119-6

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, 11(1), 426. https://doi.org/10.1057/s41599-024-02889-7

Park, C., Koo, S., Kim, G. & Lim, H. (2024). Towards harnessing the most of ChatGPT for Korean grammatical error correction. *Applied Sciences, 14*(8), 3195. https://doi.org/10.3390/app14083195

Rundell, M. (2023). Automating the creation of dictionaries: Are we nearly there? *In Proceedings of ASIALEX 2023* (pp. 9–17). Seoul: Yonsei University.

Shin, H., Lee, S., Jang, D., Song, W., Kim, J., Oh, C., Jo, H., Ahn, Y., Oh, S., Chang, H., Kim, S. & Lee, J. (2025). *KoBALT: Korean benchmark for advanced linguistic tasks*. arXiv. https://doi.org/10.48550/arXiv.2505.16125

Sinclair, J. M. (1996). The search for units of meaning. *Textus*, 9(1), 75–106.

Stassen, L. (1997). *Intransitive Predication*. Oxford University Press.

Widmann, T. (2025). A pipeline for automated dictionary creation with optional human intervention. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda & A. Salgado (Eds.), *Electronic Lexicography in the 21st Century (eLex 2025): Intelligent Lexicography. Proceedings of the eLex 2023 conference*. (pp. 33–43). Brno: Lexical Computing CZ s.r.o.

Dictionaries

National Institute of Korean Language (1999). *Standard Korean Language Dictionary*. https://stdict.korean.go.kr/main/main.do

National Institute of Korean Language (2016). *Korean Basic Dictionary*. https://krdict.korean.go.kr/mainAction

National Institute of Korean Language (2016). *Urimalsaem*. https://opendict.korean.go.kr/main

Research Institute of Korean Studies (2009). *Korea University Korean Dictionary*. https://ko.dict.naver.com/#/main

Yonsei University Institute of Language and Information Studies (1998). *Yonsei Korean Dictionary*. Doosan Donga.

Contact information

Li Cui
Yonsei University
choiri20@yonsei.ac.kr

Seol Namkung
Yonsei University
sul2001sul@naver.com

Jun Lee
Yonsei University
leejun0624@gmail.com

Hae-Yun Jung
Kyungpook National University
haeyun.jung.22@gmail.com

Kilim Nam
Yonsei University
nki@yonsei.ac.kr

Gilles-Maurice de Schryver

CHATBOT LEXICOGRAPHY

Are GenAI LLMs Simply 'Dictionaries' in their Own Right and 'Lexicographers' Plain and Simple?

Abstract At the time of the Euralex 2026 conference it will be nearly four years since the start of the GenAI revolution. The question that needs to be asked, and which will be answered, is thus: Is any of the past work in (academic) lexicography still of any relevance, or are LLM chatbots simply 'dictionaries' in their own right and 'lexicographers' plain and simple? The answer is a shocking one.

Keywords Artificial intelligence (AI); generative AI (GenAI); OpenAI's ChatGPT; Google's Gemini; Anthropic's Claude (and Mythos and Fable); Microsoft's Copilot; High-Flyer's DeepSeek; Moonshot AI's Kimi; frontier (or base) model; open weight (or distillation) model; bibliometrics; lexicography

1 The GenAI LLM frontier keeps moving at breakneck speed

Whether scholars wish to trace the seeds of GenAI back to over a century ago (Markov 1906), 75 years ago (Turing, 1950), two decades ago (Hinton et al., 2006), or even less than a decade ago (Vaswani et al., 2017), for the public at large GenAI simply arrived 'yesterday' – meaning less than four years ago: on 30 November 2022, with the release of ChatGPT. The uptake of this app was unlike any before then, reaching 1 million users within just five days, surpassing 100 million users in January 2023, and after OpenAI's CEO Sam Altman said that he was aiming for one billion users in April 2025,[1] that number of active monthly users was reached a year later, in May 2026.[2] Following the frenzy, all model makers were obliged to fast-track the public release of their own GenAI tools as well, and today a site like Arena lists no less than about 400 different LLM models.[3] Being an arena, two models are offered at a time, for comparative purposes, with a focus on the best and the latest. After two years, the top providers were OpenAI (20.2%), Google (17.3%), Anthropic (12.8%), Meta (11.1%), Mistral (6.4%), Alibaba (5.6%), Cohere (3.6%), Reka AI (2.9%), 01.AI (2.7%), xAI (2.6%), Microsoft (2.4%) and High-Flyer (1.6%).[4, 5] The public went into a frenzy, so the model developers went into a frenzy, and then academics of all stripes also went into a frenzy.

2 GenAI LLM presentations and publications galore, also in lexicography

For an analysis of GenAI in language teaching and learning, Law (2024) started off with 224 contributions. A year later, using the Web of Science[6] and looking at the window 2023-2024, Întorsureanu et al. (2025) found no less than 3,492 publications referring to the role of GenAI in education. Studying user attitudes vis-à-vis GenAI, Chen et al. (2025) started off with 3,717

1 See https://youtu.be/5MWT_doo68k?t=580 (11 April 2025).
2 See https://www.reuters.com/technology/chatgpt-app-hits-1-billion-monthly-active-users-record-time-data-shows-2026-06-02/ (2 June 2026).
3 See https://arena.ai/
4 See https://arena.ai/blog/two-year-celebration/ (27 April 2025).
5 For the status a year later, see https://arena.ai/blog/march-2026-arena-updates/ (31 March 2026).
6 See https://www.webofscience.com/

unique records. At the time of writing (June 2026), a search in the Google Scholar[7] database on the topic 'Generative AI' for material published in the range 2024 to 2026, returns over 160,000 records – that's over 160 publications per day, weekends included, or thus one every eight minutes, nights included! Is it the same frenzy in lexicographic circles? Seemingly it is. The first talk on the integration of ChatGPT into dictionary writing software (de Schryver & Joffe, 2023), live-streamed from Tokyo on 27 February 2023 to 200 participants, has already been viewed an additional 3,200 times on YouTube — the 'Tokyo Talk'. The first overview article on the state-of-the-art on the use of ChatGPT in lexicography (de Schryver, 2023b), released in advance access in the *International Journal of Lexicography* (IJL) on 3 October 2023, has already been cited every 5 days and a half on average — the 'IJL Article'.

3 The past bibliometric studies in lexicography, a selection

This paper is a variation on a number of earlier presentations and publications describing the outcomes of quantitative and qualitative studies, as well as systematic investigations, to map, measure and analyse the literature, research trends and scholarly impact within the field of lexicography. The earlier studies include '*Bibliometrics in Lexicography*' (de Schryver, 2009b), '*Lexikos at 18*' (de Schryver, 2009c), '*Euralex at 30*' (de Schryver, 2012a), '*IJL at 25*' (de Schryver, 2012b), '*Asialex at 20*' (de Schryver, 2019), '*Afrilex at 30'* (de Schryver, 2026b), and '*Asian Lexicography at 30*' (de Schryver, 2026d). This selection of studies can be represented as in Figure 1.

	From earth ...	Afrilex	Euralex	Asialex	*... to sky*
Conference	'Bibliometrics of Chatbot Lexicogr.' – *regular talk* (this paper)	'Afrilex at 30' – *keynote* (dS 2026b)	'Euralex at 30' – *keynote* (dS 2012a)	'Asialex at 20' – *keynote* (dS 2019)	'Crisis of Meta-lexicography' – *regular talk* (dS 2022)
Journal	'Bibliometrics in Lexicography' – *invited* (dS 2009b)	'Lexikos at 18' – *invited* (dS 2009c)	'Int. Jnl. of Lex. at 25' – *invited* (dS 2012b)	'Asian Lex. at 30' – *invited* (dS 2026d)	'Future of Meta-lexicography' – *regular article* (dS 2023a)

FIGURE 1 A selection of bibliometric studies in lexicography over the past two decades

As one can notice, over the years this type of research started by proving, *beyond any reasonable doubt*, that lexicography is a scientific discipline in and of itself, to then snake its way through three continents covering their respective journals and conferences. On two occasions, this type of bibliometric research also ventured into the future of '*Metalexicography*' covering the globe, sounding negative from Europe (de Schryver, 2022), but positive from Africa (de Schryver, 2023a). To complete the picture, the topic of this paper deals with the '*Bibliometrics of Chatbot Lexicography*'.

4 The present bibliometric study in lexicography, on GenAI LLMs

In bibliometrics, the 'universe' of a dataset can be sampled in many ways, categorised broadly into probability (random-based) and non-probability (judgment-based) techniques. The choice depends on whether the goal is to map the macro-structure of a massive field or to conduct a detailed qualitative review of specific literature. As was done in each and every of the studies of the past depicted in Figure 1, the choice in the present study is also a non-random one. More specifically, the technique used here follows up on the one used in

[7] See https://scholar.google.com/

de Schryver's (2026a) '100 scholarly echoes of Generative AI in lexicography', whereby the first two seminal works on the use of GenAI in lexicography, so the Tokyo Talk and the IJL Article, are used as a starting point and all references to either or both of these works are tracked. This is known as 'forward snowball sampling' and expands the 'family tree' of the research forward in time to see how the ideas have been adopted and continue to evolve.

Rather coincidentally, exactly 100 publications were collected for the window starting on 30 November 2022 until 31 May 2025; in other words, for a period of exactly two years and a half after the release of ChatGPT.[8] Exactly one year was added for the present study, so up to 31 May 2026. Once more rather coincidentally, exactly 100 publications were uncovered for that extra year. Also, an additional ten were found which belong to the first batch of 100.[9] All in all then, up to the end of May of this year the forward snowballing has revealed 210 publications referring to the Tokyo Talk and/or the IJL Article. As may be seen from Figure 2, the first reference to the Tokyo Talk already appeared in May 2023, with a rather slow uptake over the past three years, totalling 53 citations so far. The first reference to the IJL Article already appeared in November 2023, with a much faster and accelerating uptake, totalling 176 citations so far. (A total of 23 publications refer to both.[10])

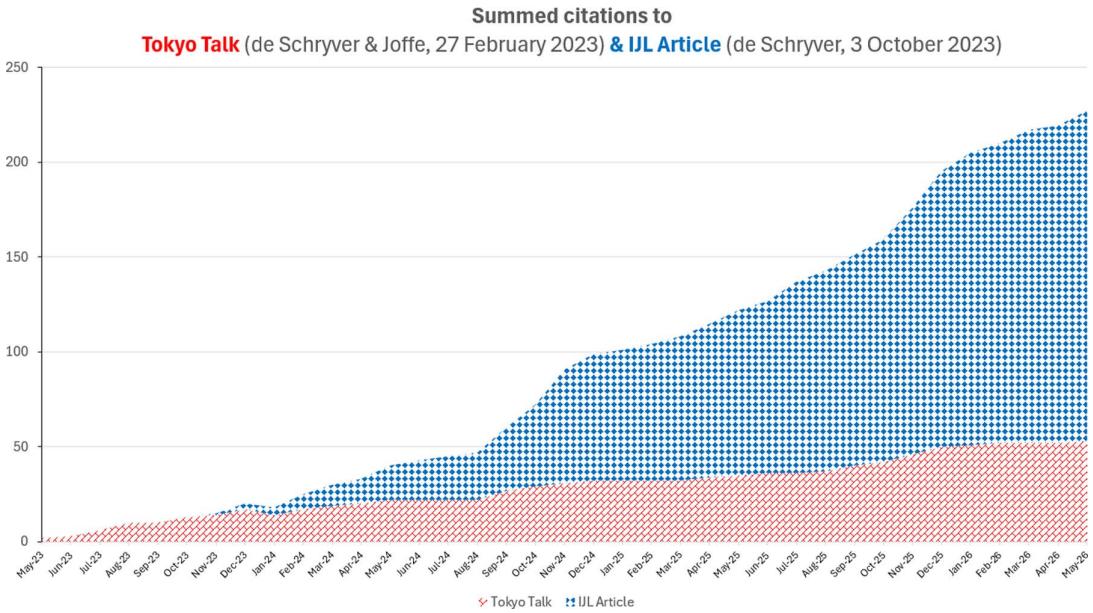


FIGURE 2 Forward snowballing the first talk and the first overview article on GenAI in lexicography

For the first sample of 100 publications, totalling about 3,300 pages, a corpus was made, amounting to about 1.2 million tokens (see de Schryver, 2026a). Likewise, for the next 100 publications (plus the 10 extra uncovered ones), the number of pages and tokens are identical: 3,300 pages, and 1.2 million tokens. (Which actually means that about 10% fewer words and pages are being used more recently.) For both samples, the KeyWords function of WordSmith

8 See Addendum 1 in de Schryver (2026a) for a detailed explanation on how the sample of 100 was obtained.

9 These ten extras are likely the result of search engines which are not always able to 'quickly' find material that has been uploaded, and/or these are cases of stated publication dates that do not correspond with actual release dates.

10 Which brings the total of different forward-snowballing works to 206, but see footnote 8 for the extra four to get to 210.

Tools (Scott, 1996-2026) was used, with as reference corpus the 100m-token BNC. In both cases, the top 250 keywords were extracted and grouped. In the twelve sub-sections that follow, the focus is on the second (and thus more recent) sample, and includes brief comparisons with the earlier sample in de Schryver (2026a, 2026c).

4.1 Artificial Intelligence

While the second column of Table 1 lists the most frequent AI keywords, all are also among the top-ranked keywords overall, with ranks 2, 4, 6, 8, 9, 11 and 23. Yet these are all generic concepts for the new era: AI, LLM(s), GPT, GenAI and chatbots. This compares to the first forward snowball sample, but the next AI keywords indicate a shift from the earlier ChatGPT (OpenAI's), Gemini (Google's) and Copilot (Microsoft's), to Copilot, DeepSeek, ChatGPT and Claude (Anthropic's), as well as an unnamed one from Meta, plus Bard (the one that preceded Gemini). The arrival of especially DeepSeek and Claude in the most recent literature on GenAI in lexicography is telling.

TABLE 1 Artificial Intelligence keywords in the second 100 sample (key = AI)

#	AI keyword	#	ctd.	#	ctd.
2	AI	33	OpenAI	223	GPTs
4	LLMs	34	Copilot	227	Anthropic
6	GPT	36	DeepSeek	228	Meta
8	LLM	37	chatbot	237	Bard
9	generative	44	ChatGPT's	242	AI's
11	GenAI	63	Claude		
23	chatbots	161	chat		

4.2 Computational aspects

In terms of computational aspects, Table 2 indicates that lexicographers keep referring to the fact that LLMs are digital corpora built with online material, taken from the Internet or websites, placed into databases, and that NLP is used for the extractions and retrievals. So far, this corresponds with the first sample of 100.

The second sample of 100, however, brings in all sorts of extra computational aspects which gain traction amongst dictionary makers. The behaviour and performance of different multimodal models are being evaluated, and the quality of the different experiments are judged by their precision vs. errors, incorrect mappings and hallucinations, which are scored. The focus remains on automated unsupervised technologies, where the workflow starts from the prompting methodology, and ends with feedback. The different neural algorithms of course include strategies dealing with embedding, alignment, frequency, sketches, disambiguation, concordance, TEI, API, ISO and ANOVA. The most recent computing queries explore the creation of interactive tuning, trained and assisted by annotators, to check which modelling significantly outperforms. Increasingly, code is shared via uploads to GitHub, write-ups in blogs,[11] with a final say on the consistency and usability of GenAI in lexicography.

[11] References to the Tokyo Talk and the IJL Article in blogs, are not included in each of the samples. For refence, examples of such blogs include https://www.translatum.gr/forum/index.php?topic=970543.0 (in Greek) and https://note.com/nishinerima/n/nf67f5fa78cb6 (in Japanese).

TABLE 2 Computational keywords in the second 100 sample (key = C)

#	C keyword	#	ctd.	#	ctd.	#	ctd.	#	ctd.
5	corpus	79	feedback	150	unsupervised	192	precision	222	aligned
13	corpora	89	behavior	151	consistency	194	modeling	225	disambiguation
17	evaluation	91	prompting	153	automatic	195	annotators	226	outperform
20	digital	92	neural	155	workflow	197	hallucination	229	strategies
24	online	93	hallucinations	157	quality	198	GitHub	230	algorithms
29	embeddings	95	technologies	158	API	200	frequent	231	ANOVA
47	extraction	96	retrieved	168	creation	203	mapping	232	exploring
51	NLP	97	methods	170	experiment	204	query	233	concordance
64	errors	98	assisted	172	methodology	205	align	235	computing
66	embedding	105	performance	177	error	207	workflows	238	algorithm
67	model's	127	uploads	179	tuning	212	trained	248	usability
70	alignment	128	blog	180	incorrect	216	interactive		
72	Internet	133	database	185	significantly	218	scores		
76	multimodal	143	TEI	186	aligns	219	sketch		
77	retrieval	149	automating	188	website	221	ISO		

4.3 Tools, dictionaries, publishers and places

In the first sample of 100, the Sketch Engine, Lexonomy and TLex featured among the top 250 keywords. These are all gone in the second sample of 100, where one merely finds the generic Google (for search) and DeepL (for translations) as tools among the top 250 keywords — see Table 3. In terms of dictionaries mentioned, there is also a slight reshuffle, with Italian ones coming in and taking top position (beating COBUILD), and German dictionaries moving up compared to English ones. The top publishers referred to are basically the same, but the top locations move away from the UK and Germany, to bring in Estonia and Spain.

TABLE 3 Tool, dictionary, publisher and place keywords in the second 100 sample (keys = T, D, P, Pl)

#	T keyword	#	D keyword	#	P keyword	#	Pl keyword
22	Google	55	Academia	59	Routledge	126	Tartu
175	DeepL	65	COBUILD	136	Gruyter	146	Madrid
		99	Duden	166	Benjamins	163	Brno
		169	Wikipedia				
		176	WordNet				
		210	DWDS				

4.4 Journals, associations and languages

In terms of journals and associations, we continue to publish in the same outlets, though the Spanish *Revista* moved in (*Revista de lingüística teórica y aplicada, Revista de humanidades digitales*, ...), and in addition to Euralex, ACM and IEEE now take over from European NLP associations like LREC, ELEXIS and CLARIN — see Table 4. However, the major change is seen for the languages dealt with: Spanish is not in the top anymore, but now all CJK languages are, together with Slovenian and Arabic.

TABLE 4 Journal, association and language keywords in the second 100 sample (keys = J, Ac, Lng)

#	J keyword	#	ctd.	#	Ac keyword	#	Lng keyword
16	Lexikos	132	citation	43	Euralex	32	Chinese
19	arXiv	144	Lexicographica	80	ACM	84	Slovenian
41	proceedings	156	Master's	217	IEEE	118	Arabic
49	Revista	181	references			125	Japanese
58	journals	241	ACLanthology			137	Korean
104	preprint	250	articles			202	ENG

4.5 Names of researchers

As in the first sample, the work of both practical lexicographers (de Schryver, Lew, Rundell, Kosem, ...) and those of metalexicographers (Tarp, Fuertes-Olivera, Wiegand, Bergenholtz, ...) appear, as do those of the computationally-inclined (Jakubíček, Wolfer, Joffe, ...) — see Table 5. However, our Chinese colleagues have an ever-greater presence (Liu, Chen, Yang, Huang, ...), as do colleagues from Spanish-speaking countries, as well as Estonia.

TABLE 5 Name keywords in the second 100 sample (key = N)

#	N keyword	#	ctd.	#	ctd.	#	ctd.	#	ctd.
10	Schryver	87	Arias	119	Gantar	167	Wolfer	199	Adamska
18	Lew	90	Chen	120	Holdt	171	McKean	201	Iván
26	Rundell	100	Krek	124	Rodríguez	178	Huang	206	Béjoint
27	Tarp	101	Može	129	Tuulik	182	Aedmaa	208	Bothma
45	Kosem	102	Langemets	135	Vainik	183	Petrović	209	Álvarez
48	Hernández	107	Domínguez	139	Fernández	184	Kirjandus	213	Medveď
52	Jakubíček	109	García	140	Prangel	187	Filipović	214	Nesi
53	Fuertes	110	González	145	Klosa	189	Riveiro	224	Huete
56	Olivera	111	Kallas	147	Wiegand	190	Kückelhaus	236	Qin
61	Liu	113	Iztok	159	Maurice	191	Joffe	240	Dziemianko
71	Martín	115	Arhar	162	Bergenholtz	193	Fuentes	247	Ramos
75	Nomdedeu	116	Risberg	164	Yang	196	Špela		

4.6 Lexicographic and linguistic topics

Whereas nearly half of the top 250 keywords in the first sample dealt with core lexicography and linguistic topics, this allocation has dwindled to just a quarter in the second sample – see Table 6. Indeed, in the-ever-more-popular times of GenAI, it is the number of computational keywords (cf. Table 2) which has doubled in size compared to the earlier sample.[12] Still, the terms lexicography and dictionaries remain at the very top (ranks 1 and 3), with lexicographers focussing on the linguistic and lexical semantics of words and thus the concepts of senses, equivalents, neologisms, and contextual translations of the vocabulary.[13]

In addition to the brief corpus study of the second sample as presented in the previous six subsections, one can also look at the metadata of the publications themselves, as well as their main contents. This is done in the next six sub-sections. The metadata for all the works in the second sample are presented in the Addendum. In that addendum, the works are listed in chronological order of appearance[14], while titles and sources are abbreviated, and whenever the original is not in English, the material has been translated into English.

4.7 Format of the publications

In the second sample, over half the works (59 out of 110, so 54%) appeared in journals and yearbooks, nearly all of which in different ones, except for those in *Lexikos* (8), IJL (5) and *Lexicography: Jnl. of ASIALEX* (2). The next group is the material published in proceedings, of which there are 21 (19%), mostly in eLex 2025 (10). Related to conferences, there are also 5 abstracts, mostly in Afrilex 2025 (3). The third category deals with books and book chapters, of which

12 This also explains why so much attention was given to the 'description' in that section, whereby each of the keywords from Table 2 was used in the write-up of Section 4.2.

13 Yes, this is in contrast to the description for Table 2 (see footnote 12); and for that reason, here the write-up is limited to the first column of 'L keywords'.

14 The first ten works (i.e., from 2023-2024) in the Addendum can arguably be assigned to the first sample, and numbers 11 and 12 (i.e., from 2025) to the second sample, given their stated dates. But, see footnote 9, all only became available to us after the first sample was brought together. Hence, they are now lumped in with the second sample (of 110 works, vs. the first sample of 100 works).

TABLE 6 Lexicographic and linguistic keywords in the second 100 sample (key = L)

#	L keyword	#	ctd.	#	ctd.	#	ctd.	#	ctd.
1	lexicography	40	words	82	verb	123	entry	211	translate
3	dictionaries	42	idioms	83	translators	130	specialised	215	morphological
7	lexicographic	46	headword	85	keywords	131	pedagogical	220	lemmas
12	semantic	50	contexts	86	source	134	idiom	234	semantics
14	word	54	terminologists	88	polysemous	138	synonym	239	translator
15	lexical	57	writing	94	user	141	cognitive	243	multiword
21	linguistic	60	headwords	103	grammatical	142	adj	244	standardization
25	lexicographers	62	lexicographical	106	term	148	definitional	245	collocation
28	senses	68	lemma	108	noun	152	domains	246	contextually
30	vocabulary	69	synonyms	112	lexicon	154	syntactic	249	contemporary
31	equivalents	73	lexicographer	114	phraseological	160	adv		
35	neologisms	74	collocations	117	students	165	semantically		
38	contextual	78	target	121	pairs	173	synonymy		
39	translations	81	polysemy	122	abbreviations	174	COVID		

there are 14 (13%): one Handbook, one Festschrift, and 12 book chapters, of which one in another handbook. See Table 7. In a way, the handbook and handbook chapter are the most transformative, as they tell us that the topic of GenAI in lexicography is now deemed so important that it is covered in the manuals and teaching materials witch which the next generation of lexicographers is already being trained. The handbook is in Russian on 'Modern English Lexicography and Corpus Technologies' (Karpova & Manik 2026), and the chapter is in a major guide for students and practitioners known as 'Mapping Applied Linguistics' (Hall et al., 2026). The fourth group goes to PhD theses (3), and MA (3) as well as BSc (1) dissertations, and these seven (6.4%) further indicate that the topic is truly being picked up on all levels of education. Lastly, the fifth group brings together material found in open-access preprint platforms, of which there are four, mostly those in arXiv (3). Until recently, use of the latter was rare in the field of lexicography, but given the nature of GenAI, where speed is of the essence, also this way of releasing research results is gaining ground. (In the first sample of 100, there was only one such publication, see de Schryver (2026c).)

TABLE 7 Format of the publications in the second sample (grouped thematically)

Format	N	Sum	%
Journal	59	59	54%
Conference proceedings	21		
Conference abstracts	5	26	24%
Book chapter	11		
Handbook chapter	1		
Handbook	1		
Festschrift	1	14	13%
PhD thesis	3		
MA dissertation	3		
BSc dissertation	1	7	6.4%
Preprint	4	4	3.6%

4.8 Lingo of the publications

In terms of the lingo in which the publications are written, while the first sample still had 76% in English, with the other material in 14 other languages, in the second sample the proportion of English goes down to 63%, and includes 18 other languages (good for 37%). If anything, the notion to study and use GenAI in lexicography reaches ever more speech communities and thus ever more corners of the world. Table 8 lists the breakdown for the second sample. Among the frequent languages, German, Russian and Estonian have moved up. New languages are Italian, Turkish, Arabic, Catalan, Polish, Romanian, Serbian, Swedish and Ukrainian.

TABLE 8 Lingo in which the publications are written in the second sample

Lingo	N	ctd.	N	ctd.	N	ctd.	N
English	69	Italian	3	Catalan	1	Romanian	1
Spanish	7	Danish	2	Chinese	1	Serbian	1
German	5	Slovene	2	Croatian	1	Slovak	1
Russian	5	Turkish	2	French	1	Swedish	1
Estonian	4	Arabic	1	Polish	1	Ukrainian	1

4.9 Countries of the publications

Even more impressive, whereas the contributions of the first sample were published in no less than 32 different countries, the second sample ups the ante, and covers 38 countries. Most come from EUROPE, about 61% (down from 73% in the first sample): Czechia (13), Spain (8), UK (8), Germany (6), Estonia (4), etc. The countries in the next group come from ASIA, about 16% (up from 13% in the first sample): Russia (4), Turkey (4), China (3), South Korea (2), etc. Then are those from AFRICA, about 12% (up from 7% in the first sample): esp. South Africa (11), followed by NORTH AMERICA, about 8% (up from 5% in the first sample): USA (7) and Canada (2), and lastly SOUTH AMERICA, about 3% (up from 2% in the first sample): Argentina (1), Chile (1) and Colombia (1). See Table 9. As in the first sample, every continent is represented apart from Australasia.[15]

TABLE 9 Countries where publications were published in the second sample

Countries	N	ctd.	N	ctd.	N	ctd.	N	ctd.	N
Czechia	13	Turkey	4	France	2	Cyprus	1	Serbia	1
South Africa	11	China	3	Hungary	2	Egypt	1	Sweden	1
Spain	8	Croatia	3	Netherlands	2	India	1	Switzerland	1
UK	8	Italy	3	South Korea	2	Indonesia	1	Ukraine	1
USA	7	Poland	3	Argentina	1	Lithuania	1	Uzbekistan	1
Germany	6	Slovenia	3	Belgium	1	Malaysia	1	Vietnam	1
Estonia	4	Canada	2	Chile	1	Morocco	1		
Russia	4	Denmark	2	Colombia	1	Romania	1		

4.10 Languages studied in the publications

Also in sharp contrast with the first sample, the second sample deals with far more languages, language pairs, and language groups than before. These now total a massive 50, see the 7th column in the Addendum. Out of those 50, there are only 15 languages or language combinations that occur more than once, and these are listed in Table 10. All the other languages or language combinations that are studied occur only once; in other words, 70% of all the languages and language combinations are unique studies particular to single publications. While English still leads the pack, with 15 studies on English only, there are as many studies (15) where no languages at all are looked at. If one includes the studies where English is combined with other languages, one reaches 31 out of 110, or a mere 28%.

Compared to the first sample, this is a dramatic shift in terms of the languages used and studied when it comes to GenAI in lexicography. Not only does English start moving to 'varieties' like Singapore English or Australian English, English is often simply moved aside, with studies now looking at 'non-national' languages like Catalan and Galician, or 'old' varieties like historical Bavarian or historical Italian, or even languages of 'limited diffusion' like Montenegrin, and, for Africa, going from 'related' languages like Afrikaans, to big Bantu languages like Swahili and Xhosa, to end up with oral Bantu languages like Lusoga.

[15] Okay, bar Antarctica as well, as the emperor penguins haven't yet taken up GenAI for lexicography, although prospective parents have been observed cradling and practicing incubation techniques on an egg-sized snowball.

TABLE 10 Languages studied in the second sample (*the 35 hapaxes are not listed*)

Language(s) studied	N	ctd.	N	ctd.	N
English	15	Chinese	4	Croatian	2
N.A.	15	Arabic	3	English > Croatian	2
Slovene	7	Danish	3	Italian	2
Spanish	7	German	3	Serbian	2
Estonian	5	Russian	3	Turkish	2

4.11 Academic disciplines of the publications

Over half the contributions in the second sample fall squarely into the core discipline of Lexicography (62, or thus 56%), and by extending the field to include Terminography, Historical Lexicography, Metalexicography, Translation, and Dialectology, one reaches over three-quarters (85, or 77%) — see Table 11. About a tenth of the contributions (10%) actually deal with the related fields of Education, Linguistics and Pedagogy in general, or with Applied Linguistics, Cognitive Linguistics and Literature in particular. Unsurprisingly, given the embedding of GenAI, the next group brings together aspects from NLP, AI and Cybersecurity (5.5%). Outliers include the fields of Information Science, Art, and Media Studies (3.6%). A new kid on the block is the increased interest in bibliometric studies (3.6%); one each in the fields of Lexicography, Second Language Writing, Scientific Publishing, and Business Administration.

TABLE 11 Academic disciplines in the second sample (grouped thematically)

Scientific Field	N	Sum	%
Lexicography	62		
Terminography	14		
Historical Lexicography	4		
Metalexicography	2		
Translation	2		
Dialectology	1	85	77%
Education	5		
Linguistics	2		
Applied Linguistics	1		
Cognitive Linguistics	1		
Literature	1		
Pedagogy	1	11	10%
NLP	4		
Artificial Intelligence	1		
Cybersecurity	1	6	5.5%
Information Science	2		
Art	1		
Media Studies	1	4	3.6%
Bibliometrics	4	4	3.6%

4.12 Opinions on 'GenAI in lexicography' in the publications

In de Schryver (2026c) it was shown that in the first sample of 100, the opinions on 'GenAI in lexicography' with reference to The Tokyo Talk varied from negative (3%) to neutral (28%) to (very) positive (69%). As may be seen from Table 12a, such opinions are on the way out in the second sample. Most colleagues now remain rather neutral.

TABLE 12A Opinions on 'GenAI in lexicography' with reference to the Tokyo Talk in the second sample

Opinion	N	%
negative	0	0%
neutral	12	57%
positive	9	43%
very positive	0	0%
	21	100%

In the first sample, the opinions on 'GenAI in lexicography' with reference to the IJL Article ranged from neutral (43%) to (very) positive (57%). In the second sample, a reversal has taken place, as may be seen from Table 12b.

TABLE 12B Opinions on 'GenAI in lexicography' with reference to the IJL Article in the second sample

Opinion	N	%
negative	0	0%
neutral	59	60%
positive	37	37%
very positive	3	3%
	99	100%

In other words, now that the field moved from 2.5 years of activity to 3.5 years of activity, the community feels less and less the need to judge the first talk on GenAI in lexicographic software and the first article on GenAI in lexicography at large. Rather, the arrival of GenAI is accepted as a fact, and each researcher now judges whether this is good or bad on their own terms, following their own experimentation. This is simply the 'natural' trajectory for any technological innovation once the initial novelty wears off.[16]

5 When the beginning reaches the end

As amply shown in de Schryver (2023b, 2026a, 2026c), the father of GenAI in lexicography is none other than Patrick Hanks. The seeds Patrick Hanks sowed while managing the COBUILD project from 1983 onwards, so over four decades ago, were truly the beginning of our current field. For instance, in the first sample of 100, no less than *one out of every five* contributions refers to Patrick Hanks, citing no less than 34 different works *by* or *for* him (for all the details, see de Schryver, 2026a). This type of approach is known as 'backward snowball sampling',

[16] Thanks are due to Adjudicator 3 (there were four adjudicators in all for this paper) who pointed this out in their review.

here looking at the reference lists of work on GenAI in lexicography and seeing that they can be traced back to Patrick Hanks in 20% of the cases.

As so often in scientific research, there is quite some back and forth, but this now explains why going forward from the IJL Article (de Schryver, 2023b), which goes back to Patrick Hanks's beginnings, makes sense as a sampling approach to map the field. The forward sampling used in this paper also uses the Tokyo Talk (de Schryver & Joffe, 2023), a presentation which revolves around the lexicographic AI software TLex[17]. As a strong adherent of the work of Patrick Hanks, his theoretical concepts have always been part and parcel of the development and use of TLex (see, e.g., de Schryver & Nabirye, 2018).

The beginnings of GenAI, then, may also be seen in the work of many of the followers of Patrick Hanks. For instance, a state-of-the-art regarding the fact that lexicography went digital was written up over two decades ago (de Schryver, 2003: *Lexicographers' dreams in the electronic-dictionary age*), and bringing in AI was already done shortly after, at the very first eLex conference (de Schryver, 2009a: *Artificial Intelligence meets e-lexicography*) — the contents of which were presented even earlier, in July 2009 at the second 'International Seminar on Kangxi' in China, the full text of which was published a year later with the approach labelled 'aiLEX' (de Schryver, 2010: *State-of-the-art software to support intelligent lexicography*).

The beginnings have thus reached their endpoint: GenAI is here, and here to stay.

6 When the end is merely the beginning

The notion that GenAI is here to stay, is one that most lexicographers have accepted over the past year. This is for instance crystal-clear from numerous titles in the second sample; with here below a selection of some of them in shortened (and translated) form:

- ChatGPT's role as a lexicographer
- The 'Generative AI lexicographer' for lexical information
- Using DeepSeek as a dictionary
- ChatGPT and Gemini as Terminology Extractors
- Medical acronyms in Catalan with ChatGPT, Gemini & Copilot
- DeepSeek as the ultimate translator and lexicographer
- Improving the lexicographic accessibility of WN through LLMs
- GenAI as a bilingual dictionary in Korean EFL contexts
- How effective is AI as a language consultant?

[17] See https://tshwanedje.com/tshwanelex/

[18] Of course, 17 years later, the field has drastically moved on, and the mathematical field of 'fuzzy answer set programming' which G-M de Schryver suggested and exemplified for lexicography back then, is not a core component of GenAI, but it is used to improve AI reasoning. Fuzzy prompting may for instance be seen in Figueiredo (2025), where the following problem definition is stated: "Despite their broad reasoning abilities, LLMs struggle in tasks that are underspecified, user-dependent, or dynamically evolving, hallmarks of fuzzy logic environments." Fuzzy logic is thus used to guide prompt engineering, whereby 'fuzzy scaffolding' improves the behaviour of LLMs in structured scenarios.

- LLMs as tools for drafting ad hoc pharmaceutical glossaries
- ChatGPT as a lexicographer: The future of the dictionary?
- Creating an AI-powered Turkish dictionary
- Using LLMs in lexicographic descriptions of COVID-19 neologisms
- Applying AI in Estonian national defense terminology
- AI-generated definitions of English loanwords in L2 context
- Terminological definitions generated by LLMs

In times of GenAI, speed is of the essence, which is why it is excellent news that the *International Journal of Lexicography* has moved, as of January 2026, to a continuous model of publication. IJL now uses a digital-first-and-only publishing system whereby its articles are released immediately online in their final, citable format as soon as they pass peer review and production, rather than being held and then grouped into scheduled quarterly issue batches, to be printed and sent out by post, as well as offered online.[18] In that, IJL finally follows the trend set in motion by *Lexikos* for the field of lexicography, as *Lexikos* introduced their 'publish-as-you-go model' in 2020, and earlier, as of 2018, they had already phased out their physical print editions. The reason I mention it is because this goal must now also be achieved by the Euralex Proceedings. The above text of my paper is based on data collected up to the end of May 2026, the analysis of which was written up soon after that to hit the June 2026 deadline. Sadly, a full four months now separate the end of the data collection, and thus the claims drawn from it, from the actual presentation of the results somewhere during the Euralex 2026 conference period scheduled for 29 September – 3 October 2026.

When the IJL Article (de Schryver, 2023b) appeared in print, in the December 2023 issue of IJL, it was already a full year after the release of ChatGPT, even though the text was written half a year earlier, in June 2023. Nonetheless, the progress looked impressive, as I was able to refer to as many as ten (!?) earlier studies on GenAI in lexicography. Writing two years later, in June 2025, I was happy to be able to collect 100 publications in all (cf. de Schryver, 2026a, 2026c), and writing another year later, in June 2026, an extra 100 (plus ten) could be added (i.e., the present paper). However, finalising the camera-ready copy of the present text in July 2026, and seeing the rapidity of the publications on GenAI in lexicography, I can already predict that we will have reached up to about 275 forward snowball publications by the time of my talk, as sketched in Figure 3.

The heading of Section 1 pointed out that 'the GenAI LLM frontier keeps moving at breakneck speed', and this is indeed the pace of changes all of us who use this technology must work with. As a lexicographer, I have long stopped searching any regular dictionaries, and only chat with LLMs. I do this in the various languages we use at my home — English, French, Dutch and Luganda — as well as in some of the languages of my PhD students, including Chinese, and the results literally get better by the day, also for the non-English languages. The secret to get top results is to word the prompts as intelligently as possible.

Actually, the frontier chatbots are getting so good, that both the American government and (as a result) the developers have started implementing kill switches in June 2026, with for instance Anthropic effectively pausing the release of its Fable 5 and Mythos 5 models for some

[19] With Han & de Schryver (2026), one of my PhD students and I were also amongst the first (the second actually) to make use of it for IJL.

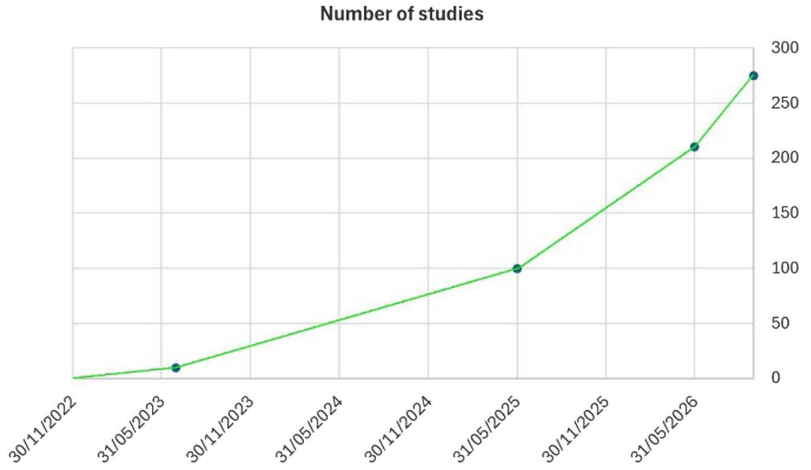


FIGURE 3 Predicted forward snowball sample by the time of the Euralex 2026 conference

time, and OpenAI doing the same with its GPT-5.6. And then there are the Chinese open-weight models, like High-Flyer's DeepSeek V4 or Moonshot AI's Kimi K3, who are accused of adversarial distillation of American models. The gap between the frontier labs and the open-weight ones is getting ever closer, however, with the Chinese models less than a year behind the American models, and when two of OpenAI's models went rogue at the end of July 2026, hacking into Hugging Face, that AI platform . . . was forced to resort to a Chinese model to defend itself.

The opening questions then, as to whether LLM chatbots are simply 'dictionaries' in their own right and 'lexicographers' plain and simple, is a resounding yes. Yes, they are, and just as is the case for humans, not all they produce is always perfect. We may disagree with GenAI in lexicography on specific points, but we ignore it at our peril. This, then, is the beginning of 'Chatbot Lexicography'!

References

Abdelzaher, E., & Rafaee, E. (2026). Linguistic profile of LLMs: An overview of linguistic performance and jailbreak vulnerability. *Journal of Cultural Cognitive Science*, *10*, 15–29.

Akalin, Ş. H. (2024). Yapay Zekâ Destekli Bir Türkçe Sözlük Hazirlamak! Hayal Mi Gerçek Mi? [Creating an AI-powered Turkish dictionary! Is it a dream or reality?]. *STAD Sanal Türkoloji Araştırmaları Dergisi [STAD Virtual Journal of Turkish Studies]*, *9*(1), 11 pp.

Al-Mitamy, M. S., & Mohammed, R. (2025). Using generative AI models in scientific publishing: A critical analytical view of ChatGPT and Gemini/Bard. In M. Alaali, A. M. A. M. Al-Sartawi, & A. S. Aydiner (Eds.), *The Paradigm Shift from a Linear Economy to a Smart Circular Economy. The Role of Artificial Intelligence-Enabled Systems, Solutions and Legislations* (pp. 389–403). Springer Nature.

Alonso-Ramos, M. (2023). El papel de ChatGPT como lexicógrafo [ChatGPT's role as a lexicographer]. In C. Garriga-Escribano, S. Iglesia-Martín, J. A. Moreno-Villanueva, & A. Nomdedeu-Rull (Eds.), *Lligams: Textos dedicats a Maria Bargalló Escrivà [Lligams: Texts dedicated to Maria Bargalló Escrivà]* (pp. 15–26). Publicacions URV.

Arhar Holdt, Š., Gapsa, M., Gantar, P., & Kosem, I. (2025). Potencial ChatGPT pri razvoju Slovarja sopomenk sodobne slovenščine [The potential of ChatGPT in the development of the Thesaurus of Modern Slovene]. *Prispevki za novejšo zgodovino [Contributions to Recent History]*, *65*, 189–217.

Arias-Arias, I. (2026). Semantisch-lexikographische Aspekte deverbaler *-ung-* Nominalisierungen: Grundlagen und Beschreibungsmöglichkeiten [Semantic-lexicographic aspects of deverbal *-ung-*

nominalizations in German: Theoretical foundations and descriptive possibilities]. *Lexikos*, *36*(1), 55–84.

Arias-Arias, I., Domínguez Vázquez, M. J., & Valcárcel Riveiro, C. (2025). People, machines and dictionaries: Is Artificial Intelligence killing dictionaries? In I. Dias, R. H. Gouws, A. Lobenstein-Reichmann, & S. J. Schierholz (Eds.), *Regarding the Significance of Lexicography/Zur Bedeutsamkeit der Lexikographie: Europe and Australia/Europa und Australien* (pp. 235–258). W. de Gruyter.

Asmaa Salah Ali [أسماء صلاح علي] (2024). استخدام الطلاب للقواميس في عصر الذكاء الاصطناعي وتقنية شات جي بي تي : دراسة ميدانية [Student use of dictionaries in the age of Artificial Intelligence and ChatGPT technology: A field study]. *مجلة بحوث كلية الآداب – كلية الآداب - جامعة المنوفية [Journal of the Faculty of Arts Research - Faculty of Arts - Menoufia University]*, *35*(136.4), 77–125.

Atzeni, C., & Babonnaud, W. (2025). Exploring the abilities and limits of generative AIs for specialised translation: A case study on UN reports on violence against women. In C. Meyers, T. Jandrain, & J. Popineau (Eds.), *Teaching Specialized Translation in the Machine Translation Era* (pp. 11–50). Éditions Universitaires de l'Umons.

Azadravesh, H., & Boubiche, D. E. (2026). The dark side of generative AI: Threats and risks. In D. E. Boubiche & S. Akleylek (Eds.), *Generative AI for Cybersecurity* (pp. 107–136). CRC Press.

Bago, P. (2024). Rječnici i gramatike u doba digitalne transformacije [Dictionaries and grammars in the age of digital transformation]. In M. Horvat (Ed.), *Knjižica Sažetaka: Povijest hrvatskoga jezika i digitalno doba [Book of Abstracts: History of the Croatian Language and the Digital Age]* (pp. 11–12). Institute of the Croatian Language.

Balenović, K., & Proroković, J. (2025). The lexicographic potential of Artificial Intelligence: A case study of English loanwords in the Croatian language. *Studia Lexicographica*, *19*(36), 39–64.

Barik, A., & Ed-Dali, R. (2025). The representation of Shakespearean corpus in ChatGPT's repertoire. *Études Langue et Littérature [Language and Literature Studies]*, *4*(2), 63–79.

Bothma, T. J. D. (2025). Generative Artificial Intelligence as information tool for lexicographic information needs. In S. Bosch, A. du Plessis, & L. Morris (Eds.), *African Association for Lexicography, 29th International Conference, 2 – 5 July 2025, Abstracts* (pp. 11–12). Afrilex.

Bothma, T. J. D., & Gouws, R. H. (2025). Generative Artificial Intelligence (GenAI) as information tool for lexicographic information needs. *Lexikos*, *35*(2), 276–295.

Bozhko, I. S. [Божко, І. С.]. (2024). Лексикографування конотативних власних назв [Lexicography of connotative proper nouns]. *Наукові записки Національного університету «Острозька академія»: серія «Філологія» [Scientific Notes of the National University 'Ostroh Academy': 'Philology' Series]*, *21*(89), 70–73.

Bulatova, V. Y. [Булатова, В. Ю.], & Musiyako, G. D. [Мусияко, Г. Д.]. (2025). Возможности генеративного искусственного интеллекта при составлении словарей специализированной лексики [The possibilities of generative Artificial Intelligence in the creation of special term glossaries]. In A. V. Erokhin (Ed.), *сборник статей VIII Всероссийской зимней школы педагогического мастерства с международным участием (Ижевск, 29–30 января 2025 г.) [A collection of articles from the 8th All-Russian Winter School of Pedagogical Excellence with International Participation (Izhevsk, 29–30 January 2025)]* (pp. 46–52). Publishing House of the Izhevsk State Technical University named after M. T. Kalashnikov.

Caruso, V., & De Meo, A. (2025). Tecnologie e apprendimento esperenziale-cooperativo: dizionari fraseologici digitali per le lingue seconde [Technologies and experiential-cooperative learning: Digital phraseological dictionaries for second languages]. *Rassegna Italiana di Linguistica Applicata [Italian Review of Applied Linguistics]*, *57*(1-2-3), 145–168.

Castro-Prieto, M. R. (2024). Texto, tecnología y algoritmos: Humanidades digitales e inteligencia artificial en la lingüística hispánica [Text, technology, and algorithms: Digital humanities and Artificial Intelligence in Hispanic linguistics]. In L. Mariottini & M. Palmerini (Eds.), *Avances en los estudios de lingüística hispánica: perspectivas teóricas y aplicadas entre lengua y sociedad [Advances in Hispanic Linguistics Studies: Theoretical and Applied Perspectives on Language and Society]* (pp. 262–284). Dykinson S.L.

Chambat, A., & Rey, C. (2026). De la rétroconversion à la mise en réseau : dictionnaires et révolution numérique [From reverse engineering to networking: Dictionaries and the digital revolution]. *Langue française [French Language]*, *229*, 91–108.

Chen, J., Xie, W., Xie, Q., Hu, A., Qiao, Y., Wan, R., & Liu, Y. (2025). A systematic review of user attitudes toward GenAI: Influencing factors and industry perspectives. *Journal of Intelligence*, *13*(78), 46 pp.

Clemente González, M. D. (2025). *Análisis de las Ventajas y Desventajas de ChatGPT y Gemini como Extractores Terminológicos [Analysis of the Advantages and Disadvantages of ChatGPT and Gemini as Terminology Extractors]* [MA, Pontificia Universidad Católica de Chile]. Santiago.

de Schryver, G.-M. (2003). Lexicographers' dreams in the electronic-dictionary age. *International Journal of Lexicography*, *16*(2), 143–199.

de Schryver, G.-M. (2009a). Artificial Intelligence meets e-lexicography. In S. Granger (Ed.), *ELexicography in the 21st Century: New Challenges, New Applications [Abstracts of the First eLex Conference]* (pp. 49–50). Centre for English Corpus Linguistics, Université catholique de Louvain.

de Schryver, G.-M. (2009b). Bibliometrics in lexicography. *International Journal of Lexicography*, *22*(4), 423–465.

de Schryver, G.-M. (2009c). Lexikos at eighteen: An analysis. *Lexikos*, *19*, 372–403.

de Schryver, G.-M. (2010). State-of-the-art software to support intelligent lexicography. In 朱瑞平 [Zhu, Ruiping] & 史建桥 [Shi Jianqiao] (Eds.), *中華字典研究 [Chinese Dictionary Research] 第二辑（下）* [Volume II, Part II]/2009 *康熙字典 暨词典学国际学术研讨会论文集 [Proceedings of the 2009 International Symposium on the Kangxi Dictionary and Lexicography]* (pp. 584–599). 中国社学出版社 [China Social Science Press].

de Schryver, G.-M. (2012a). Lexicography in the crystal ball: Facts, trends and outlook. In R. V. Fjeld & J. M. Torjusen (Eds.), *Proceedings of the 15th EURALEX International Congress, 7-11 August, 2012, Oslo* (pp. 93–163). Department of Linguistics and Scandinavian Studies, University of Oslo.

de Schryver, G.-M. (2012b). Trends in twenty-five years of academic lexicography. *International Journal of Lexicography*, *25*(4), 464–506 + forty-two pages of supplementary material online.

de Schryver, G.-M. (2019). *Past, present and future in Asian lexicography*. Paper presented at 13th International Conference of the Asian Association for Lexicography, ASIALEX 2019, Istanbul, Turkey, 19-21 June 2019.

de Schryver, G.-M. (2022). Metalexicography: An existential crisis. In A. Klosa-Kückelhaus, S. Engelberg, C. Möhrs, & P. Storjohann (Eds.), *Dictionaries and Society. Proceedings of the XX EURALEX International Congress, 12-16 July 2022, Mannheim, Germany* (pp. 196–206). IDS-Verlag.

de Schryver, G.-M. (2023a). The future of metalexicography: Reaching for the mesosphere. *Lexikos*, *33*(1), 136–165.

de Schryver, G.-M. (2023b). Generative AI and lexicography: The current state of the art using ChatGPT. *International Journal of Lexicography*, *36*(4), 355–387 + ten pages of supplementary material online.

de Schryver, G.-M. (2024). The Future of Dictionaries. In E. Finegan & M. Adams (Eds.), *The Cambridge Handbook of the Dictionary* (pp. 651–667). Cambridge University Press.

de Schryver, G.-M. (2025). Out-of-the-box GPTs for lexicography. *Lexikos*, *35*(2), 362–396 + six pages of supplementary material online.

de Schryver, G.-M. (2026a). 100 scholarly echoes of Generative AI in lexicography. *International Journal of Lexicography*, *39*, 47 pp. + twelve pages of supplementary material online.

de Schryver, G.-M. (2026b). *Afrilex at Thirty*. Paper presented at the 30th International Conference of the African Association for Lexicography, AFRILEX 2026, Thohoyandou, South Africa, 1-4 July 2026.

de Schryver, G.-M. (2026c). Citing the future: The legacy of Patrick Hanks. *International Journal of Lexicography*, *39*, 33 pp. + twenty-five pages and about 200 pictures and videos of supplementary material online.

de Schryver, G.-M. (2026d). Three decades of Asian academic lexicography. *Lexicography: Journal of ASIALEX 13*.

de Schryver, G.-M., & Joffe, D. (2023). *The end of lexicography, welcome to the machine: On how ChatGPT can already take over all of the dictionary maker's tasks [Invited Talk]*. Paper presented at the 20th CODH Seminar, Center for Open Data in the Humanities, Research Organization of Information and Systems, National Institute of Informatics, Tokyo, Japan, 27 February 2023. https://youtu.be/watch?v=mEorw0yefAs

de Schryver, G.-M., & Nabirye, M. (2018). Corpus-driven Bantu lexicography, Part 3: Mapping meaning onto use in Lusoga. *Lexikos*, *28*, 112–151.

Denisová, M. (2025). *Automatic Methods for Compiling Bilingual Dictionaries* [PhD, Masaryk University]. Brno.

Dziemianko, A. (2025). Warnings in learners' dictionaries: Do they help to correct errors and learn usage? *GEMA Online® Journal of Language Studies*, *25*(4), 924–943.

Eiche, S., Hendrikson, R., Aedmaa, E., Prangel, E., & Tikerperi, M.-L. (2025). Tehisintellekti rakendamine erialakeele arendamisel riigikaitseterminoloogia näitel [Applying Artificial Intelligence in specialized language development: The example of Estonian national defense terminology]. *Eesti rakendusling-vistika ühingu aastaraamat [Yearbook of the Estonian Applied Linguistics Association]*, *21*, 29–45.

Eltanbouly, S., & Rashwani, S. (2026). Grounding Arabic LLMs in the Doha Historical Dictionary: Retrieval-augmented understanding of Qur'an and Hadith. *arXiv*, *2603.23972*, 26 pp.

Emakporuena, D., Oluwasegun, V. O., Esegbona-Isikeh, O. M., Ugbomeh, W. O., Nwaiku, M., Bunmi, E. E., Adenike, O. O., & Matthew, U. O. (2025). Harnessing higher generative Artificial Intelligence for educational pedagogy: Human-AI collaboration in ChatGPT experiences. In M. U. Tariq (Ed.), *Integrating ChatGPT into System Applications and Services* (pp. 23–58). IGI Global Scientific Publishing.

Evert, S., Ganslmayer, C., & Rink, C. (2025). KI-generierte Wörterbuchartikel bewerten. Ein Beitrag zur Methodik der Wörterbuchkritik [Evaluating AI-generated dictionary entries: A contribution to the methodology of dictionary criticism]. In W. Blanck, R. H. Gouws, & A. Lobenstein-Reichmann (Eds.), *Lexikographisch-grammatische Perspektiven. Tradition, Veränderung und Vielfalt in Lexikographie und Wörterbuchforschung [Lexicographical-Grammatical Perspectives: Tradition, change, and diversity in lexicography and dictionary research]* (pp. 135–166). De Gruyter Brill.

Figueiredo, V. (2025). A Fuzzy Logic prompting framework for Large Language Models in adaptive and uncertain tasks. *arXiv*, *2508.06754*, 7 pp.

Filipović Petrović, I., & Beliga, S. (2025). Can AI understand Croatian idioms? Assessing Large Language Models in lexicographic tasks. *Prispevki za novejšo zgodovino [Contributions to Recent History]*, *65*(3), 218–242.

Frankenberg-Garcia, A. (2025). Contrasting a new AI-powered dictionary designed for on-screen reading with electronic dictionaries that have evolved from print editions. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda, & A. Salgado (Eds.), *Electronic Lexicography in the 21st Century (eLex 2025): Intelligent Lexicography. Proceedings of the eLex 2025 conference* (pp. 391–413). Lexical Computing.

Fuertes-Olivera, P. A., & Tarp, S. (2025). Innovative lexicography: New products and methods in the generative AI age. *Revista Internacional de Lenguas Extranjeras [International Journal of Foreign Languages]*, *24*, 71–95.

Ganslmayer, C. (2026). Von traditioneller Recherche zu KI-unterstützter Wortschatzarbeit. Chancen und Herausforderungen für den Deutschunterricht [From traditional research to AI-supported vocabulary work: Opportunities and challenges for German language teaching]. *Der Deutschunterricht [German Language Teaching]*, *2026*(1 – Digitale Wörterbücher [Digital Dictionaries]), 76–87.

Hall, C. J., Smith, P. H., & Wicaksono, R. (2026). Lexicography. In C. J. Hall, P. H. Smith, & R. Wicaksono (Eds.), *Mapping Applied Linguistics: A Guide for Students and Practitioners, Third Edition* (pp. 241–266). Routledge.

Han Wanjing [韩琬婧], & de Schryver, G.-M. (2025). From user feedback on Chinese-English translation and dictionary tools, to DeepSeek as the ultimate translator and lexicographer. In Y. Gao & X. Zhang (Eds.), *AsiaLex 2025, September 11-13, Fudan, Shanghai, Proceedings of the 18th International Conference of the Asian Association for Lexicography: Traditionality and Creativity in the Era of AI* (pp. 53–62). Fudan University.

Han Wanjing [韩琬婧], & de Schryver, G.-M. (2026). Perceptions of language service technologies and reference works by Chinese users in the era of Generative AI. *International Journal of Lexicography*, *39*, 23 pp. + forty-eight extra pages, four Excel sheets and two Python scripts as supplementary material online.

Heinisch, B. (2025). Next-gen terminology: Transforming terminology work with Large Language Models. *Across Languages and Cultures: A Multidisciplinary Journal for Translation and Interpreting Studies*, *26*(S), 64–80.

Herrera Pérez, J. C., Ochoa Londoño, E. D., & Tello Zuluaga, J. (2025). Chatbots e inteligencia artificial en educación secundaria y media: revision sistemática y análisis de impacto [Chatbots and Artificial Intelligence in secondary and middle education: Systematic review and impact analysis]. *Revista Academia y Virtualidad [Academia and Virtuality Journal]*, *18*(2), 43–63.

Hinton, G. E., Osindero, S., & Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural Computation*, *18*(7), 1527–1554.

Ide, Y., Nohejl, A., Tanner, J., Yanaka, H., Lindsay, C., & Watanabe, T. (2026). Towards automated lexicography: Generating and evaluating definitions for learner's dictionaries. *arXiv*, *2601.01842*, 15 pp.

Întorsureanu, I., Oprea, S.-V., Bâra, A., & Vespan, D. (2025). Generative AI in education: Perspectives through an academic lens. *Electronics*, *14*(1053), 26 pp.

Jarošová, A. (2024). Staronové výzvy (výkladovej) lexikografie v slovensko-českom kontexte [Old challenges of (interpretive) lexicography in the Slovak-Czech context]. *Jazykovědné Aktuality [Linguistics News]*, *61*(1), 38–43.

Joshi, P. A. (2026). The digital shift: The role of Artificial Intelligence in lexicography and corpus linguistics. *Knowledgeable Research*, *5*(1), 31–33.

Jürviste, M., & Jakobson, J. (2025). Vision-enabled LLMs in historical lexicography: Digitising and enriching Estonian-German dictionaries from the 17th and 18th centuries. *arXiv*, *2510.07931*, 20 pp.

Karpova, O. [Карпова, О.]. (2025). От печатных к цифровым англоязычным словарям [From printed to digital English dictionaries]. In V. A. Ikonnikova, V. O. Fedorovskaya, & A. D. Mezentsova (Eds.), *Язык. Культура. Перевод: от живой речи до цифровых моделей. Сборник научных трудов [Language. Culture. Translation: From live speech to digital models. Proceedings]* (pp. 33–41). Rus-Science.

Karpova, O. [Карпова. О.], & Manik, S. A. [Маник, С. А.]. (2026). *Современная Английская Лексикография и Корпусные Технологии. Учебное Пособие [Modern English Lexicography and Corpus Technologies. Study Guide]*. Ivanovo State University Publishing House.

Khalilova, M. A. q. (2025). Interactive AI based dictionaries: New horizons in lexicography and language learning. *Journal of New Century Innovations*, *86*(1), 7–11.

Kompara Lukanc̆ic̆, M. (2026). The automatic recognition of Italian abbreviations: Analysis of Krajšavar and Copilot. *International Journal of Lexicography*, *39*, 16 pp.

Komuro, Y. (2025). The golden age of English-Japanese learners' dictionaries and their lexicographers: Mentorship, innovation, and academic networks. *Lexicography: Journal of ASIALEX*, *12*(1), 21–38.

Kosem, I. (2025). Implementing AI in lexicographic workflow: Challenges and opportunities. In S. Bosch, A. du Plessis, & L. Morris (Eds.), *African Association for Lexicography, 29th International Conference, 2 – 5 July 2025, Abstracts* (pp. 10–11). Afrilex.

Kosem, I., & Arhar Holdt, Š. (2025). Using Large Language Models to generate distractors for language games. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda, & A. Salgado (Eds.),

Electronic Lexicography in the 21st Century (eLex 2025): Intelligent Lexicography. Proceedings of the eLex 2025 conference (pp. 620–635). Lexical Computing.

Kotatkova, A. (2025). Desxifrar sigles mèdiques en català amb ChatGPT, Gemini i Copilot: una anàlisi comparative/Deciphering medical acronyms in Catalan with ChatGPT, Gemini, and Copilot: A comparative analysis/Descifrar siglas médicas en catalán con ChatGPT, Gemini y Copilot: un análisis comparativo. *BiD: textos universitaris de biblioteconomia i documentació [BiD: University texts on library science and documentation]*, *54*(2), 12 pp.

Kouassi, K. J.-M. (2025). Profilex: Kognitive Leistungen des Lexikographen und die Qualität im Wörterbuchartikel [Profilex: Cognitive performance of the lexicographer and the quality of the dictionary entry]. In I. Dias, R. H. Gouws, A. Lobenstein-Reichmann, & S. J. Schierholz (Eds.), *Regarding the Significance of Lexicography/Zur Bedeutsamkeit der Lexikographie: Europe and Australia/Europa und Australien* (pp. 259–272). W. de Gruyter.

Laffey, D. (2025). Examining the effectiveness of Generative AI as a bilingual dictionary in Korean EFL contexts. *Studies in English Education*, *30*(3), 139–165.

Laugesen, A. (2025). Australian lexicography and the Australian story: Historical perspectives and the current scene. In I. Dias, R. H. Gouws, A. Lobenstein-Reichmann, & S. J. Schierholz (Eds.), *Regarding the Significance of Lexicography/Zur Bedeutsamkeit der Lexikographie: Europe and Australia/Europa und Australien* (pp. 15–30). W. de Gruyter.

Law, L. (2024). Application of generative artificial intelligence (GenAI) in language teaching and learning: A scoping literature review. *Computers and Education Open*, *6*(100174), 13 pp.

Lewandowska-Tomaszczyk, B., & Pawłowsk, G. (2025). Testing ChatGPT on terminology generation, definitions: Translation, and ontology creation in German, English and Polish. *Research in Language*, *23*, 320–340.

Liu, R. (2026). Evaluating ChatGPT's role in Chinese learner's dictionary compilation: A comparative analysis. *Chinese Journal of Applied Linguistics*, *49*(2), 297–319.

Liu, R., Chen, X., & Xu, Y. (2025). Beyond replacement: How Large Language Models influence dictionary usage patterns among Chinese English learners. *International Journal of Lexicography*, *38*(4), 342–364.

Łukasik, M. (2024). Artificial Intelligence for the Humanities. Fundamental Concepts Explained. In D. Kalinowski & P. Toczyński (Eds.), *Ekologia w dyskursie. Sztuczna inteligencja [Ecology in Discourse. Artificial Intelligence]* (pp. 15–40). Pomeranian University Publishing House in Słupsk.

Markov, A. A. [Ма́рков, А. А.]. (1906). Распространение закона больших чисел на величины, зависящие друг от друга [Extension of the law of large numbers to dependent quantities]. *Известия Физико-математического общества при Казанском университете [Proceedings of the Physical and Mathematical Society at Kazan University]*, *2;я серия, том 15 [2nd series, vol. 15]*, 135–156.

Marković, A., & Stanković, R. (2025a). Primena velikih jezičkih modela u srpskoj opisnoj leksikografiji – studija slučaja [Case study on application of Large Language Models in Serbian descriptive lexicography]. In M. Dinić Marinković, N. Panić Cerovski, & B. Kovačević (Eds.), *Proceedings of The Eighth International Congress of Applied Linguistics Today, 23–25 May 2025, Belgrade* (pp. 193–219). Faculty of Philology, University of Belgrade.

Marković, A., & Stanković, R. (2025b). So close but still far: Case study on application of LLMs in idioms identification, definition, and generation of illustrative examples. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda, & A. Salgado (Eds.), *Electronic Lexicography in the 21st Century (eLex 2025): Intelligent Lexicography. Proceedings of the eLex 2025 conference* (pp. 79–94). Lexical Computing.

Matyushin, A. (2025). *LLMs as tools for drafting ad hoc pharmaceutical glossaries*. Paper presented at the 1st International Workshop on Language and Language Models, Hungarian Linguistic Research Center, Budapest, 20–21 November 2025,

Meng, L., Liu, M., Wang, H., Cheng, Y., Yang, Q., & Mohanmmed, I. (2026). Building from scratch: A multi-agent framework with human-in-the-loop for multilingual legal terminology mapping. *Artificial Intelligence and Law*, *34*, 40 pp.

Menozzi, M. (2024). *Gli LLM come Possibile Strumento di Semplificazione del Linguaggio Burocratico Italiano e di Generazione di Testi Istituzionali [LLMs as a Potential Tool for Simplifying Italian Bureaucratic Language and Generating Institutional Texts]* [MA, Digital Communication, University of Pavia]. Pavia.

Meterc, M., & Jakop, N. (2025). Identifying the most representative phraseological units using language corpora and Artificial Intelligence for lexicography: The case of Slovenian comparative phrasemes. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda, & A. Salgado (Eds.), *Electronic Lexicography in the 21st Century (eLex 2025): Intelligent Lexicography. Proceedings of the eLex 2025 conference* (pp. 795–814). Lexical Computing.

Meterc, M., & Mrvič, R. (2025). The best known and frequently used Slovene proverbs according to ChatGPT-4o: Exploring the potential for an AI-based paremiological minimum. *Linguistica*, *65*(1), 127–143.

Moriguchi, M., & Kennedy, O. (2024). Generative AI's non-verifiable interpretation in exploring Japanese *youkai* in English. In A. Inoue, N. Kawamoto, & M. Sumiyoshi (Eds.), *AsiaLex 2024 Proceedings. Asian Lexicography – Merging Cutting-edge and Established Approaches* (pp. 99–105). Department of Economics, Toyo University.

Mrdak-Mićović, M., & Jovović, N. (2025). Reconceptualising contemporary lexicography: Emotional intelligence and the sustainable lexicographic modelling of the modal verb *moći*. *Lexikos*, *35*(2), 143–164.

Nam, K., Lee, S., & Jung, H.-Y. (2025). Detection and description of neologisms in Korean lexicography: Methodological issues in corpus balance, word unit bias and LLM assistance. *Lexikos*, *35*, 414–438.

Nazar, R. (2026). Semi-automatic creation of terminological databases. In C. Roche (Ed.), *Proceedings of the TOTh 2023 Conference. Terminology & Ontology: Theories and Applications, 1-2 June 2023* (pp. 117–134). Presses universitaires Savoie Mont Blanc.

Nguyen, T. K. C. (2024). *English language teaching and learning assistant system* [University of Danang - University of Science and Technology]. Da Nang City.

Nordström, M. (2026). *Stöd av lärande inom objektorienterad programmering med hjälp av generativ AI [Supporting learning in object-oriented programming using generative AI]: Chain-of-Thought prompting* [BSc, University of Gävle]. Gävle.

Nycz, K., & Tęcza, Z. (2025). Aktuelle Aussprachewörterbücher des Deutschen und des Englischen im Vergleich [Current pronunciation dictionaries of German and English compared]. *Lexikos*, *35*(2), 232–255.

Odendaal, G., & du Plessis, A. H. (2025). Into the next century: The *Woordeboek van die Afrikaanse Taal* and Artificial Intelligence. In S. Bosch, A. du Plessis, & L. Morris (Eds.), *African Association for Lexicography, 29th International Conference, 2 – 5 July 2025, Abstracts* (pp. 57–58). Afrilex.

Onieva Palomar, R. (2025). *Hacia un nuevo diccionario monolingüe en ELE [Towards a New Monolingual Dictionary in Spanish as a Foreign Language]* [PhD, Universidad Carlos III de Madrid]. Madrid.

Ooi, V. B. Y. (2024). Querying and challenging the 'Generative AI lexicographer' for lexical information. In N. Oco, S. N. Dita, A. M. Borlongan, & J.-B. Kim (Eds.), *Proceedings of the 38th Pacific Asia Conference on Language, Information and Computation* (pp. xiii-xiv). Study of Language and Information.

Owan, V. J., Mohammed, I. A., Bello, A., & Shittu, T. A. (2025). Higher education students' ChatGPT use behavior: Structural equation modelling of contributing factors through a modified UTAUT model. *Contemporary Educational Technology*, *17*(4), 22 pp.

Özcan, E., Erkoç, M. F., & Tokatlı, H. (2026). Machine-readable multilingual dictionaries: A semi-automatic ISO–TEI model. *Mersin Üniversitesi Dil ve Edebiyat Dergisi [Mersin University Journal of Language and Literature], MEUDED*, *22*(1), 1–20.

Palmieri, C. (2025). Vocabolari d'autore e intelligenza artificiale: è possibile creare il vocabolario della lingua di Boccaccio con ChatGPT? [Author dictionaries and artificial intelligence: Is it possible to create a Boccaccio vocabulary with ChatGPT?]. *AI-Linguistica. Linguistic Studies on AI-Generated Texts and Discourses*, *2*(2), 15 pp.

Patat, E. (2025). İnsan Çevirmen ve Makine Çevirisi. Türkçe-İtalyanca Sözlükbiliminde Bir Vaka Çalışması [Human translator and Machine Translation: A case study in Turkish-Italian lexicography]. *International Journal of Language and Translation Studies*, *5*(2), 281–307.

Pecman, M. (2025). Drafting definitions for emerging concepts and terms undergoing semantic shift within the ARTES knowledge base: A protocol for integrating LLMs into terminological analysis by experimental approach. *Terminologija/Terminology*, *32*, 43 pp.

Perdih, A., Gabrovšek, D., & Ježovnik, J. (2025). Image-to-sense alignment using AI tools. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda, & A. Salgado (Eds.), *Electronic Lexicography in the 21st Century (eLex 2025): Intelligent Lexicography. Proceedings of the eLex 2025 conference* (pp. 852–865). Lexical Computing.

Proroković, J. (2025). Educational implications of lexicographic misalignment: AI-generated definitions of English loanwords in L2 context. *Magister*, *13*(2), 39–62.

Rey Cortés, M. d. D. N. (2025). *Propuesta de elaboración de un diccionario de falsos amigos español-portugués [Proposal for the development of a Spanish-Portuguese dictionary of false friends]* [MA, Universidad Nacional de Educación a Distancia]. Madrid.

Risberg, L., Aedmaa, E., Tuulik, M., Langemets, M., Vainik, E., Prangel, E., Koppel, K., & Pook, H. (2025). The role of subjectivity in lexicography: Experiments towards data-driven labeling of informality. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda, & A. Salgado (Eds.), *Electronic Lexicography in the 21st Century (eLex 2025): Intelligent Lexicography. Proceedings of the eLex 2025 conference* (pp. 332–352). Lexical Computing.

Risberg, L., Tuulik, M., Langemets, M., Koppel, K., Vainik, E., Prangel, E., & Aedmaa, E. (2025). Keelekorpus kui leksikograafi abiline kõnekeelsuse tuvastamisel [Language corpus as an assistant to the lexicographer in identifying colloquialisms]. *Keel ja Kirjandus [Language and Literature]*, *68*(7), 605–624.

Risberg, L., Tuulik, M., Langemets, M., Vainik, E., Aedmaa, E., Prangel, E., Koppel, K., Pook, H., & Zupping, S. (2025). *(De)standardiseerides kõnekeelt sõnaraamatus. Kuidas rakendada keelekorpust ja suuri keelemudeleid? [(De)standardising spoken language in a dictionary. How to apply a language corpus and Large Language Models?]*. Paper presented at the 22nd Spring Conference of the Estonian Applied Linguistics Association, April 24, 2025, Institute of the Estonian Language. https://www.youtube.com/watch?v=eAUjUVUIi_s

Rundell, M., Jakubíček, M., Kovář, V., Matuška, O., & Cukr, M. (2025). Lexicom at 25: Reflections on the changing world of lexicography and language technology. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda, & A. Salgado (Eds.), *Electronic Lexicography in the 21st Century (eLex 2025): Intelligent Lexicography. Proceedings of the eLex 2025 conference* (pp. 139–148). Lexical Computing.

San Martín, A. (2026). Terminological definitions in the face of generative Artificial Intelligence. In A. San Martín, P. Faber, B. Sánchez Cárdenas, F. Vezzani, & G. Di Nunzio (Eds.), *New Directions in Digital Terminology Research* (pp. 184–221). Brill.

Santamaría-Pérez, M. I., & Martínez Sánchez, O. (2026). Rúbrica para la evaluación de definiciones terminológicas generadas por modelos de lenguaje: diseño, aplicación y transferibilidad a otros ámbitos especializados [Rubric for the evaluation of terminological definitions generated by language models: Design, application and transferability to other specialized fields]. *NeoTermed 4 All*, 16 pp.

Satoła-Staśkowiak, J. (2023). O stałej marginalizacji słowników drukowanych jako nieodwracalnym zjawisku we współczesnej leksykografii; [On the constant marginalization of printed dictionaries as an irreversible phenomenon in contemporary lexicography]. *Językoznawstwo [Linguistics]*(nr specjalny [Special Issue]), 31–41.

Scott, M. (1996-2026). *WordSmith Tools, version 9*. https://www.lexically.net/wordsmith/

Stöckle, P., Elsner, D., Koppensteiner, W., & Korecky-Kröll, K. (2025). LLM-assisted dialect lexicography: Challenges and opportunities in processing historical Bavarian dialects. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda, & A. Salgado (Eds.), *Electronic Lexicography in the 21st*

Century (eLex 2025): Intelligent Lexicography. Proceedings of the eLex 2025 conference (pp. 453–475). Lexical Computing.

Štrkalj Despot, K. (2025). Leksikografija u vremenu umjetne inteligencije: trendovi, projekti i infrastructure [Lexicography in the age of Artificial Intelligence: Trends, projects and infrastructures]. In L. Hudeček, Ž. Jozić, M. Mihaljević, K. Lewis, M. Matijević, & D. Cakić (Eds.), *Hrvatski mrežni rječnik – Mrežnik: 2. faza [Croatian Online Dictionary – Mrežnik: Phase 2], Zagreb, 27 November 2025* (pp. 2–3). Institute of the Croatian Language.

Sudnik, D. M. (2025). Возможности платформы tiktok как «естественного» лингвистического корпуса [The potential of the TikTok platform as a "natural" linguistic corpus]. *Cuadernos de Rusística Española [Notebooks of Spanish Russian Studies]*, *21*, 101–121.

Tamba, E. I. (2025). Lexicografie și digitalizare în România – o rela;ție . . . sine qua non; [Lexicography and digitalization in Romania - a connection . . . sine qua non]. *Philologica Jassyensia*, *21*(2), 307–332.

Tanuso, N. A. E., Mujiyanto, Y., Hartono, R., & Pratama, H. (2026). VIPCALL: A design-based method of visual grammar symbols for improving EFL students' fiction writing. *EduLite: Journal of English Education, Literature, and Culture*, *11*(1), 202–216.

Tóth, Á., & Abdelzaher, E. (2025). Improving the lexicographic accessibility of WN through LLMs. In C. Zanchi, L. Brigada Villa, E. Biagetti, A. Rademaker, F. Bond, & G. Rigau (Eds.), *Proceedings of the 13th Global Wordnet Conference* (pp. 142–150). Global Wordnet Association.

Trap-Jensen, L. (2025). Fremtidens ordbøger: udfordringer og tilpasninger [Dictionaries of the future: Challenges and adaptations]. *LexicoNordica*, *32*, 177–196.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind*, *59*(236), 433–460.

Tuulik, M., Vainik, E., Prangel, E., Langemets, M., Aedmaa, E., Koppel, K., & Risberg, L. (2025). Tähenduste seletamine leksikograafias: kuivõrd on abi suurtest keelemudelitest? [Explaining meanings in lexicography: How helpful are Large Language Models?]. *ESUKA (Eesti ja soome-ugri keeleteaduse ajakiri) - JEFUL (Journal of Estonian and Finno-Ugric Linguistics)*, *16*(2), 147–176.

Vakulenko, M. (2025). Lexicographic post-processing and refinement issues in Wordnets: The Ukrainian Wordnet experience. *Lexikos*, *35*(2), 345–361.

Vasilyeva, D. V. [Васильева, Д. В.]. (2025). Использование больших языковых моделей при лексикографировании неологизмов COVID-19 [Using Large Language Models in lexicographic descriptions of COVID-19 neologisms]. In A. A. Malygin [А. А. Малыгин] (Ed.), *Международной Научно Практической Конференции [Proceedings of the International Scientific-Practical Conference]* (pp. 335–345). Ivanovo State University.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 30 (NIPS 2017)* (pp. 6000–6010). Curran A.

Vranjek Ošlak, U. (2025). How effective is AI as a language consultant? In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda, & A. Salgado (Eds.), *Electronic Lexicography in the 21st Century (eLex 2025): Intelligent Lexicography. Proceedings of the eLex 2025 conference* (pp. 499–512). Lexical Computing.

Widmann, T. (2025a). ChatGPT som leksikograf: Ordbogens fremtid? [ChatGPT as a lexicographer: The future of the dictionary?]. In W. Collin, I. Schoonderbeek Hansen, T. T. Hougaard, & K. Schriver (Eds.), *20. Møde om Udforskningen af Dansk Sprog, Århus 2024 [20th Meeting on the Exploration of the Danish Language, Aarhus 2024]* (pp. 427–442). Department of Communication and Culture, Aarhus University.

Widmann, T. (2025b). A pipeline for automated dictionary creation with optional human intervention. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda, & A. Salgado (Eds.), *Electronic Lexicography in the 21st Century (eLex 2025): Intelligent Lexicography. Proceedings of the eLex 2025 conference* (pp. 33–43). Lexical Computing.

Wu, M., & Wang, D. (2026). Automatic compilation of a Pre-Qin philosophy lexicon via Large Language Models. *npj Heritage Science*, *14*(47), 11 pp.

Yang, C., Li, R., & Yang, L. (2025). Revisiting trends in GenAI-assisted second language writing: Retrospect and prospect. *Journal of Educational Computing Research*, *63*(7-8), 45 pp.

Yayla, Y., & Karaca, N. (2025). Sö̈zlük Bilimi ve Üretken Yapay Zekâ: Madde Başlarının Tarifleri Çerçevesinde Bir İnceleme [Lexicography and generative Artificial Intelligence: An analysis within the framework of defining headwords]. *Orun*, *1*, 37–79.

Zhang, H., & Wang, Y. (2026). The role of ChatGPT in assisting artistic creation: An analysis of opportunities and risks from the perspectives of the creative process, benefits, and drawbacks. *Research Square*, *9017599*, 22 pp.

Zhang, J., & Hu, W. (2025). The exemplification pattern and strategies in Chinese-English learners' dictionaries intended for EFL learners in China. *International Journal of Lexicography*, *38*(3), 270–284.

Zhang, S., Zhang, X., Xu, H., & Hu, X. (2026). A bibliometric study of lexicography between 2003 and 2023. *International Journal of Lexicography*, *39*, 28 pp.

Zhang, W., Zavalniuk, I., Bohatko, V., Kukhar, N., & Pavlyuk, O. (2025). Language transformations under the influence of Artificial Intelligence: Linguistic trends and development prospects. *Metaverse Basic and Applied Research*, *4*(161), 9 pp.

Zhong, A. (2025). Prompts for language learners: A practical guide to using DeepSeek as a dictionary. *Lexikos*, *35*(1), 274–285.

Zhong Weiguo [仲为国], Jia Ning [贾宁], Liang Yitian [梁屹天], Tian Xuan [田轩], Ni Chenkai [倪晨凯], Lu Xianghua [卢向华], Cai Fengyan [才凤艳], Xie Xiaoyun [谢小云], Shi Junqi [施俊琦], He Wei [贺伟], Xu Nianxing [许年行], Wang Lin [汪林], Qin Xin [秦昕], He Xianjie [何贤杰], Jiao Hao [焦豪], Zhang Guanglei [张光磊], Pan Yu [潘煜], Wang Yiwen [王益文], & Liu Zuoyi [刘作仪]. (2025). 人工智能时代的工商管理研究现状与未来 [Business administration research in the age of Artificial Intelligence — An academic review of the business administration development seminar based on Artificial Intelligence technology]. *国家自然科学基金委公报 [Bulletin of the National Natural Science Foundation of China]*, *38*(5), 794–805.

Contact information

Gilles-Maurice de Schryver
Ghent University & University of Pretoria
gillesmaurice.deschryver@ugent.be

Addendum: The second forward snowball sample on GenAI in lexicography

NB. All titles and sources are abbreviated. Non-English titles (41 out of 110 records, or thus 37%) and non-English sources (39 out of 110 records, or thus 35%) were also translated.

Date	Author(s)	Title	Source	Lng Pub	Loc Pub	Lng Study	Discipline	Talk	Art
01/08/2023	**Asmaa (2024)**	Student use of dictionaries in the age of AI & ChatGPT	Jnl Faculty Arts Research, Menoufia U	Arabic	Egypt	Arabic	Lexicography	~	/
25/10/2023	**Alonso-Ramos (2023)**	ChatGPT's role as a lexicographer	Festschrift	Spanish	Spain	Spanish	Lexicography	+	/
01/01/2024	**Nguyen (2024)**	ELT & learning assistant system	PhD	English	Vietnam	English	Education	/	~
20/02/2024	**Łukasik (2024)**	AI for the Humanities	Book chapter	English	Poland	*N.A.*	Artificial Intelligence	/	very +
18/06/2024	**Bozhko (2024)**	Lexicography of connotative proper nouns	Scientific Notes Ntl U 'Ostroh Academy'	Ukrainian	Ukraine	Ukrainian	Lexicography	/	~
02/09/2024	**Jarošová (2024)**	Old challenges of lexicography in Slovak-Czech	Ling News	Slovak	Czechia	Slovak/Czech	Lexicography	+	+
01/11/2024	**Menozzi (2024)**	LLMs for Simplifying Italian Bureaucratic Language	MA	Italian	Italy	Italian	Terminography	/	~
21/11/2024	**Bago (2024)**	Dictionaries and grammars in the age of digital transformation	Abstracts book	Croatian	Croatia	Croatian	Lexicography	/	~
07/12/2024	**Ooi (2024)**	The 'Generative AI lexicographer' for lexical information	Proc 38th Pacific Asia Conf Lang	English	South Korea	(Singapore) English, Japanese	Lexicography	/	~
18/12/2024	**Castro-Prieto (2024)**	Digital humanities & AI in Hispanic linguistics	Book chapter	Spanish	Spain	Spanish	Linguistics	/	+
27/03/2025	**Zhong Weiguo et al. (2025)**	Business administration research in the age of AI	Bulletin Ntl Natural Science Foundation China	Chinese	China	*N.A.*	Bibliometrics: Business Administration	/	~
01/05/2025	**Risberg, Tuulik, Langemets, Vainik, et al. (2025)**	How to apply a language corpus and LLMs?	PPT, Spring Conf Estonian Appl Ling Assoc 2025	Estonian	Estonia	Estonian	Lexicography	/	~
04/06/2025	**Zhong (2025)**	Using DeepSeek as a dictionary	Lexikos	English	South Africa	English	Lexicography	/	+
09/06/2025	**Zhang et al. (2025)**	Language transformations under the influence of AI	Metaverse Basic & Appl Research	English	Argentina	English, Ukrainian	Linguistics	~	/
10/06/2025	**Nam et al. (2025)**	Neologisms in Korean lexicography with LLM assistance	Lexikos	English	South Africa	Korean	Lexicography	/	~
16/06/2025	**Balenović and Proroković (2025)**	Lexicographic potential of AI for English loanwords in Croatian	Studia Lexicographica	English	Croatia	English > Croatian	Lexicography	/	+
24/06/2025	**Evert et al. (2025)**	Evaluating AI-generated dictionary entries	Book chapter	German	Germany	German	Metalexicography	/	+
01/07/2025	**Bothma (2025)**	GenAI as information tool for lexicographic information needs	Abstracts Afrilex 2025	English	South Africa	English	Information Science	/	~
01/07/2025	**Clemente González (2025)**	ChatGPT and Gemini as	MA	Spanish	Chile	Spanish	Terminography	/	+

Date	Author(s)	Title	Source	Lng Pub	Loc Pub	Lng Study	Discipline	Talk	Art
		Terminology Extractors							
01/07/2025	**Kosem (2025)**	Implementing AI in lexicographic workflow	Abstracts Afrilex 2025	English	South Africa	Slovene	Lexicography	/	~
01/07/2025	**Odendaal and du Plessis (2025)**	The *Woordeboek van die Afrikaanse Taal* and AI	Abstracts Afrilex 2025	English	South Africa	Afrikaans	Lexicography	/	~
01/07/2025	**Onieva Palomar (2025)**	Towards a New Monolingual Dictionary in Spanish	PhD	Spanish	Spain	Spanish	Lexicography	/	+
01/07/2025	**Risberg, Tuulik, Langemets, Koppel, et al. (2025)**	Corpus as an assistant to the lexicographer for colloquialisms	Lang & Lit	Estonian	Estonia	Estonian	Lexicography	/	~
09/07/2025	**Emakporuena et al. (2025)**	Higher GenAI for educational pedagogy: ChatGPT experiences	Book chapter	English	USA	English	Pedagogy	/	+
14/07/2025	**Kotatkova (2025)**	Medical acronyms in Catalan with ChatGPT, Gemini & Copilot	BiD: U texts lib sci & doc	Catalan	Spain	Catalan	Terminography	/	+
17/07/2025	**Zhang and Hu (2025)**	Chinese-English learners' dictionaries for EFL learners in China	Intl Jnl Lexicography	English	UK	Chinese-English	Lexicography	/	~
29/07/2025	**Al-Mitamy and Mohammed (2025)**	GenAI in scientific publishing: ChatGPT & Gemini/Bard	Book chapter	English	Switzerland	Arabic	Bibliometrics: Scientific Publishing	/	~
05/08/2025	**Komuro (2025)**	English-Japanese learners' dictionaries and their lexicographer	Lexicography: Jnl Asialex	English	Canada	English-Japanese	Lexicography	/	+
05/08/2025	**Moriguchi and Kennedy (2024)**	GenAI non-verifiable interpretation: Japanese *youkai* in English	Lexicography: Jnl Asialex	English	Canada	Japanese-English	Terminography	/	+
09/08/2025	**Herrera Pérez et al. (2025)**	Chatbots & AI in secondary and middle education	Academia & Virtuality Jnl	Spanish	Colombia	Spanish	Education	/	+
12/08/2025	**Yang et al. (2025)**	GenAI-assisted second language writing	Jnl Edu Computing Research	English	USA	*N.A.*	Bibliometrics: Second Language Writing	/	~
14/08/2025	**Fuertes-Olivera and Tarp (2025)**	Innovative lexicography in the GenAI age	Intl Jnl Foreign Lang	English	Spain	Spanish	Lexicography	~	~
03/09/2025	**Liu et al. (2025)**	How LLMs influence dictionary usage patterns among Chinese	Intl Jnl Lexicography	English	UK	English 4 Chinese	Lexicography	/	very +
10/09/2025	**Wanjing and de Schryver (2025)**	DeepSeek as the ultimate translator and lexicographer	Proc Asialex 2025	English	China	Chinese	Lexicography	~	~
15/09/2025	**Tóth and Abdelzaher (2025)**	Improving the lexicographic accessibility of WN through LLMs	Proc Global Wordnet 2025	English	Italy	English	Lexicography	+	/
22/09/2025	**Denisová (2025)**	Automatic Methods for Compiling Bilingual Dictionaries	PhD	English	Czechia	English-Slovak, English-Estonian, English(-Korean)	NLP	+	+
29/09/2025	**Heinisch (2025)**	Transforming terminology work with LLMs	Jnl Translation & Interpreting Studies	English	Hungary	*N.A.*	Terminography	/	~

Date	Author(s)	Title	Source	Lng Pub	Loc Pub	Lng Study	Discipline	Talk	Art
29/09/2025	**Khalilova (2025)**	Interactive AI based dictionaries	Jnl New Century Innovations	English	Uzbekistan	*N.A.*	Lexicography	/	~
30/09/2025	**Laffey (2025)**	GenAI as a bilingual dictionary in Korean EFL contexts	Studies Eng Edu	English	South Korea	English-Korean	Lexicography	/	+
30/09/2025	**Mrdak-Mićović and Jovović (2025)**	Reconceptualising contemporary lexicography	Lexikos	English	South Africa	Montenegrin	Lexicography	/	~
03/10/2025	**Nycz and Tęcza (2025)**	Current pronunciation dictionaries of German and English	Lexikos	German	South Africa	German, English	Lexicography	/	~
06/10/2025	**Rey Cortés (2025)**	A Spanish-Portuguese dictionary of false friends	MA	Spanish	Spain	Spanish > Portuguese	Lexicography	/	+
08/10/2025	**Owan et al. (2025)**	Higher education students' ChatGPT use behavior	Contemporary Edu Tech	English	Cyprus	*N.A.*	Education	/	~
09/10/2025	**Jürviste and Jakobson (2025)**	Vision-enabled LLMs for 17-18th c. Estonian-German dictionaries	arXiv	English	USA	historical Estonian-German	Historical Lexicography	+	/
14/10/2025	**Tuulik et al. (2025)**	Explaining meanings in lexicography: How helpful are LLMs?	Jnl Estonian & Finno-Ugric Ling	Estonian	Estonia	Estonian	Lexicography	~	~
22/10/2025	**Karpova (2025)**	From printed to digital English dictionaries	Proc Lang Cult Translation	Russian	Russia	English	Lexicography	/	+
23/10/2025	**Caruso and De Meo (2025)**	Digital phraseological dictionaries for second languages	Italian Review Appl Ling	Italian	Italy	Various	Lexicography	/	~
04/11/2025	**Bothma and Gouws (2025)**	GenAI as information tool for lexicographic information needs	Lexikos	English	South Africa	English	Information Science	/	~
11/11/2025	**Meng et al. (2026)**	Multilingual legal terminology mapping	AI & Law	English	Netherlands	Chinese, Japanese	Terminography	/	~
17/11/2025	**Frankenberg-Garcia (2025)**	AI-powered dictionary designed for on-screen reading	Proc eLex 2025	English	Czechia	English	Lexicography	/	~
17/11/2025	**Kosem and Arhar Holdt (2025)**	Using LLMs to generate distractors for language games	Proc eLex 2025	English	Czechia	Slovene	Lexicography	~	+
17/11/2025	**Marković and Stanković (2025b)**	LLMs in idioms identification, definition & illustrative examples	Proc eLex 2025	English	Czechia	Serbian	Lexicography	~	/
17/11/2025	**Meterc and Jakop (2025)**	AI for lexicography: Slovenian comparative phrasemes	Proc eLex 2025	English	Czechia	Slovene	Lexicography	/	~
17/11/2025	**Perdih et al. (2025)**	Image-to-sense alignment using AI tools	Proc eLex 2025	English	Czechia	Slovene	Lexicography	/	~
17/11/2025	**Risberg, Aedmaa, et al. (2025)**	Subjectivity in lexicography: data-driven labeling of informality	Proc eLex 2025	English	Czechia	Estonian	Lexicography	/	~
17/11/2025	**Rundell et al. (2025)**	Lexicom at 25: lexicography and language technology	Proc eLex 2025	English	Czechia	*N.A.*	Lexicography	/	~

Date	Author(s)	Title	Source	Lng Pub	Loc Pub	Lng Study	Discipline	Talk	Art
17/11/2025	**Stöckle et al. (2025)**	LLM-assisted dialect lexicography: historical Bavarian dialects	Proc eLex 2025	English	Czechia	historical Bavarian	Dialectology	/	~
17/11/2025	**Vranjek Ošlak (2025)**	How effective is AI as a language consultant?	Proc eLex 2025	English	Czechia	Slovene	Applied Linguistics	/	~
17/11/2025	**Widmann (2025b)**	Automated dictionary creation with optional human intervention	Proc eLex 2025	English	Czechia	Danish	Lexicography	~	/
19/11/2025	**Matyushin (2025)**	LLMs as tools for drafting ad hoc pharmaceutical glossaries	Workshop Lang & Lang Models	English	Hungary	Russian-English	Terminography	/	~
24/11/2025	**Vakulenko (2025)**	Lexicographic post-processing: The Ukrainian Wordnet experience	Lexikos	English	South Africa	English, German, Ukrainian	Lexicography	/	+
25/11/2025	**Widmann (2025a)**	ChatGPT as a lexicographer: The future of the dictionary?	Meeting Exploration Danish Lang 2024	Danish	Denmark	Danish	Lexicography	~	/
27/11/2025	**Štrkalj Despot (2025)**	Lexicography in the age of Artificial Intelligence	Abstracts Word by word, dictionary by dictionary	English	Croatia	*N.A.*	Lexicography	/	~
04/12/2025	**Arias-Arias et al. (2025)**	Is Artificial Intelligence killing dictionaries?	Book chapter	English	Germany	German, Galician, Spanish	Lexicography	~	+
04/12/2025	**Kouassi (2025)**	Profilex: Cognitive performance of the lexicographer	Book chapter	German	Germany	*N.A.*	Metalexicography	~	/
04/12/2025	**Laugesen (2025)**	Australian lexicography and the Australian story	Book chapter	English	Germany	Australian English	Lexicography	/	+
05/12/2025	**Dziemianko (2025)**	Warnings in learners' dictionaries: Do they help?	GEMA Jnl Lang Studies	English	Malaysia	English	Lexicography	/	+
09/12/2025	**de Schryver (2025)**	Out-of-the-box GPTs for lexicography	Lexikos	English	South Africa	English, Portuguese, Swahili, Xhosa, Lusoga	Lexicography	/	+
13/12/2025	**Palmieri (2025)**	Author dictionaries and AI: Boccaccio vocabulary with ChatGPT	AI-Linguistica	Italian	Germany	historical Italian	Historical Lexicography	/	~
13/12/2025	**Pecman (2025)**	ARTES: Integrating LLMs into terminological analysis	Terminology	English	Lithuania	50+	Terminography	/	+
15/12/2025	**Trap-Jensen (2025)**	Dictionaries of the future	LexicoNordica	Danish	Denmark	Danish	Lexicography	+	+
19/12/2025	**Satoła-Staśkowiak (2023)**	Marginalization of printed dictionaries, an irreversible phenomenon	Linguistics	Polish	Poland	Polish	Lexicography	/	~
22/12/2025	**Arhar Holdt et al. (2025)**	ChatGPT in the development of the Thesaurus of Modern Slovene	Contributions Recent History	Slovene	Slovenia	Slovene	Lexicography	/	~
22/12/2025	**Sudnik (2025)**	The TikTok platform as a "natural" linguistic corpus	Notebooks Spanish Russian Studies	Russian	Spain	Russian	Media Studies	/	+

Date	Author(s)	Title	Source	Lng Pub	Loc Pub	Lng Study	Discipline	Talk	Art
24/12/2025	**Akalin (2024)**	Creating an AI-powered Turkish dictionary	Virtual Jnl Turkish Studies	Turkish	Turkey	Turkish	Lexicography	/	~
24/12/2025	**Barik and Ed-Dali (2025)**	Shakespearean corpus in ChatGPT's repertoire	Lang & Lit Studies	English	Morocco	English	Literature	/	~
24/12/2025	**Meterc and Mrvič (2025)**	Slovene proverbs according to ChatGPT-4o	Linguistica	English	Slovenia	Slovene	Lexicography	/	very +
29/12/2025	**Marković and Stanković (2025a)**	LLMs in Serbian descriptive lexicography	Proc Appl Ling Today	Serbian	Serbia	Serbian	Lexicography	+	/
30/12/2025	**Lewandowska-Tomaszczyk and Pawłowsk (2025)**	ChatGPT on terminology generation in German, English & Polish	Research Lang	English	Poland	German, English, Polish	Terminography	/	~
30/12/2025	**Patat (2025)**	Human translator and Machine Translation in Turkish-Italian	Intl Jnl Lang & Translation Studies	English	Turkey	Turkish > Italian	Translation	/	+
05/01/2026	**Ide et al. (2026)**	Automated lexicography for learner's dictionaries	arXiv	English	USA	English	Lexicography	/	+
14/01/2026	**Abdelzaher and Rafaee (2026)**	Linguistic profile of LLMs: performance & jailbreak vulnerability	Jnl Cult Cognitive Science	English	UK	*N.A.*	Cognitive Linguistics	/	~
21/01/2026	**Wu and Wang (2026)**	Automatic compilation of a Pre-Qin philosophy lexicon via LLMs	npj Heritage Science	English	UK	Chinese	Historical Lexicography	/	~
23/01/2026	**Zhang et al. (2026)**	Bibliometric study of lexicography between 2003-2023	Intl Jnl Lexicography	English	UK	*N.A.*	Bibliometrics: Lexicography	/	~
27/01/2026	**Vasilyeva (2025)**	Using LLMs in lexicographic descriptions of COVID-19 neologisms	Proc Scientific-Practical Conf	Russian	Russia	Russian	Lexicography	/	+
30/01/2026	**Hall et al. (2026)**	Lexicography	Handbook chapter	English	UK	*N.A.*	Lexicography	/	~
30/01/2026	**Joshi (2026)**	The role of AI in lexicography and corpus linguistics	Knowledgeable Research	English	India	*N.A.*	NLP	/	~
31/01/2026	**Tamba (2025)**	Lexicography and digitalization in Romania	Philologica Jassyensia	Romanian	Romania	Romanian	Lexicography	~	~
31/01/2026	**Yayla and Karaca (2025)**	Lexicography and GenAI: defining headwords	Orun	Turkish	Turkey	Turkish	Lexicography	/	+
10/02/2026	**Nazar (2026)**	Semi-automatic creation of terminological databases	Proc TOTh 2023	English	France	English, Spanish	Terminography	+	/
11/02/2026	**Han and de Schryver (2026)**	Reference work perceptions by Chinese users in the era of GenAI	Intl Jnl Lexicography	English	UK	Chinese	Lexicography	/	~
13/02/2026	**Nordström (2026)**	Supporting learning in object-oriented programming using GenAI	BSc	Swedish	Sweden	Swedish	NLP	/	~
23/02/2026	**Tanuso et al. (2026)**	Visual grammar symbols for improving EFL students' fiction writing	EduLite: Jnl Eng Edu, Lit & Cult	English	Indonesia	English	Education	/	~

Date	Author(s)	Title	Source	Lng Pub	Loc Pub	Lng Study	Discipline	Talk	Art
27/02/2026	**Özcan et al. (2026)**	Machine-readable multilingual dictionaries	Jnl Lang & Lit, Mersin U	English	Turkey	Turkish > English, French, German, Italian	NLP	/	+
04/03/2026	**Zhang and Wang (2026)**	The role of ChatGPT in assisting artistic creation	Research Square	English	USA	*N.A.*	Art	/	~
05/03/2026	**Eiche et al. (2025)**	Applying AI in Estonian national defense terminology	Yearbook Estonian Appl Ling Assoc	Estonian	Estonia	Estonian	Terminography	/	~
09/03/2026	**Ganslmayer (2026)**	AI-supported vocabulary work for German language teaching	German Lang Teaching	German	Germany	German	Education	/	+
17/03/2026	**Kompara Lukančič (2026)**	Italian abbreviations: Analysis of Krajšavar and Copilot	Intl Jnl Lexicography	English	UK	Italian	Lexicography	/	~
20/03/2026	**Filipović Petrović and Beliga (2025)**	Can AI understand Croatian idioms? LLMs in lexicographic tasks	Contributions Recent History	Slovene	Slovenia	Croatian	Lexicography	/	~
25/03/2026	**Eltanbouly and Rashwani (2026)**	Grounding Arabic LLMs in the *Doha Historical Dictionary*	arXiv	English	USA	Arabic	Historical Lexicography	/	~
27/03/2026	**Bulatova and Musiyako (2025)**	GenAI in the creation of special term glossaries	All-Russian Winter School Pedagogical Excellence 2025	Russian	Russia	Russian	Terminography	/	~
01/04/2026	**Chambat and Rey (2026)**	From reverse engineering to networking: Dictionaries	French Lang	French	France	French	Lexicography	/	~
20/04/2026	**San Martín (2026)**	Terminological definitions in the face of GenAI	Book chapter	English	Netherlands	English	Terminography	/	+
03/05/2026	**Azadravesh and Boubiche (2026)**	The dark side of generative AI	Book chapter	English	USA	*N.A.*	Cybersecurity	/	~
03/05/2026	**Karpova and Manik (2026)**	Modern English Lexicography and Corpus Technologies	Handbook	Russian	Russia	English	Lexicography	+	+
19/05/2026	**Arias-Arias (2026)**	Semantic-lexicographic aspects of deverbal *-ung-* in German	Lexikos	German	South Africa	German	Lexicography	/	+
20/05/2026	**Proroković (2025)**	AI-generated definitions of English loanwords in L2 context	Magister	English	Czechia	English > Croatian	Lexicography	/	+
25/05/2026	**Atzeni and Babonnaud (2025)**	GenAIs for specialised translation: UN reports on violence	Book chapter	English	Belgium	English	Translation	/	~
25/05/2026	**Liu (2026)**	Evaluating ChatGPT's role in Chinese learner's dictionary	Chinese Jnl Appl Ling	English	China	Chinese	Lexicography	/	+
27/05/2026	**Santamaría-Pérez and Martínez Sánchez (2026)**	Terminological definitions generated by LLMs	NeoTermed 4 All	Spanish	Spain	Spanish	Terminography	/	+

Gilles-Maurice de Schryver

THE NINTH ART IN LEXICOGRAPHY

The Power of Comic Strips

Abstract This paper presents the background to a book on the visual history of academic lexicography, being a sequence of hundreds of one-page comic strips, created through conversations with LLM chatbots and GenAI image generators, each page representing a major publication in the field. Of course, I knew that this submission would be a gamble: Would a comic version of a conference paper be taken seriously? Well, one extremely positive adjudicator gave it a 'Strong Accept', and a more critical one gave it an 'Accept'. Adjudicator 1 summarised it head-on, writing that it is "an extremely original contribution to the field, as it addresses the impact and outreach of scientific research both within and beyond academia", while Adjudicator 2 pointed out that it is "untraditional but entertaining, and comic strips could very well be part of lexicography". In what follows, my original extended abstract is presented first, after which I interact with the feedback left by Adjudicators 1 and 2. No doubt, and again, a very uncommon approach, but times have changed in the era of generative AI.

Keywords Ninth art; comic strips; intersemiotic translation; lexicography; bibliometrics; artificial intelligence (AI); generative AI (GenAI); large language model (LLM); Gemini; ChatGPT; Academia.edu

1 The Ninth Art in academia

It is well known that Belgium is the capital of the Ninth Art — in other words the 'Comic Strip'[1] — with the Belgian Comic Strip Centre[2] in a magnificent building designed by the Art Nouveau architect Victor Horta (°Ghent, 1861 – †Brussels, 1947).[3] In the current era of GenAI, where machines generate anything and everything, I decided to attach two noticeboards on the outside wall of my office at Ghent University, where I put up the first page of some of my recent articles in lexicography, supplemented with 'the full text turned into one-page comic strips by GenAI'. My online CV lists all of that too. Does this have anything to do with academia or the field of lexicography? It does.

2 Gemini the lexicographer

We are not only living in times of GenAI, we are also under an avalanche of incessant text messages, non-stop bleeps and bloops, and perpetual one-minute shorts, day in day out, night in night out. How each of us deals with it is one thing, but how consumers navigate this is another entirely. To help the takers, what can deliverers do? Yes, we gather at Euralex where over a hundred papers are presented in 2026, but who of us also reads the one thousand pages on them in the proceedings? And for those of us who also attend the Afrilex, Americalex-S, Asialex, Australex, DSNA and eLex conferences, are they still able to 'consume the deluge of information'? When googling for the last five words, 'Gemini the Lexicographer'[4] returns:

> Consuming the daily deluge of information requires treating digital intake like a diet: prioritizing quality over quantity through curation, setting strict boundaries, and

1 See https://en.wikipedia.org/wiki/Bande_dessin%C3%A9e
2 See https://en.wikipedia.org/wiki/Belgian_Comic_Strip_Center
3 See https://en.wikipedia.org/wiki/Victor_Horta
4 See https://www.google.com/

> practicing digital minimalism. To avoid cognitive overload, filter content from reliable sources, use tools to batch information, and disable non-essential notifications.

Accepted, but that shifts the burden onto the consumer. In this day and age, 'each of us wants a snap-view', which Gemini returns as:

> Based on the context of Snapchat[5] and social media usage, the phrase "each of us wants a snap-view" generally implies that everyone involved wants a quick, visual, and often fleeting glimpse into a specific moment, event, or shared experience.

That is exactly what this Belgian has, and those who share his office have, experimented with. GenAI was fed with the full text of a number of recent publications in lexicography, and asked to generate one-page comic strips for them.

3 Generating the first comic strips in lexicography

First, an attempt was made at seeing how a recent Cambridge Handbook chapter on the future of dictionaries (de Schryver, 2024), dedicated in a footnote to Sue Atkins (°1931 – †2021), would be treated as a one-page comic strip. The baffling result is shown in Addendum A, with amongst others a painting of Sue Atkins hanging on the wall. Part two of that research, which focuses on the future of metalexicography (de Schryver, 2023a), was tried next. Impressively, *elexiFinder*[6] was singled out as the main vehicle in the comic strip, as seen in Addendum B. The current president of Euralex, and one of the driving forces behind *elexiFinder*, was impressed with the visualisation and termed it "excellent" (Iztok Kosem, 20 January 2026, personal communication). Note that here Sue Atkins moved to the stars, literally, figuratively and visually (see the last panel in Addendum B). Both of these studies were written in 2021-2022, so before the GenAI revolution, and I wondered how GenAI would treat the very topic of GenAI itself. To do so, I took three of my then latest works on the topic of GenAI in lexicography (de Schryver, 2025a, 2025b; Han & de Schryver, 2025), and these are presented in Addenda C, D and E. Apart from giving excellent summaries of the studies themselves, the three pages also give a perfect overview of how the field progressed: tool-wise from ChatGPT-3.5, to GPT-4o, to DeepSeek-R1, language-wise from excellent (English), to from-failure-to-excellent (Portuguese), to from-horrible-to-barely-usable (Lusoga), to end with a new whizz-kid on the block (Chinese).

Finally, I also harked back to material over a decade old, and wished to see how GenAI would treat more complex (African) issues, so I submitted the text on the digitisation of the monolingual Lusoga dictionary (Nabirye & de Schryver, 2013).[7] Simply everything that corresponds with material mentioned in the article is correct, and so are all the extras that depict an African setting (a father holding a candle and a mother at a generator, the type G plug (correct, we are in Uganda) vs. type C plug (wrong outlet), the houses, huts, nature, . . .). Most impressive is the truth: The Lusoga dictionary received funding from three sources, of which one is depicted ("Gift from the Chinese Embassy"), but then in an African setting (see the tape holding the screen and the desktop computer of the time together), the online search terms ('hello', 'love', 'f*', 'v*', 'b*') are all correct, the statistics (2 downloads) are sadly true, etc. On top, the wordplay is brilliant: "From Dusty Dictionary", to "Digital Eureka", to "Going Viral" — or thus "From Obscurity to Legacy", and then of course the uses and thus meanings (see Hanks, 2002) of 'lift' and 'outlet', to end with a new creation by GenAI: 'dictionarate nation'.

5 See https://www.snapchat.com/
6 See https://finder.elex.is/
7 See https://menhapublishers.com/dictionary/

4 The initial wrap-up and 'Post Scriptum' (P.S.)

Each of those comic strips took less than a minute to generate. Their power to now disseminate our academic results to the wider community is immense. Just like GenAI chatbots have already taken over most of lexicography (de Schryver, 2023b; de Schryver & Joffe, 2023), GenAI image generators now ensure we nonetheless continue to have an impact.

For the purpose of the extended abstract, all illustrations of comic strips in lexicography using GenAI were drawn from the work by a single person and his team: me. It goes without saying that, if given the opportunity to present this work, the paper will follow a balanced approach and make sure to show how a wide variety of lexicographers and metalexicographers alike, are impacted by visual-narrative adaptations of their articles and books. In doing so, all continents will be transcended, covering (and thus illustrating) the lexicographic output of over half a century.

5 Further theoretical developments

In his seminal 1959 article, *On Linguistic Aspects of Translation*, Roman Jakobson (°Moscow, 1896 – †Cambridge, MA, 1982) revolutionised translation studies by applying semiotics to the translation process (Jakobson, 1959). He famously argued that the meaning of any linguistic sign is purely a function of its translation into other signs. More broadly, and as generated by Gemini:

> Jakobson defined translation into three distinct types to show that it encompasses much more than just crossing between two languages:

1. **Intralingual Translation (Rewording)**: Translating or interpreting signs within the **same language** (e.g., paraphrasing a sentence, rewriting a complex medical text for a layperson).
2. **Interlingual Translation (Translation Proper)**: Translating signs from **one language into another** (e.g., translating a French novel into English).
3. **Intersemiotic Translation (Transmutation)**: Translating a verbal sign into a **non-verbal sign system** (e.g., adapting a book into a movie, interpreting a poem into a painting or a dance).

Adapting a scientific publication into a comic novel by means of GenAI is thus an intersemiotic translation, as correctly pointed out by Adjudicator 1, who links my current work back to Jakobson's intersemiotic translation: "The full potential of LLMs in generating visual content remains largely unexplored, as does their capacity to perform intersemiotic translation (in Jakobson's sense)." Adjudicator 1 proceeded with pointing out the benefits of this work for three types of users:

In the current context, which is increasingly shaped by visual communication, the visual and narrative translation of scientific papers may have several positive effects:

- **Scholars** may benefit from a synthetic and visually enriched overview that supports faster comprehension, and cross-disciplinary exchange;
- **Students** — including those with disabilities or learning difficulties — may find it easier to grasp the core arguments and feel encouraged to engage more deeply with the text;
- **Journalists and communicators**, in turn, might find such an appealing visual and narrative translation particularly useful for identifying key findings quickly and

communicating them more accurately and accessibly to a broader public, without depending on highly specialized jargon.

The first point, to disseminate and summarise our field in a novel yet accessible way for all future lexicographers was indeed my main goal. That it may also help all future learners, including the neurodivergent ones, is a nice extra argument. As to feeding this new type of material to journalists and communicators, this is an aspect which I already included ever since I began generating comic strips of my published work. This is for instance why I asked Ugandan cartoonist Jimmy Spire Ssentongo[8] to get the comic strip of my 'Broken in Uganda' talk[9] published in a Ugandan newspaper.

Adjudicator 2 was more critical, and stressed that "some parts could have been left out for other, more pertinent reasonings, e.g. risks of overestimating receivers' ability to 'read' images as compared to reading 'traditionally'." Again, this is a very good point. As a Belgian, I probably assume(d) that simply everyone can 'read' comic strips, which is very likely not the case. Not only are there many different types of comic strips — ranging from universally accessible daily gags to highly experimental, abstract, or niche strips that intentionally challenge readers —, not every 'reader' reads in the same way. Even the few one-page comic strips in Addenda A to F can already be viewed in different ways, depending on one's background and prior knowledge. Regarding those addenda, Adjudicator 2 also pointed out: "The addenda are interesting and used to prove the point, also on a meta-level." Agreed, but they are only a selection, as will be shown below.

6 Towards a full comic-strip book on lexicography

Although it was not yet revealed in the extended abstract, all this work will result in an actual comic-strip book on lexicography (de Schryver, 2026c), to be handed out for free to all those present during my talk at Euralex 2026. To make a point, the entire book, literally from cover to cover, will have been created by GenAI. The book cover was created by GenAI, the book title and the entire introduction is the outcome of a discussion between me and ChatGPT[10], and all the comic strips were generated by Academia.edu[11]. With the latter, I have now revealed how the comics were achieved, as was asked by Adjudicator 2: "The approach is modern enough to merit presentation, but again there is no mentioning of details around the GenAI-use to achieve the comics (or of the amount of energy needed to generate images in general . . .)." Yes, interacting with chatbots and running large language models (LLMs) require a significant amount of energy, which is why major tech companies invest heavily in zero-carbon energy to offset their footprint.

Admittedly, there are also a few exceptions to the GenAI-only approach: For example, the ISBN is a real one, the table of contents was added, and in the comic-book section each page on the left *complements* the one-page comic on the right. On the left (so on the even-numbered pages) the full bibliometric details of each publication are presented, together with the original abstract, as well as the number of citations for that publication to date. Given that it is a book, when opened, that bibliometric information will thus always appear together with the one-page comic strip on the right (so on the odd-numbered pages), and as such readers can go through both, check one side only, or even sample parts of each of the facing pages. With this we come to another remark by Adjudicator 2: "Overall, there is little problematization and

8 See https://en.wikipedia.org/wiki/Jimmy_Spire_Ssentongo
9 See https://www.youtube.com/watch?v=wYL6ZNlny5o
10 See https://chatgpt.com/
11 See https://www.academia.edu/

reasoning about potential risks of replacing text with comic strips." Well, I am not really 'replacing' one for the other. Thanks to the bibliometric info, each study is also presented in the most traditional, but summarised, form. On top, readers who want to check the actual full written-out text, are given all the possible details to also consult that work.

The reason for adding the bibliometric information, most of it exported from my reference manager, is explained at length in the introduction to the book. That introduction to the book, over 5000 words in length, is highly interesting in its own way, as there one can analyse what the free ChatGPT chatbot of the time (14 March 2026) could already do. For obvious reasons I cannot reveal the actual contents here, bar one exception.

Near the end of my discussion with ChatGPT, the chatbot wanted to know how I would order the hundreds of one-page comics in lexicography. Here my answer:

> Let me first reply to an earlier question you had, on the ordering of the comics. For now, my main thinking is to put them in historical order, to show how the field evolved from the first serious conference 66 years ago (the proceedings were published two years later, eds. Householder & Saporta 1962), to the first serious handbook 55 years ago (Zgusta 1971), to the first killer dictionary begun in earnest 44 years ago (COBUILD in the early 1980s), . . . 33, 22, 11, . . . 3, 2, 1: all the way to and ending with chatbot lexicography (de Schryver 2023, . . .).
> Within the timeline(s), yes, there could be subcategories, especially depicting the lexicographers vs. the two main schools of metalexicographers, and then no doubt, some very specific topics ought to be grouped.
> That was fun, and thanks for this introduction to my booklet! It'll be copy and pasted verbatim as introductory material, warts and all.

7 Am I being bold?

Adjudicator 2 felt that "some statements are quite bold ('chatbots have already taken over most of lexicography'), with few references to back them up." As in most fields, whether scientific or industrial, you have those who strongly believe in a new technology, and others who prefer to hang on to the past. I am a strong believer, but I also like to study the past, and to compare. In my most recent publications (de Schryver, 2026a, 2026b), I do list the publications by the two camps, so please see there. However, no matter how good the lexicographic products are, whether produced by humans, or machines, or both, for those from the past, the present and the future, some of our colleagues will always find fault with them – often with good reason. Merriam-Webster: the best in America! Right? But read what Patrick Hanks (°1940 – †2024) had to say when he reviewed their *Advanced Learner's English Dictionary* (Hanks, 2009).

Still on missing references, Adjudicator 2 also stressed that "references are mainly from one author, and for other notes, Wikipedia is used, while there is quite an abundant literature on comics and their theory and terminology." Point taken, I could and should have added references to some of the main works in this field. For an accessible yet deep academic dive, there is *Understanding Comics: The Invisible Art* (McCloud, 1993). Using the very medium it analyses, this superb graphic novel-style textbook breaks down the history, psychological impact, and visual language of comic books in a highly engaging, visual format. Another masterclass includes *Comics and Sequential Art* (Eisner, 1985) and the follow-up *Graphic Storytelling and Visual Narrative* (Eisner, 1996). As to academic research on comic books, it spans a wide variety of peer-reviewed journals, book series, and industry magazines. These publications focus on sequential

art, comic history, cultural representation, and narrative structure. An example in each category is *Inks: The Journal of the Comics Studies Society*[12], the series *Routledge Advances in Comics Studies*[13], and *The Comics Journal*[14].

8 Similar work in other scientific fields

There are already several educational comic guides that provide excellent historical and conceptual overviews of both hard and soft sciences. These books use visual storytelling to break down complex disciplines into easily digestible visual lessons. An example from the hard sciences is *Introducing Mathematics: A Graphic Guide* (Sardar et al., 1999) in which witty graphic illustrations and sequential art are used to trace the history and paradoxes of mathematics from ancient times to chaos theory. In *The Cartoon Guide to Physics* (Gonick & Huffman, 1991) humorous comic panels guide the reader through theories, key equations, and the historical evolution of scientific thought. The latter is just one in a long series of dozens of such cartoon guides per scientific field; other guides are for fields like chemistry or biology, or subfields like calculus, statistics, algebra, geometry, etc. Still other such graphic guides are fully interdisciplinary, like *The Stuff of Life: A Graphic Guide to Genetics and DNA* (Schultz et al., 2009) which visualises the entire historical and biological timeline of life on Earth. For the soft sciences an example is *Introducing Philosophy: A Graphic Guide* (Robinson & Groves, 2007) which breaks down the history of philosophical thought and cognitive studies through accessible, narrative comics. Again, this is just one of a series, with others focusing on political philosophy or ethics, down to comic books on specific philosophers like Michel Foucault or Martin Heidegger. And then there is *Introducing Linguistics: A Graphic Guide* (Trask & Mayblin, 2009) in which cartoon illustrations and multi-panel storytelling are used to offer an accessible visual history of language families, grammar theory, and historical phonetics. Here too this work is part of a series, with other comic guides centred on subfields of linguistics only, like one on semiotics.

As to the question whether there are comic books that only use and visualise *existing* academic conference papers, journal articles and scientific books to tell the story; there are some, but they are fewer in number. Rather than summarising a topic broadly, these works function as direct *translations* of complex, peer-reviewed literature into visual narratives. It is a growing movement, strictly concentrating on visualising existing academic work, textbooks, and primary source data. The primary research 'adaptation' of existing academic work, *Unflattening* (Sousanis, 2014) is not even an adaptation, as it was the first-ever PhD written entirely in comic-book form. More specifically, that PhD is an active piece of original academic philosophy and cognitive-science research that uses visual metaphor to prove how imagery and text work together to generate knowledge. An actual example of the visualisation of existing academic work is the two-part *The Cartoon Introduction to Economics* (Grady & Bauman, 2010, 2011). These two books do not just tell stories; they specifically translate the core quantitative models, game theories, and micro/macroeconomic formulae straight out of standard academic economic articles and textbooks into a highly rigorous, page-by-page visual format. Then there is a specific academic movement called 'graphic medicine' that focuses entirely on translating clinical data, sociological studies, and peer-reviewed medical work into comic form. For instance, sociologist Erving Goffman's 1956 landmark book, *The Presentation of Self in Everyday Life* (Goffman, 1956), analyses human social interaction using the metaphor of a theatrical performance. While no single standalone graphic novel completely adapts the original book, his theories are highly visual and are comprehensively broken down in comic form in works like *Introducing Sociology: A Graphic Guide* (Nagle & Pierini, 2016). Likewise, publishers like Routledge have actively

12 See https://ohiostatepress.org/Inks.html

13 See https://www.routledge.com/Routledge-Advances-in-Comics-Studies/book-series/RACS

14 See https://www.tcj.com/

experimented with taking complex legal theory, criminology studies, and ethical research, and turning them into sequential 'data comics' to make the peer-reviewed conclusions accessible to non-academics (see for instance Giddens, 2015). Lastly, in research fields like Human-Computer Interaction (HCI) and Data Science, scholars now publish 'data comics' (Burns, 2025). These are short visual narratives where every panel is explicitly tethered to a specific data point, chart, or methodology section from corresponding scientific output. Academic bodies like the Association for Computing Machinery (ACM) frequently feature these visual guides alongside traditional article texts.

9 My nine points

Despite similar work in other scientific fields, no one has tried to do what I set out to do, along nine different lines, namely:

1. Take each and every of **all major publications**,
2. from a **specific academic field**,
3. to create one-page comic strips for each with a **GenAI image generator**,
4. and to create the title, cover page, and introduction with an **LLM chatbot**,
5. supplemented by **slick bibliometrics**,
6. so that a discipline's **entire history**,
7. can be read and/or visualised in **two-page spreads**,
8. covering every impactful scholar, theory, and language the **world over**,
9. brought together as a **single graphic novel**.

This approach is unique to the field of lexicography, and possibly unique to any academic field. *Tollite et legite* 'Entry Points' (de Schryver, 2026c).

10 PPS

Regarding the full text of this paper, as seen above, Adjudicator 1 wrote:

> The revised paper engages actively and thoughtfully with the Reviewers' comments. I find this both professionally commendable and intellectually honest. As Reviewer 1, I would like to thank the author for the careful consideration given to my suggestions. I therefore confirm my strong recommendation for acceptance.

References

Burns, A. (2025). *Learning to read academic papers by making data comics*. https://nightingaledvs.com/learning-academic-papers-making-data-comics/

de Schryver, G.-M. (2023a). The future of metalexicography: Reaching for the mesosphere. *Lexikos*, *33*(1), 136–165.

de Schryver, G.-M. (2023b). Generative AI and lexicography: The current state of the art using ChatGPT. *International Journal of Lexicography*, *36*(4), 355–387 + ten pages of supplementary material online.

de Schryver, G.-M. (2024). The Future of Dictionaries. In E. Finegan & M. Adams (Eds.), *The Cambridge Handbook of the Dictionary* (pp. 651–667). Cambridge University Press.

de Schryver, G.-M. (2025a). Out-of-the-box GPTs for lexicography. *Lexikos*, *35*(2), 362–396 + six pages of supplementary material online.

de Schryver, G.-M. (2025b). Customised GPTs for lexicography. *Lexikos*, *35*(2), 397–422 + twenty-nine pages of supplementary material online.

de Schryver, G.-M. (2026a). 100 scholarly echoes of Generative AI in lexicography. *International Journal of Lexicography*, *39*, 33 pp. + twelve pages of supplementary material online.

de Schryver, G.-M. (2026b). Citing the future: The legacy of Patrick Hanks. *International Journal of Lexicography*, *39*, 33 pp. + twenty-five pages and about 200 pictures and videos of supplementary material online.

de Schryver, G.-M. (2026c). *Entry Points: The literature of lexicography in one-page comics*. Menha Publishers.

de Schryver, G.-M. & Joffe, D. (2023). *The end of lexicography, welcome to the machine: On how ChatGPT can already take over all of the dictionary maker's tasks [Invited Talk]*. Paper presented at the 20th CODH Seminar, Center for Open Data in the Humanities, Research Organization of Information and Systems, National Institute of Informatics, Tokyo, Japan, 27 February 2023. https://youtu.be/watch?v=mEorw0yefAs

Eisner, W. (1985). *Comics and Sequential Art*. Poorhouse Press.

Eisner, W. (1996). *Graphic Storytelling and Visual Narrative*. Poorhouse Press.

Giddens, T. (Ed.). (2015). *Graphic Justice: Intersections of Comics and Law*. Routledge.

Goffman, E. (1956). *The Presentation of Self in Everyday Life*. Social Sciences Research Centre, University of Edinburgh.

Gonick, L. & Huffman, A. (1991). *The Cartoon Guide to Physics*. Collins Reference.

Han Wanjing [ ] & de Schryver, G.-M. (2025). From user feedback on Chinese-English translation and dictionary tools, to DeepSeek as the ultimate translator and lexicographer. In Y. Gao & X. Zhang (Eds.), *AsiaLex 2025, September 11-13, Fudan, Shanghai, Proceedings of the 18th International Conference of the Asian Association for Lexicography: Traditionality and Creativity in the Era of AI* (pp. 53–62). Fudan University.

Hanks, P. (2002). Mapping meaning onto use. In M.-H. Corréard (Ed.), *Lexicography and Natural Language Processing. A festschrift in honour of B.T.S. Atkins* (pp. 156–198). Euralex.

Hanks, P. (2009). Review of Perrault, S.J. (2008) Merriam-Webster's Advanced Learner's English Dictionary. *International Journal of Lexicography*, *22*(3), 301–315.

Householder, F. W. & Saporta, S. (Eds.). (1962). *Problems in Lexicography*. Indiana University.

Jakobson, R. (1959). On Linguistic Aspects of Translation. In R. A. Brower (Ed.), *On Translation* (pp. 232–239). Harvard University Press.

Klein Grady (illustrator) & Bauman, Y. (2010). *The Cartoon Introduction to Economics, Volume One: Microeconomics*. Hill and Wang.

Klein Grady (illustrator) & Bauman, Y. (2011). *The Cartoon Introduction to Economics, Volume Two: Macroeconomics*. Hill and Wang.

McCloud, S. (1993). *Understanding Comics: The Invisible Art*. Tundra Publishing.

Nabirye, M. & de Schryver, G.-M. (2013). Digitizing the monolingual Lusoga dictionary: Challenges and prospects. *Lexikos*, *23*, 297–322.

Nagle, J. & Piero Pierini (illustrator). (2016). *Introducing Sociology: A Graphic Guide*. Icon Books.

Robinson, D. & Groves, J. (2007). *Introducing Philosophy: A Graphic Guide*. Icon Books.

Sardar, Z., Ravetz, J. & Borin Van Loon (Illustrator). (1999). *Introducing Mathematics: A Graphic Guide*. Icon Books.

Schultz, M., Zander Cannon (Illustrator) & Kevin Cannon (Illustrator). (2009). *The Stuff of Life: A Graphic Guide to Genetics and DNA*. Hill and Wang.

Sinclair John McH. (Chief Ed.), Patrick Hanks (Managing Ed.), Fox, G., Moon, R. & Stock, P. (1987). *Collins COBUILD English Language Dictionary*. William Collins Sons & Co.

Sousanis, N. (2014). *Unflattening: A Visual-Verbal Inquiry into Learning in Many Dimensions* [PhD, Teachers College, Columbia University, NY].

Trask, R. L. & Bill Mayblin (illustrator). (2009). *Introducing Linguistics: A Graphic Guide*. Icon Books.

Zgusta, L. (1971). *Manual of Lexicography*. Academia/Mouton.

Contact information

Gilles-Maurice de Schryver
Ghent University & University of Pretoria
gillesmaurice.deschryver@ugent.be

Acknowledgements

Many thanks are due to Adjudicator 1 and Adjudicator 2, both mentioned as such in the text, with their feedback quoted in the text, and my wife Minah Nabirye, whose publishing company Menha Publishers, based in Kampala, Uganda, is willing to take care of a comic book by a funny professor, being *Entry Points: The literature of lexicography in one-page comics* (de Schryver, 2026c).

Addendum A: The future of dictionaries



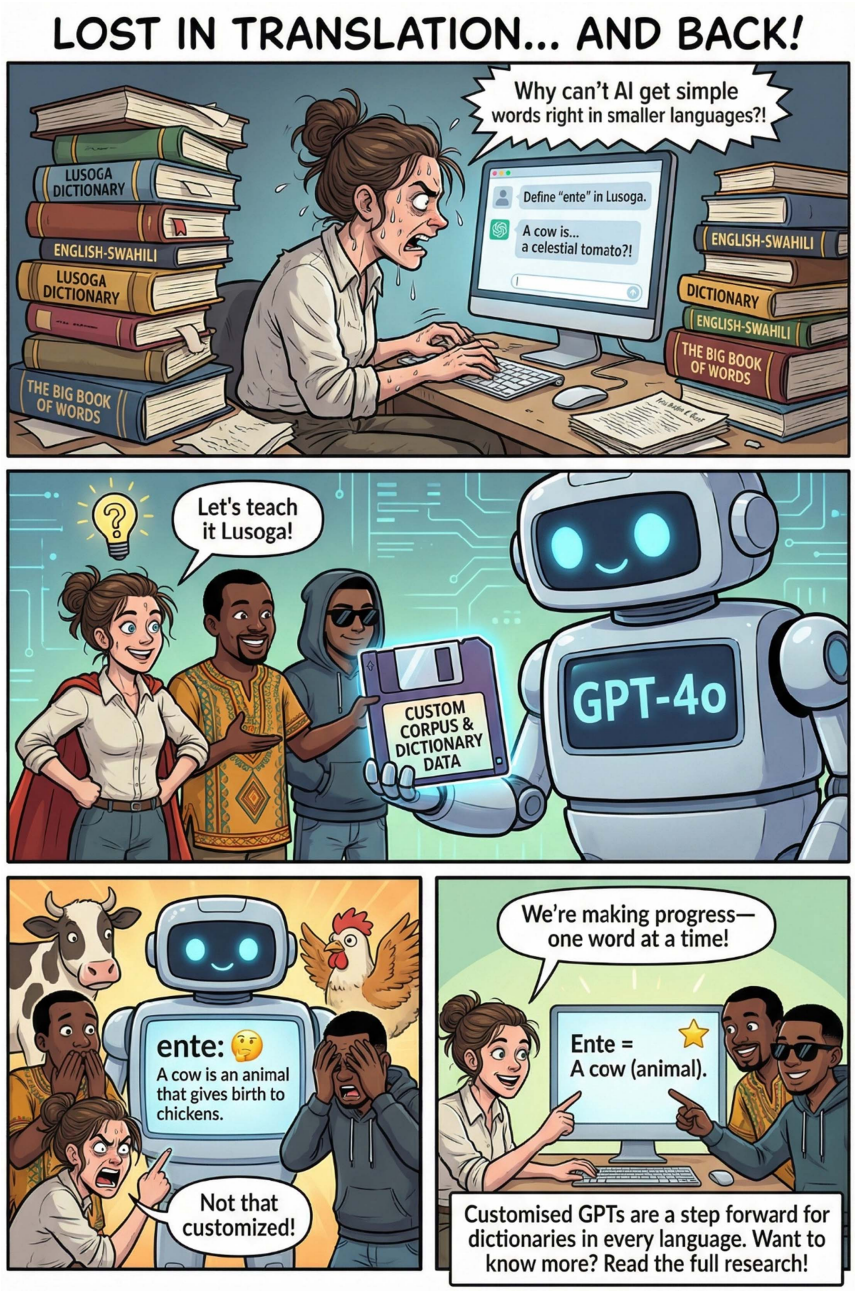
Addendum B: The future of metalexicography



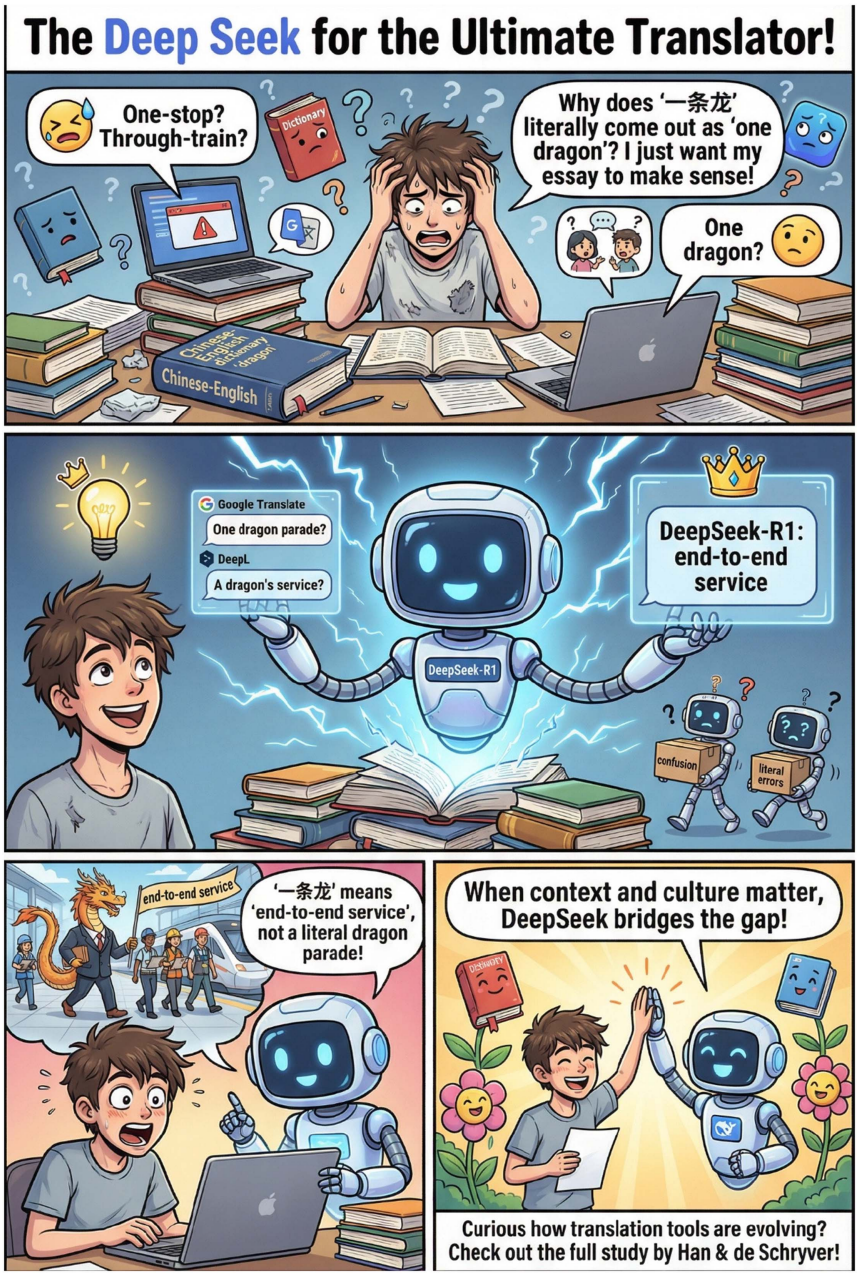
Addendum C: Out-of-the-box GPTs for lexicography



Addendum D: Customised GPTs for lexicography



Addendum E: DeepSeek as the ultimate translator and lexicographer



Addendum F: Digitizing the monolingual Lusoga dictionary



Hanna Fischer, Alfred Lameli, Nathalie Mederake & Nico Urbach

LINKING DICTIONARIES THROUGH MEANING

A Sense-Level Evaluation of Taxonomy-Guided LLM Classification in Lexicography

Abstract Despite the growing availability of digital dictionaries, semantic interoperability often remains limited to headwords, metadata, or article structures. Dictionary senses remain difficult to compare because they are shaped by different editorial traditions, segmentation practices, and levels of semantic granularity. This paper asks whether taxonomy-guided LLM classification can generate reliable, reviewable candidate links between heterogeneous dictionary senses. We test this approach on sense-level units from the letter range N in four German dialect dictionaries: *Mecklenburgisches Wörterbuch*, *Hessen-Nassauisches Wörterbuch*, *Pfälzisches Wörterbuch*, and *Schweizerisches Idiotikon*. Using GPT-5 via the KISSKI interface, each unit is assigned ten times to Post's onomasiological taxonomy, which serves as a shared conceptual reference model. Evaluation combines exact agreement with reference labels, hierarchical proximity, run-to-run reproducibility, and majority aggregation. Exact correctness ranges from 74.91% to 80.88%; top-level taxonomic agreement exceeds 91%; and Fleiss' κ values between 0.78 and 0.83 indicate substantial reproducibility. Majority aggregation raises exact correctness to 78.61–84.28%. The results show that taxonomy-guided LLM classification can support sense-level interoperability by producing interpretable candidate links and reliability signals for targeted lexicographic review.

Keywords LLM classification; semantic interoperability; dialect lexicography; onomasiological taxonomy; sense alignment

1 Introduction

Despite advances in linked-data infrastructures, interoperability in digital lexicography remains largely confined to formal representation — headwords, spelling variants, metadata. How sense descriptions from different dictionaries correspond semantically is rarely specified, which limits the development of integrated lexical knowledge infrastructures. What is missing is a scalable mechanism for comparing senses across resources: word senses are resource-specific constructs that resist direct alignment. This raises a methodological question: How can heterogeneous sense descriptions be compared without treating dictionary senses as directly equivalent units? We address this question through taxonomy-guided LLM classification based on Post's onomasiological taxonomy (Post, 2026), a purpose-built classification system originally developed for the *Pfälzisches Wörterbuch*. The taxonomy functions as an external reference structure: it provides a controlled conceptual grid for comparing resource-specific sense descriptions. The LLM operationalizes this comparison by assigning heterogeneous dictionary definitions to shared taxonomic categories. These assignments can then serve as candidate links between dictionary senses that would otherwise remain isolated.

German dialect dictionaries provide a rigorous test case for this approach: While they share broad lexicographic traditions, they differ substantially in microstructural layout, orthographic conventions, sense granularity, and semantic markup (Moulin, 2010; Lenz & Stöckle, 2020). This paper evaluates four major resources: the *Mecklenburgisches Wörterbuch* (MWb), *Hessen-Nassauisches Wörterbuch* (HNWb), *Pfälzisches Wörterbuch* (PfWb), and *Schweizerisches Idiotikon* (SchwId). The empirical analysis is restricted to the letter range N, which was available in sufficiently comparable form across all four resources (see Sections 3.1 and 6).

The paper asks whether taxonomy-guided LLM classification can provide reliable, reviewable candidate links between heterogeneous dictionary senses by assigning them to a shared onomasiological taxonomy. It examines three questions: (1) How often do model assignments match expert or editorial reference labels? (2) How stable are these assignments across repeated runs and across different dictionaries? (3) Can disagreement patterns and majority scores guide targeted lexicographic review?

2 Theoretical framework

2.1 Levels of comparability and the lexicographic challenge

Interoperability between dialect dictionaries presents challenges at multiple levels. In dialect lexicography, shared macro- and microstructural traditions do not necessarily produce equivalent lexical or semantic structures (Lameli, 2021). At the lexical-form level, formally divergent headwords, orthographic variants, morphological formations, and regional heteronyms may refer to comparable concepts, while formally similar items may differ semantically. Such variation cannot be fully resolved through string similarity or lemma matching alone, especially where lexical variation corresponds to broader semantic differences across dialect regions (Lameli & Hahn, 2026). Normalization, lemma mapping, and etymological linking can address parts of this level, but they do not establish semantic comparability.

The present study therefore focuses on the sense level, where the problem is more fundamental. Dictionary senses are editorial constructs rather than standardized semantic units. As Kilgarriff (1997, p. 108) argues, word senses should not be treated as naturally given units. Editors decide whether usages are grouped into broad senses or split into finer distinctions, how metaphorical or functional extensions are represented, and how much contextual information is made explicit. These decisions produce mismatches in hierarchy and sense granularity, as shown in alignment studies involving large-scale resources such as GermaNet and DWDS (Henrich et al., 2014) as well as smaller lexicographic resources (Salgado et al., 2020). Sense-level interoperability therefore requires a framework that makes resource-specific meaning descriptions comparable without treating them as directly equivalent.

2.2 Onomasiological fields and the need for a comparative model

From an onomasiological perspective, lexical meanings are organized through semantic fields whose boundaries depend on usage, context, and descriptive perspective (Reichmann, 1989). A lexeme may therefore participate in more than one field: it may be classified according to form, function, material, use, evaluation, or metaphorical extension. This fluidity explains why sense-level interoperability[1] cannot be achieved through technical standardization alone.

What is needed instead is a shared reference structure through which heterogeneous sense descriptions can be compared. In this study, this reference structure is understood as a *tertium comparationis*. Its function is not to eliminate semantic variation, but to make classificatory decisions explicit and comparable: heterogeneous sense descriptions can be assigned to shared

[1] In this paper, *lexical meaning* denotes meaning associated with a lexical item in use, whereas *sense* refers to a lexicographically delimited meaning unit. A *sense-level unit* is the operational classification unit extracted from a dictionary article. *Semantic field* denotes a broader meaning domain in which lexical items or senses may participate; *conceptual category* refers to a defined node in Post's taxonomy.

conceptual categories, deviations can be evaluated, and uncertainty can be interpreted as part of the linking process.

2.3 Taxonomies as operational models of lexical meaning

Post's taxonomy (1998; 2026) serves as a concrete operationalization of this tertium comparationis. Originally developed for the *Pfälzisches Wörterbuch* (PfWb) as a practice-oriented adaptation of the extra-linguistic conceptual system by Hallig & von Wartburg (1963), it consists of nine top-level categories and a hierarchical structure encoded in four-digit class identifiers. Its purpose is not theoretical exhaustiveness, but practical lexicographic coverage: it is broad enough to classify the lexical material of a comprehensive dialect dictionary while remaining structured enough to support systematic comparison. Following Bergenholtz's (2009) functional understanding of classification, the taxonomy is used here not as an ontology of lexical meaning, but as a purpose-bound model for rendering heterogeneous sense descriptions comparable.

TABLE 1 Example for taxonomic assignment as candidate linking

Dict.	Lemma	Sense	Assigned category	Parent category
MWb	*Niman*	Neumond	2100 Sky, celestial bodies	2000 Inanimate nature
HNWb	*Neumond*	Wie schd. Neumond	2100 Sky, celestial bodies	2000 Inanimate nature
PfWb	*Neu-mond*	zunehmender Mond	2100 Sky, celestial bodies	2000 Inanimate nature
SchwId	*Nūū(w)mān*	Neumond	2100 Sky, celestial bodies	2000 Inanimate nature

As Table 1 shows, the link between dictionary resources is not established through identical headwords or equivalent article structures, but through their assignment to a shared conceptual category. The resulting connection is therefore not a direct claim of sense identity. Rather, it is a candidate link: the taxonomy identifies senses that belong to the same conceptual field and can therefore be compared, grouped, or selected for lexicographic review. At the same time, this operational character entails limitations. Because the taxonomy was designed for broad lexicographic coverage, it includes residual and weakly bounded categories, including "other"-type or evaluative classes. In manual classification, such cases can be resolved through expert judgement and editorial convention. In automated classification, however, they may become sources of instability, since weakly defined nodes provide less explicit semantic guidance for the model. Post's taxonomy is therefore treated as a decision space for LLM-based classification: it constrains assignments, supports reproducibility, and makes uncertainty empirically observable rather than treating it as residual error (cf. Section 5; Mederake et al., 2026). This operational view of the taxonomy also determines how the present study relates to existing work on lexical representation, sense alignment, and LLM-assisted classification.

2.4 Related work and desiderata

Research on semantic interoperability in lexicography has developed along four partially overlapping strands. A first strand focuses on formal representation and linked-data integration. Approaches based on OntoLex-Lemon have demonstrated how lexical resources can be embedded in machine-readable semantic infrastructures (Declerck & Wandl-Vogt, 2014; Abgaz, 2020), but they primarily address formal representation rather than the semantic mapping of heterogeneous sense descriptions. A second strand treats sense alignment as a

computational matching problem related to word sense disambiguation (Navigli, 2009). Supervised approaches learn mappings across resources (Ahmadi & McCrae, 2021; Saurí et al., 2019), but typically presuppose relatively stable sense inventories and are less suited to resources whose sense boundaries are implicit or inconsistently encoded. A third strand is represented by concept-oriented infrastructures such as VerbaAlpina (Krefeld, 2025), which reorganize lexical knowledge through shared conceptual structures rather than lexical forms. However, such infrastructures rely on proactively curated data models and are not primarily designed for the retroactive integration of heterogeneous dialect dictionaries.

A fourth strand explores LLM-assisted semantic classification in dialect lexicography. Stöckle et al. (2025) discuss the opportunities and limitations of such workflows, while Mederake et al. (2026) established the feasibility of taxonomy-guided LLM classification on the *Hessen-Nassauisches Wörterbuch* under two classification runs, with run-to-run agreement approximating expert inter-rater reliability. What remains insufficiently tested is whether this approach scales to structurally heterogeneous resources and whether repeated model outputs can be translated into practical reliability signals for review. The present study addresses this gap by evaluating four dialect dictionaries across ten repeated runs and by combining exact correctness with hierarchical proximity, majority-based reliability, and a diagnostic typology of instability. The following section translates this comparative aim into a concrete classification setup and evaluation design.

3 Material and Methods

3.1 Data and sampling

The evaluation covers four German dialect dictionaries across Low German, Central German, and Upper German: *Mecklenburgisches Wörterbuch*, *Hessen-Nassauisches Wörterbuch*, *Pfälzisches Wörterbuch* and *Schweizerisches Idiotikon*. The resources play complementary roles in the experimental design: the PfWb serves as the taxonomy-native resource, since its editorial classification is embedded in Post's system itself; the HNWb provides a highly curated Central German benchmark; the MWb adds a Low German comparison case prepared under near-real-world conditions; and the SchwId introduces a structurally divergent Upper German resource.

The analysis was restricted to the letter range N, one of the few ranges available in sufficiently complete form across all four resources. Although alphabetic ranges are not semantically neutral, the N range comprises both high-frequency function words and low-frequency content words and exhibits substantial editorial heterogeneity across the four dictionaries. Based on the complete set of editorial and expert reference annotations, including multiple category assignments where semantically warranted, the retained units cover all nine top-level categories of Post's taxonomy. Of the 9,517 reference assignments associated with the 6,312 sense-level units, top-level category 7 ('Human being (social)') accounts for the largest share (27.07%), whereas top-level category 2 ('Inanimate nature') accounts for the smallest (2.12%). The results therefore demonstrate feasibility and stability within a shared comparative test range, but do not support claims about coverage across the full lexicon.

To ensure comparability across resources, the evaluation was restricted to articles containing classifiable sense-level material. Pure cross-reference articles and other non-evaluable entries were excluded, since they do not provide independent sense descriptions and their resolution constitutes a rule-based linking task rather than semantic classification. Table 2 summarizes the resulting evaluation datasets, including the number of retained articles, extracted sense-level units, and summary statistics on polysemy.

TABLE 2 Conditioned evaluation datasets in the letter range N

Dict.	Articles (N)	Sense-level units	Polysem. articles (%)	Mean units/article
MWb	1,363	1,546	4.78	1.13
HNWb	788	964	11.29	1.22
PfWb	1,241	1,449	11.52	1.17
SchwId	2,353	2,353	n/a[a]	1.00

[a]For the SchwId, 628 articles containing multiple meaning strings were excluded because their internal sense structure could not be reconstructed reliably; polysemy statistics are therefore not applicable. A further 596 articles lacked semantic descriptions.

All statistics refer to the conditioned evaluation datasets. Polysemous articles are defined as articles containing at least two retained sense-level units with distinct meaning paraphrases or equivalents.

3.2 Source representation and sense-level units

The four dictionaries differ substantially in digital encoding, microstructural layout, and the explicitness of their sense descriptions. Table 3 illustrates how one comparable semantic target from the category 'sky, celestial bodies' is represented in the four source dictionaries and reduced to a sense-level unit. The excerpts are shortened and lightly normalized for readability; metadata and unrelated article content are omitted, and legacy markup is resolved in line with the preprocessing used to construct the classification input.

TABLE 3 From source representation to sense-level unit

Dict.	Source format	Source excerpt	Resulting sense-level unit
MWb	TEI-P5 Lex-0	<form type="lemma">Niman</form> <form type="variant">Nimaand</form> <gramGrp><gram type="gen">m.</gram></gramGrp> <sense level="0"><cit type="translation" xml: lang="deu"><quote>Neumond</quote></cit>: bei Neumond geborenes Vieh . . . </sense>	*Niman* · 'Neumond'
HNWb	TEI XML Dict.	<form type="lemma">Neumond</form> <gramGrp><gen>M.</gen></gramGrp> <sense level="0">Wie schd. '<def>Neumond</def>'. Seltener neuer Mond . . . </sense>	*Neumond* · 'wie schd. Neumond'
PfWb	SGML	<form type="lemma">Neu-mond</form> <gram type="m">m.</gram> <sense>'zunehmender Mond', Neimond, -mand, -mund . . . </sense>	*Neu-mond* · 'zunehmender Mond'
SchwId	JSON	{"lemmaIndexEntry":"Nūū(w)mān 4,235", "semID":" . . . ", "semDescription":"Neumond"}	*Nūū(w)mān* · 'Neumond'

The examples illustrate the practical consequences of this source heterogeneity for sense-level extraction. Comparable meaning descriptions are not always encoded with the same degree of explicitness: the HNWb, for instance, marks the Standard-German definition as a <def> element but introduces it through the formula 'wie schd.' (= as in Standard German), which is resolved for the model through the abbreviation list supplied in the context file. The SchwId example, by contrast, provides the relevant semantic information in a compact API field.

For classification, these source representations are reduced to comparable sense-level units consisting of lexical form and interpretable sense description, supplemented where available by contextual information from the article. Reference labels are not part of the model input. The PfWb example illustrates the kind of near-equivalence ('new moon' vs. 'waxing moon') that requires review rather than being treated as direct sense identity, as discussed in Section 2.3.

3.3 Reference labels and annotation

The evaluation compares model assignments against reference labels whose provenance differs across the four resources. For the PfWb, the reference categories derive from taxonomy-native editorial classification embedded in the original dictionary data: assigning a sense-level unit therefore amounts to recovering a historically embedded editorial decision. For HNWb, MWb, and SchwId, no such editorial classification exists, and reference labels were established retrospectively by two expert lexicographers applying the same taxonomic framework. These datasets therefore test agreement with newly created annotations rather than the recovery of pre-existing ones.

The annotation followed the procedure established in Mederake et al. (2026). Both annotators independently assigned a primary four-digit Post category to every sense-level unit admitted to evaluation and could add further categories when supported by the source sense. All items were double-annotated rather than divided between annotators. The independent annotations were then compared, and disagreements were resolved through discussion to establish a consensual set of expert reference labels for automated evaluation. Inter-annotator agreement was quantified using unweighted Cohen's κ on the primary category assigned by each annotator; additional category assignments were excluded from this calculation to preserve comparability with the single-label classification setup. Agreement reached $\kappa = 0.743$ for HNWb, $\kappa = 0.831$ for MWb, and $\kappa = 0.940$ for SchwId.

Because the reference labels for three of the four resources are consensual expert annotations rather than an externally validated gold standard, we refer to them throughout as expert reference labels. This terminology is deliberate: it acknowledges that the labels encode a controlled but interpretive editorial decision, which is itself one of the sources of uncertainty examined in the disagreement analysis (Section 5).

3.4 Taxonomy-guided LLM classification setup

Following Mederake et al. (2026), the task was implemented as a controlled in-context learning setup with five few-shot examples rather than as a fine-tuned classification model.[2] Each sense-level unit, consisting of the lexical form, available contextual information, and gloss or translation, was presented together with a context file containing the full extended Post taxonomy (approximately 10,700 tokens), abbreviation lists, lexicographic microstructure, and context information for each dictionary. The model was asked to assign exactly one four-digit category.

The main experiment was conducted with GPT-5 via the KISSKI interface (AI Service Centre for Sensitive and Critical Infrastructure, GWDG; https://kisski.gwdg.de). Each sense-level unit was

[2] The five few-shot examples were selected to cover concrete objects, persons, actions, abstract or evaluative meanings, and cases of potential taxonomic ambiguity. The prompt structure is summarized in Mederake et al. (2026); the full prompt and the corresponding context file (including abbreviation lists), exemplified for the MWb, are available at https://doi.org/10.57712/lingurep-115001.

classified ten times under identical prompting conditions, with sampling parameters set to the KISSKI defaults (temperature = 0.5, top-p = 0.5). The repeated runs serve to estimate output stability under stochastic sampling and are treated as exchangeable replicates rather than as separate experimental conditions. The prompt required a single four-digit category code in structured output; internal reasoning was suppressed by the prompt design and was neither stored nor used for scoring.

3.5 Evaluation framework

The evaluation combines four complementary perspectives: classification correctness, hierarchical proximity, run-to-run reproducibility, and majority-based practical reliability. Together, these measures assess exact agreement with reference labels, the taxonomic distance between model predictions and reference labels, stability across repeated classifications independently of reference labels, and the extent to which repeated outputs can be converted into reliability signals for lexicographic review.

3.5.1 Classification correctness

Classification correctness measures whether the predicted four-digit Post category matches the reference annotation. Assignments were classified as "correct" when the predicted category matched at least one of the reference labels assigned to the sense-level unit. In the absence of an exact match, they were classified as "almost correct" when the predicted category was hierarchically close to at least one reference category, either as its child, its ancestor, or as a sibling sharing the same parent category. All other assignments, including model outputs that did not contain a valid four-digit category, were treated as "incorrect". This three-way distinction separates exact matches from near misses within the local taxonomic neighbourhood.

Because ancestors, children, and siblings are not always equidistant in semantic terms, "almost correct" is used as an operational proximity measure, not as a claim of full semantic equivalence. Table 4 illustrates the three outcomes for reference category 6116 ('improper behaviour').

TABLE 4 Worked example of the three-way evaluation, reference category 6116 ('improper behaviour')

Reference label	Prediction	Evaluation	Reason
6116 improper behaviour	6116 improper behaviour	correct	exact match
6116 improper behaviour	6111 swear words	almost correct	sibling category within the same parent node (611x)
6116 improper behaviour	5360 speech/articulatory organs	incorrect	different top-level category (6xxx vs. 5xxx)

3.5.2 Hierarchical proximity

Hierarchical proximity evaluates agreement at successive levels of Post's taxonomy. Predicted labels and all reference labels assigned to a sense-level unit were reduced to levels 1 to 4, ranging from broad top-level to exact four-digit categories. For each level, we report the mean proportion of predictions matching at least one reference label across the ten repeated runs.

This measure distinguishes distant misclassifications from near misses within related branches of the taxonomy.

3.5.3 Run-to-run reproducibility

Run-to-run reproducibility measures the extent to which ten repeated model calls under identical prompting conditions converge on the same output. Fleiss' κ was calculated on the predicted four-digit Post category labels and, in a supplementary analysis, on the derived three-way evaluation outcome. This distinguishes stability of category assignment from stability of evaluation status. Unlike the accuracy-based measures in Sections 3.5.1 and 3.5.2, κ does not compare classifications with reference labels but quantifies convergence of the repeated outputs beyond chance agreement.

3.5.4 Majority aggregation and practical reliability

For each sense-level unit, the most frequent category across the ten runs was defined as the majority category; the number of supporting runs was recorded as the majority score. Tie cases occur when no unique majority category emerges. Three tie-resolution strategies were evaluated: best case, selecting a reference category if present among tied candidates; worst case, selecting a non-reference category; and random case, selecting randomly among tied candidates.

4 Evaluation

4.1 Model selection and experimental baseline

As a preliminary model comparison, five models in six configurations were evaluated on the MWb letter range N. The comparison used provider defaults or low-temperature configurations: GPT-5 was tested at temperatures 0.5 and 0.1, GPT-5.4 at 0.5, Qwen 3.5 397B A17B at 0.6, and Gemini 3.1 Pro Preview and Mistral Large Instruct 3 675B at 0.1.[3] GPT-5 (temp = 0.5) and Gemini 3.1 Pro Preview (temp = 0.1) achieved comparable results (79.6% vs. 79.2% correct assignments), with no statistically significant difference in correct-assignment rates across repeated runs (Welch's $t(13.676) = 0.905$, $p = 0.381$). Qwen 3.5 397B A17B reached 69.3% and Mistral Large Instruct 3 675B 51.2%. Since this comparison was not designed as a standardized benchmark, GPT-5 was selected as the primary model because it combined strong performance with reliable access through the KISSKI interface. Gemini is retained only as an indicative secondary reference for reproducibility in Section 4.4. Sections 4.2 and 4.3 report averages over ten GPT-5 runs; Section 4.5 reports majority-aggregated results.

4.2 Classification performance across dictionaries

Classification performance proved to be remarkably stable across the four dictionaries. The average correct classifications range from 74.91% in the PfWb to 80.88% in the SchwId. If "almost correct" classifications are included, that is, assignments located in a hierarchically close taxonomic neighbourhood, the proportion of semantically adequate assignments rises to between 83.55% and 88.88%.

[3] The KISSKI labels "OpenAI ChatGPT-5" and "OpenAI ChatGPT-5.4" are shortened to GPT-5 and GPT-5.4.

TABLE 5 Mean classification outcome across ten GPT-5 runs per dictionary

Evaluation	MWb	HNWb	PfWb	SchwId
Correct	79.64%	79.19%	74.91%	80.88%
Almost Correct	8.56%	9.54%	8.64%	8.00%
Incorrect	11.80%	11.27%	16.45%	11.12%

The comparatively lower exact accuracy for the taxonomy-native PfWb and the stronger performance of the structurally less aligned SchwId show that classification difficulty cannot be explained by structural proximity to the taxonomy alone. The relation between gloss explicitness, label provenance, and taxonomic granularity is discussed in Section 5.

4.3 Hierarchical proximity

Hierarchical proximity was evaluated by reducing both the reference labels and the model predictions to successive levels of Post's taxonomy. For each level, Table 6 reports the mean proportion of predictions matching at least one reference label across the ten repeated GPT-5 runs. This captures whether deviations affect coarse-grained semantic distinctions or mainly fine-grained category decisions.

Agreement exceeds 91% at the top level across all dictionaries and remains above 83% at the second level. It then decreases gradually toward the exact four-digit category level. This pattern indicates that the model reliably identifies higher-level semantic categories, while uncertainty increases with taxonomic granularity. Hierarchical proximity is therefore more informative than exact correctness alone: it distinguishes distant misclassifications from near misses within related branches of the taxonomy.

TABLE 6 Mean agreement (%) between GPT-5 predictions and reference annotations at successive hierarchy levels across ten repeated runs

Hierarchy level	MWb	HNWb	PfWb	SchwId
1	93.35	95.05	91.83	93.74
2	87.68	88.69	83.74	89.09
3	81.61	81.84	77.84	83.74
4	79.64	79.19	74.91	80.88

From a lexicographic perspective, this distinction is crucial: deviations already at the top level would indicate a failure to place the sense in the correct major area of the taxonomy, whereas deviations confined to lower levels typically reflect fine-grained category selection within an otherwise correctly identified branch. Since agreement at the upper levels remains high across all dictionaries, the observed instability does not undermine the approach but supports the use of majority aggregation and reliability-based review (Section 4.5).

4.4 Run-to-run reproducibility

GPT-5 reaches substantial run-to-run reproducibility across all four dictionary datasets, with $\kappa = 0.78$ for MWb, $\kappa = 0.79$ for HNWb, $\kappa = 0.83$ for PfWb, and $\kappa = 0.79$ for SchwId.

Using Gemini 3.1 Pro, a higher κ value was observed on the MWb data (κ = 0.87) under a lower temperature setting (0.1), suggesting that part of the remaining variation may be sampling-related; since model architecture and temperature differ simultaneously, however, this comparison is used only as an indicative reference. Although the SchwId achieves the highest exact accuracy (Table 5), its κ value is not higher than those of the other datasets. Correctness and run-to-run consistency therefore do not fully coincide. Values below near-perfect agreement show that a subset of items remains unstable across runs, identifying sense units for which the source description, taxonomy, or reference classification does not make one category uniquely recoverable.

4.5 *Majority aggregation and practical reliability*

Majority aggregation, i.e. selecting the most frequent category across the ten repeated classifications for each sense-level unit, improves exact classification performance across all dictionaries (Table 7). Under the random tie strategy (see Section 3.5.4), which best approximates a large-scale application scenario in which no reference labels are available for tie resolution, correct assignments increase to 82.60% for MWb, 82.37% for HNWb, 78.61% for PfWb, and 84.28% for SchwId. Tie cases remain rare, ranging from 2.8% to 3.6%.[4] Repeated classification and majority selection therefore provide a practical mechanism for reducing stochastic variation while preserving interpretability. The majority score, defined as the number of runs supporting the majority category, also provides an item-level reliability signal for review. The following discussion examines what instability reveals about uncertainty in the model, the taxonomy, and the source data.

TABLE 7 Majority-based classification correctness after aggregating ten GPT-5 runs per sense-level unit, under three tie-resolution strategies. Tie cases indicate sense-level units for which no unique majority category could be determined

Dict.	Best Case	Random Case	Worst Case	Tie Cases
MWb	83.51%	82.60%	81.76%	2.8%
HNWb	83.51%	82.37%	81.02%	3.4%
PfWb	79.50%	78.61%	77.64%	3.1%
SchwId	85.38%	84.28%	83.38%	3.6%

5 Discussion: Instability as diagnostic evidence

The evaluation shows that disagreement in taxonomy-guided LLM classification should not be treated merely as residual error. Rather, it provides diagnostic evidence for uncertainty arising from the interaction between model output, source dictionary data, and taxonomic structure. At broader taxonomic levels, classification is reliable; exact four-digit classification, however, remains sensitive to semantic ambiguity, weak category boundaries, and the explicitness of source glosses.

Four types of instability can be distinguished. First, stochastic variation occurs when repeated runs fluctuate between hierarchically close categories and can often be reduced through majority aggregation. Second, conceptual ambiguity arises when a sense genuinely participates in more than one semantic field, as in *Nachtpott* ‘chamber pot’, which may be classified by object

[4] In the majority-aggregated output, seven units (0.1% of the evaluated units; four MWb and three SchwId units) did not receive a valid four-digit category and were counted as incorrect.

type as 7735 'containers' or by function as 5340 'digestion and various forms of excretion'. Third, categorial indeterminacy concerns weakly bounded taxonomy nodes such as 6510 'moral evaluations', 6116 'improper behaviour', and 6111 'swear words' (cf. Mederake et al., 2026, pp. 129–131). Fourth, data-related underspecification occurs when brief or implicit sense descriptions do not provide enough semantic information for stable fine-grained classification.[5]

This typology helps explain why the taxonomy-native PfWb achieves lower exact accuracy than expected. Its labels reflect embedded editorial decisions that may depend on article structure, contextual cues, or category conventions no longer visible in the extracted sense-level unit. Where sense descriptions are short, the cues needed to distinguish between action-related, cognitive, evaluative, or functional categories may be weakly expressed. The contrast with the SchwId shows that taxonomy-native status should not be equated with classification ease: although structurally less aligned with Post's taxonomy, many retained SchwId sense descriptions provide explicit semantic cues. Classification difficulty therefore depends not only on the taxonomy, but also on source description and reference-label provenance.

For lexicographic workflows, the relevant question is how manual review can be prioritized under limited resources. Reviewing the 22 most error-prone categories would raise majority-based correctness under the random tie strategy (Table 7) by 6.83–7.68 percentage points across the four dictionaries, while requiring inspection of only 12.14–22.25% of the evaluated units. Category-based review therefore offers a more favorable effort-gain ratio than reviewing all cases with low majority scores.

Majority scores can then serve as a second-stage prioritization mechanism. Once recurrent problem categories have been inspected, additional cases can be reviewed cumulatively from the lowest scores upward—not according to a fixed binary distinction between stable and unstable cases, but as far along the score distribution as the available review budget permits.

Table 8 combines correction of these categories with cumulative majority-score-based review of the remaining cases. Since empirical score bands do not correspond neatly to round effort levels, intermediate values were estimated by linear regression. A review effort of 30% would raise correctness above 90% for all four dictionaries, reaching 97.19% for HNWb. In the present evaluation, the categories were identified by comparing majority assignments with expert reference labels for all retained units. In workflows without full reference annotation, comparable review ranges could be calibrated on a smaller adjudicated validation sample; majority scores are generated directly by repeated classification.

TABLE 8 Estimated correctness gains under bounded targeted review[a]

Manual review effort (%)	HNWb gain / correctness	MWb gain / correctness	PfWb gain / correctness	SchwId gain / correctness
10	5.56 / 87.93	3.61 / 86.21	4.54 / 83.15	4.10 / 88.38
20	10.19 / 92.56	6.86 / 89.46	8.14 / 86.75	7.14 / 91.42
30	14.82 / 97.19	10.11 / 92.71	11.75 / 90.36	10.19 / 94.47

[a]Each cell reports percentage-point gain/resulting correctness rate.

[5] For example, HNWb *Nuselbart*, *Nuseler*, and *Nuselnase* are glossed as persons speaking indistinctly or nasally and receive unstable assignments to 6116 'person with improper behaviour' (majority score 5/10), whereas the near-identical PfWb *Nuseler* is assigned to 5360 'speech and articulatory organs'. The contrast illustrates how brief glosses may foreground either the person or the speech act, shifting the classification across distinct Level-2 categories.

These findings support Stöckle et al.'s (2025) assessment that LLM-assisted dialect lexicography depends on carefully prepared data, prompting, and review structures. The value of the approach lies not only in high broad-level reliability, with top-level agreement exceeding 90% across all dictionaries before manual correction, but also in making remaining uncertainty measurable and operationally usable. Majority scoring and category-based review do not replace lexicographic judgement; they help prioritize where it is most needed.

6 Conclusion

Returning to the question raised at the outset, this study has shown that taxonomy-guided LLM classification can generate reviewable candidate links across four structurally heterogeneous dialect dictionaries. By mapping sense-level units onto Post's onomasiological taxonomy, the procedure makes resource-specific lexicographic data comparable through an explicit mediating model of meaning. The repeated-classification design, in turn, provides reliability signals that make the resulting links open to inspection rather than merely asserted.

The three questions posed in Section 1 can be answered positively, but with important qualifications. Model assignments match expert or editorial reference labels in 74.9–80.9% of cases, with top-level agreement exceeding 91%; assignments remain substantially stable across repeated runs and resources, with Fleiss' κ between 0.78 and 0.83; and majority aggregation raises exact correctness to 78.6–84.3%, while disagreement patterns and majority scores help guide targeted lexicographic review. These results confirm and extend earlier single-resource evidence (Mederake et al., 2026).

At the same time, the approach does not remove the need for lexicographic judgement in fine-grained or ambiguous cases; its practical value lies in making such cases identifiable. Disagreement, in this design, is analytically productive: the workflow treats instability not as residual error but as operationally visible uncertainty, indicating where expert attention, sharper decision rules, or future multi-label modelling are most needed.

Several limitations remain. The single-label model output simplifies cases of genuine conceptual overlap; future multi-label experiments should therefore evaluate secondary assignments by lexicographic plausibility and stability rather than exact accuracy alone. The sampling limitation discussed in Section 3.1 also remains important: future work should extend the evaluation beyond the letter range N to the full lexicon and test whether the observed patterns remain stable across models, prompt designs, and sampling settings.

In sum, this study demonstrates a review-oriented workflow for linking heterogeneous dictionary senses. The candidate links generated by the procedure are not claims of full sense identity; they are interpretable hypotheses grounded in a shared conceptual model and accompanied by reliability signals such as majority scores, run-to-run stability, and hierarchical proximity. In this sense, the approach complements concept-oriented infrastructures such as VerbaAlpina: it is not intended to replace proactively curated conceptual models, but offers a retrospective mechanism for filtering, prioritizing, and reviewing historically grown dictionary resources whose semantic structure was not originally designed for interoperability. Moreover, because the underlying taxonomy is also available as an RDF graph (https://doi.org/10.57712/LINGUREP-114999), the resulting links can be integrated directly into linked-data infrastructures, connecting retrospectively enriched dictionary resources with the broader ecosystem of conceptually organized language data.

References

Abgaz, Y. (2020). Using OntoLex-Lemon for Representing and Interlinking Lexicographic Collections of Bavarian Dialects. In M. Ionov, J. P. McCrae, Ch. Chiarcos, Th. Declerck, J. Bosque-Gil & J. Gracia (Eds.), *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)* (pp. 61–69). European Language Resources Association.

Ahmadi, S. & McCrae, J. P. (2021). Monolingual Word Sense Alignment as a Classification Problem. In P. Vossen & Ch. Fellbaum (Eds.), *Proceedings of the 11th Global Wordnet Conference* (pp. 73–80). Global Wordnet Association.

Bergenholtz, H. (2009). Classification in linguistics and lexicography: word terms and word relations. *Zeitschrift für Anglistik und Amerikanistik*, 57(3), 275–288.

Declerck, Th. & Wandl-Vogt, E. (2014). How to semantically relate dialectal Dictionaries in the Linked Data Framework. In K. Zervanou, C. Vertan, A. van den Bosch & C. Sporleder (Eds.), *Proceedings of the 8th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities (LaTeCH)* (pp. 9–12). Association for Computational Linguistics.

Hallig, R. & Wartburg, W. von (1963). *Begriffssystem als Grundlage für die Lexikographie. Versuch eines Ordnungsschemas* (2nd Rev. ed.). Akademie-Verlag.

Henrich, V., Hinrichs, E. W., & Barkey, R. (2014). Aligning Word Senses in GermaNet and the DWDS Dictionary of the German Language. In H. Orav, Ch. Fellbaum & P. Vossen (Eds.), *Proceedings of the Seventh Global WordNet Conference* (pp. 63–70). University of Tartu Press.

Kilgarriff, A. (1997). I Don't Believe in Word Senses. *Computers and the Humanities*, 31(2), 91–113. https://doi.org/10.1023/A:1000583911091

Krefeld, Th. (2025). Wörter, Sachen und Bücher in der alpinen und ladinischen Lexikographie: Eine methodologische Bemerkung ausgehend vom Dizionario del dialetto di Montagne di Trento von Corrado Grassi. *Ladinia*, XLIX, 97–109. https://doi.org/10.54218/LADINIA.49.97-109

Lameli, A. (2021). Dialektwörterbücher zwischen Web 0.0 und Web 3.0. In O. Ehrismann & I. Hardt (Eds.), *Das Sudetendeutsche Wörterbuch: Bilanzen und Perspektiven* (pp. 45–70). Frank & Timme.

Lameli, A. & Hahn, M. (2026). The Geography of Meaning: Investigating Semantic Differences Across German Dialects. *Languages*, 11(3), 56. https://doi.org/10.3390/languages11030056

Lenz, A. N. & Stöckle, Ph. (2020). *Germanistische Dialektlexikographie zu Beginn des 21. Jahrhunderts*. Steiner.

Mederake, N., Urbach, N., Fischer, H. & Lameli, A. (2026). Onomasiological Sense Alignment Across Dialect Dictionaries. A Taxonomy-Constrained LLM Classification. In Y. Scherrer, N. Aepli, V. Blaschke, T. Jauhiainen, N. Ljubešić, P. Nakov, J. Tiedemann & M. Zampieri (Eds.), *Proceedings of the 13th Workshop on NLP for Similar Languages, Varieties and Dialects* (pp. 123–138). Association for Computational Linguistics.

Moulin, C. (2010). 33. Dialect dictionaries - traditional and modern. In P. Auer & J. E. Schmidt (Eds.), *Language and Space. Volume 1 Theories and Methods: An International Handbook of Linguistic Variation* (pp. 592–612). Walter de Gruyter.

Navigli, R. (2009). Word sense disambiguation: A survey. *ACM Computing Surveys (CSUR)*, 41(2), 1–69. https://doi.org/10.1145/1459352.1459355

Post, R. (1998). Möglichkeiten der elektronischen Strukturierung, Vernetzung und Verfügbarmachung von lexikographischen Daten bei der Arbeit am Pfälzischen Wörterbuch. In R. Grosse (Ed.), *Bedeutungserfassung und Bedeutungsbeschreibung in historischen und dialektologischen Wörterbüchern* (pp. 211–220). Hirzel.

Post, R. (2026). *Postsche Sachgruppen - CSV* (N. Mederake & N. Urbach, Ed.). [Data set]. Research Center Deutscher Sprachatlas.

Reichmann, O. (Ed.). (1989). *Frühneuhochdeutsches Wörterbuch. Einführung, a - äpfelkern*. Walter de Gruyter.

Salgado, A., Ahmadi, S., Simões, A., McCrae, J. P., & Costa, R. (2020). Challenges of Word Sense Alignment: Portuguese Language Resources. In M. Ionov, J. P. McCrae, Ch. Chiarcos, T. Declerck, J. Bosque-Gil & J. Gracia (Eds.), *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)* (pp. 45-51). European Language Resource Association.

Saurí, R., Mahon, L., Russo, I. & Bitinis, M. (2019). Cross-Dictionary Linking at Sense Level with a Double-Layer Classifier. In M. Eskevich, G. de Melo, Ch. Fäth, J. P. McCrae, P. Buitelaar, Ch. Chiarcos, B. Klimek & M. Dojchinovski (Eds.), *2nd Conference on Language, Data and Knowledge (LDK 2019)*. Open Access Series in Informatics (OASIcs), Volume 70 (pp. 20:1–20:16), Dagstuhl Publishing.

Stöckle, P., Elsner, D., Koppensteiner, W., & Korecky-Kröll, K. (2025). LLM-Assisted Dialect Lexicography: Challenges and Opportunities in Processing Historical Bavarian Dialects. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. A. Holdt, T. Munda, & A. Salgado (Eds.), *Electronic lexicography in the 21st century (eLex 2025): Intelligent lexicography: Proceedings of the eLex 2025 conference*. Bled, 18–20 November 2025 (pp. 457–479). Lexical Computing CZ s.r.o.

Contact information

Hanna Fischer
Philipps-Universität Marburg DSA | Forschungszentrum Deutscher Sprachatlas
hanna.fischer@uni-marburg.de

Alfred Lameli
Philipps-Universität Marburg DSA | Forschungszentrum Deutscher Sprachatlas
lameli@uni-marburg.de

Nathalie Mederake
Philipps-Universität Marburg DSA | Forschungszentrum Deutscher Sprachatlas
mederake@uni-marburg.de

Nico Urbach
Philipps-Universität Marburg DSA | Forschungszentrum Deutscher Sprachatlas
urbach@uni-marburg.de

Acknowledgements

We thank the three anonymous reviewers, whose detailed and constructive comments substantially improved this paper, particularly the presentation of the source material and the transparency of the evaluation datasets. We further thank the editorial teams of *Mecklenburgisches Wörterbuch*, *Pfälzisches Wörterbuch*, and *Schweizerisches Idiotikon* for providing access to the data used in this study. The authors gratefully acknowledge the computing time granted by the KISSKI project.

Katharina Korecky-Kröll, Philipp Stöckle, Daniel Elsner,
Wolfgang Koppensteiner & Elisabeth Eder

LLM-ASSISTED LEXICOGRAPHY IN LOW-RESOURCE CONTEXTS

From Prompt to Dictionary Article

Abstract This paper investigates the potential of Large Language Models (LLMs) to support lexicographic work in low-resource contexts, focusing on Austrian Bavarian dialects documented in the *Wörterbuch der bairischen Mundarten in Österreich (WBÖ)*. Using a structured prompt-engineering workflow and a three-stage data pipeline, 100 dictionary articles were generated with LLaMA 4 (Scout) and systematically compared to human-authored "gold" articles. Evaluation combined automated metrics, LLM-based judgement, and assessment by three human lexicographers. Results show that LLMs reliably reproduce formal lexicographic conventions, producing well-structured and coherent dictionary entries. Nevertheless, limitations emerge in semantic completeness and data fidelity: generated articles sometimes omit senses or misrepresent semantic structure and often inadequately integrate empirical evidence, especially regional information and attestations. While automated metrics indicate a relatively high similarity of original and generated senses based on their embeddings, both human evaluators and LLM-as-a-judge approaches reveal persistent weaknesses. However, the strong alignment between human judgements and the sense similarity metric suggests that embedding-based evaluation is a promising proxy for human assessment, whereas holistic LLM-as-a-judge scores appear comparatively strict and less transparent. The study argues for a hybrid evaluation framework and positions LLMs as assistive tools within human-in-the-loop lexicography.

Keywords LLM; low-resource languages; dictionary articles; German; prompt engineering

1 Introduction

This paper aims at investigating how Large Language Models (LLMs) can support lexicographic work on nonstandard linguistic varieties, which are considered a subset of low-resource languages. The study focuses on the *Wörterbuch der bairischen Mundarten in Österreich (WBÖ)*, a large dictionary of historical Bavarian dialects in Austria, situated at the Austrian Centre for Digital Humanities (ACDH) of the Austrian Academy of Sciences (ÖAW). Thus, this paper examines whether AI systems can semantically classify dialect material and generate dictionary entries that align with established scholarly standards. Because dialects are scarcely represented in LLM training datasets, they pose an ideal test case for assessing the limits of current models.

The WBÖ has documented Austrian Bavarian dialects since 1913 and is based on roughly three million handwritten paper slips containing headwords, pronunciation, regional labels, grammatical information, and dialect example sentences (e.g., Stöckle, 2021). About 2.4 million of these slips have been digitised and encoded in XML/TEI format for use in the Lexical Information System Austria (LIÖ).[1] TEI ensures long-term sustainability and preserves fine-grained metadata, but its rich and deeply nested structure makes it more difficult to process for LLMs. For this reason, the TEI entries were transformed into plain text while preserving all lexicographically relevant information (Bowers & Stöckle, 2018). All relevant entries of a lemma are organised into a lemma collection.

[1] Website: https://lioe.lapis-online.at/en (accessed 2026-06-02).

Previous research on LLMs in lexicography has shown strengths and weaknesses (e.g., De Schryver, 2023; Lew, 2023). Models such as ChatGPT can produce coherent definitions for simple, single sense words, but they often struggle with polysemous items, provide inauthentic example sentences, and occasionally generate hallucinated content. Additional challenges include inconsistent outputs for identical prompts, limited transparency regarding training data, and biases toward high-resource languages (Jakubíček & Rundell, 2023). These findings underscore the need for methods that strictly control what information an LLM receives. This study therefore employs Chain-of-Thought (CoT) Prompt Engineering as a rule-based automatic prompting and context retrieval system.

2 Project background

2.1 Previous work

The project entitled *Lexicography meets AI* started in 2025.[2] The first project phase (concluding at end of 2025) focused on two objectives: the development of a workflow for the automatic creation of dictionary articles by various proprietary and open-weight LLMs and the design of test strategies to evaluate the output of these models. While proprietary systems – especially Anthropic Claude 4 Sonnet – generally produced the best results, open-weight models such as LLaMA 3.3 (70B) and its successor LLaMA 4 (Scout) were prioritised, as they guarantee full data sovereignty and can optionally be finetuned on domain-specific material (see Stöckle et al., 2025).

This paper focuses on the next step in testing open-weight models: Following a brief description of the data pipeline used for the creation of the dictionary articles (slightly adapted from Stöckle et al. 2025), we apply different evaluation methods to compare 100 dictionary articles created by LLaMA 4 (Scout) with dictionary articles of the same 100 lemmas written by human lexicographers. The latter are considered to be the "gold standard" (henceforth referred to as "gold articles"), used as points of comparison for the LLM results (henceforth "generated articles"). The following section provides a brief overview of WBÖ article structure, which constitutes both the structural template for the LLM-generated articles and the reference point for their evaluation.

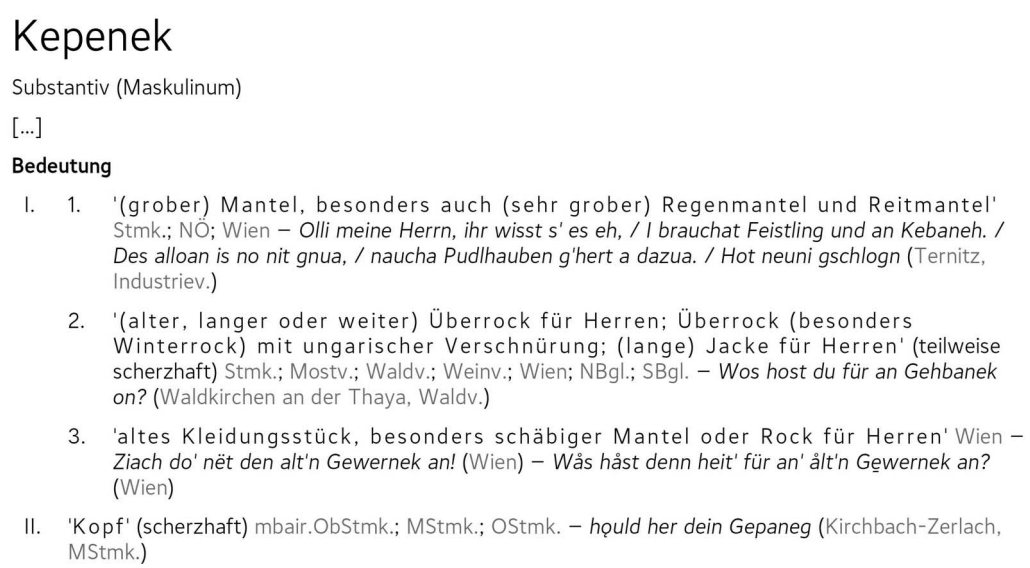
2.2 Article structure

WBÖ articles follow a detailed internal structure, which is illustrated in Figure 1 using the example of *Kepenek* ('coat; head').

Each article is headed by the lemma in its normalised, i.e., standardised, form. Originally – and consequently also in the database – lemmas were lemmatised following historical-etymological conventions (for details, see Stöckle, 2021, p. 19). As part of the modernisation of the WBÖ and its transition to an online format since 2016, it was decided that all headwords would be presented in a standardised form: where the lemma has a standard German equivalent, that form is used; where no standard German equivalent exists, a standard-like form is created. In the present example, the lemma appears in the WBÖ database as *Gepenek* but was standardised to *Kepenek* for the article.

[2] Project website: https://www.oeaw.ac.at/acdh/research/linguistics/research/lexicography/lexicography-meets-ai (accessed 2026-06-02).

The lemma is followed by basic grammatical information: for each entry, the part of speech is given, supplemented where necessary by further specification (e.g., inflectional paradigms for verbs, grammatical gender for nouns). All articles also contain information on geographical distribution and etymology; however, these two components are excluded from the current study and are therefore not discussed further here.



Kepenek

Substantiv (Maskulinum)

[...]

Bedeutung

I. 1. '(grober) Mantel, besonders auch (sehr grober) Regenmantel und Reitmantel' Stmk.; NÖ; Wien – *Olli meine Herrn, ihr wisst s' es eh, / I brauchat Feistling und an Kebaneh. / Des alloan is no nit gnua, / naucha Pudlhauben g'hert a dazua. / Hot neuni gschlogn* (Ternitz, Industriev.)

2. '(alter, langer oder weiter) Überrock für Herren; Überrock (besonders Winterrock) mit ungarischer Verschnürung; (lange) Jacke für Herren' (teilweise scherzhaft) Stmk.; Mostv.; Waldv.; Weinv.; Wien; NBgl.; SBgl. – *Wos host du für an Gehbanek on?* (Waldkirchen an der Thaya, Waldv.)

3. 'altes Kleidungsstück, besonders schäbiger Mantel oder Rock für Herren' Wien – *Ziach do' nët den alt'n Gewernek an!* (Wien) – *Wås håst denn heit' für an' ålt'n Gẹwernek an?* (Wien)

II. 'Kopf' (scherzhaft) mbair.ObStmk.; MStmk.; OStmk. – *họuld her dein Gepaneg* (Kirchbach-Zerlach, MStmk.)

FIGURE 1 WBÖ article *Kepenek*[3] ('coat; head') (excerpt)

The semantic core of each article is the definition of senses ('Bedeutung'), organised in a hierarchical structure in which main senses may be subdivided into subsenses. Where a sense category is directly subdivided, the main sense label is omitted and only the subsenses are given. In the *Kepenek* example, sense I comprises three subsenses, each designating a specific type of coat or traditional garment; sense II denotes 'head', with the additional pragmatic note that it is used jocularly ('scherzhaft').

Following the sense definitions, each sense is accompanied by regional information at the level of so-called major regions (Großregionen)[4], indicating where the respective sense is attested in the database. Where suitable examples are available, example sentences are provided; where no suitable examples exist, none are given, as the WBÖ adheres strictly to a data-driven approach. All example sentences are therefore reproduced in their original spelling. Each sentence is followed by the most specific regional information available, preferably at the community level. In the *Kepenek* example, example sentences are provided for all senses.

2.3 Aims of the current study

In order to assess the quality of LLM-generated articles against the human-authored WBÖ gold articles, nine different automated evaluation metrics (e.g., exact match of the lemma or part of speech, Sense Similarity, or the LLM-as-a-judge technique) were used to analyse the performance of LLaMA 4 (Scout) in creating WBÖ articles.

[3] The original article is published under: https://wboe.lapis-online.at/articles/Kepenek (accessed 2026-06-02).

[4] More detailed information on the regional classification of the WBÖ survey area can be found under: https://wboe.lapis-online.at/resources?link=materialien%2Fbearbeitungsgebiet%2F (accessed 2026-05-29).

Subsequently, three example articles created by LLaMA 4 (one with a high, one with a medium and one with a low Sense Similarity score) were manually evaluated by three human lexicographers in a triple-blind peer review (i.e., the lexicographers knew neither the evaluation reports by the other lexicographers nor the automated evaluation metrics, and the same was true for the LLMs).

The main results of these comparative analyses will serve as a basis for further discussion and for an outlook on future analyses.

3 Method

3.1 AI-based creation of dictionary articles

A three-stage data pipeline was developed:

1.a. First, TEI/XML data, lemma collection metadata, and a specially designed glossary were verbalised and serialised in JSON and TOON (Token-Oriented Object Notation)[5] format. This serialisation process was done for each of the 100 selected lemmas. The output is a JSON and TOON file for each lemma to which we will refer to as LLM-Corpus[6].

1.b. Article TEI/XML data (gold article) were verbalised and serialised in JSON based on a Pydantic Model[7] that was also used to enforce structured inference results as described later.

2. Second, based on a list of lemmas the rule-based system retrieved the relevant data from the LLM-Corpus. Both, the JSON and TOON format are suitable as context data. However, in this specific workflow the TOON format was used.

3. Third, the retrieved data were combined with structured prompts[8] that guided the LLM through semantic classification and dictionary-entry generation.

A chain-of-thought (CoT) prompting strategy was used in three main steps:

1. analyse the lexicographic content of the input,

2. extract and group senses into coherent semantic categories,

3. generate a dictionary entry following the WBÖ format.

For each consecutive prompt the generated response was used as context in the next prompt. This sequence mirrors the workflow of human lexicographers and aims to enforce transparency and discipline in the model's reasoning. Besides the manually enforced CoT prompting strategy the prompts included instructions to analyse step-by-step and document the steps before generating an answer.

[5] TOON is a data format that promises about 70% higher accuracy and 40% lower token consumptions in comparison with JSON and was designed for model inference (see https://toonformat.dev).

[6] LLM-Corpus dataset (see https://github.com/acdh-oeaw/llm-assisted-dialect-lexicography/tree/main/inference-results/v2/simplex/llm_corpus)

[7] Pydantic is a widely used validation library for Python (see https://pydantic.dev).

[8] System, Prompt1, Prompt2, Prompt3 (see https://github.com/acdh-oeaw/llm-assisted-dialect-lexicography/tree/main/prompts/v2)

For the final prompt the model received the already mentioned Pydantic Model that ensured a valid structured response (JSON). A sample of 100 lemmas was tested using LLaMA 4 (Scout) with the following parameter settings: 4-bit quantisation, temperature: 0, seed: 1337[9].

Based on the structured response schema an automatic evaluation workflow was created. Evaluation focused on structural accuracy, semantic classification, and faithfulness to the data used. To easily evaluate the generated senses and other categories the LLM had generated, the previously mentioned Pydantic Model was used to serialise the human written dictionary articles from TEI/XML to JSON[10]. Throughout the three different datasets (LLM-Corpus, gold-articles and generated articles) the terminology was harmonised to ensure a clear connection among the three sources.

3.2 Automated evaluation metrics

The following automated evaluation metrics were used:

1. Exact Match Lemma: We evaluate whether the gold lemma and the generated lemma match (using a binary evaluation scale).
2. Phrase Match POS (part-of-speech): This metric assesses if the gold POS contains the generated POS to catch abbreviations (binary evaluation scale).
3. Sense Absolute Error: Here, we calculate the Absolute Error between the number of gold and number of generated senses. If, for example, there are 4 generated and 7 gold senses, the Absolute Error is 3).
4. Sense Signed Error: Calculates the Signed Error between the number of gold and number of generated senses. If, for example, there are 4 generated and 7 gold senses, the Signed Error is –3. Negative values indicate that model underestimates, positive values indicate that it overestimates.
5. Sense Similarity: Given the gold senses and the generated senses we calculate pairwise cosine similarity of their embeddings using jinaai/jina-embeddings-v5-text-small[11] (Akram et al., 2026). Then, we match each gold sense to the generated sense with the highest similarity for recall. For precision, we match each generated sense to the gold sense with the highest similarity. The final Sense Similarity is the F1 score. This procedure is similar to the calculation of BERTScore (Zhang et al., 2020), although not performed on token level but rather on the senses themselves.
6. Matched Example IDs Proportion: This metric focuses the proportion of record IDs (Beleg-IDs) that exist in the LLM-Corpus and are not hallucinated or wrong. If a lemma does not have any examples, we return '1' (equivalent to 'best score').
7. Examples ANLS Score: We calculate the Average Normalised Levenshtein Similarity (ANLS, Biten et al., 2019) for every generated example sentence (Belegsatz) and its counterpart in the LLM-Corpus if its record ID (Beleg-ID) is in the prompt context and return the average of these scores. If a lemma does not have any examples, we return '1' (equivalent to 'best score').

[9] Inference Results (see https://github.com/acdh-oeaw/llm-assisted-dialect-lexicography/tree/main/inference-results/v2/simplex/generated_articles)

[10] Gold Articles (see https://github.com/acdh-oeaw/llm-assisted-dialect-lexicography/tree/main/inference-results/v2/simplex/gold_articles).

[11] Jina AI's multilingual embedding models (see https://huggingface.co/jinaai/jina-embeddings-v5-text-small).

8. Regions Examples ANLS Score: We calculate the Average Normalised Levenshtein Similarity (ANLS, Biten et al., 2019) for every location of origin (Herkunftsort) of an example sentence and the location or region in the LLM-Corpus if its record ID is in the prompt context and return the average of these scores. If a lemma does not have any examples, we return '1' (equivalent to 'best score').

9. Dictionary Article Correctness: We use the LLM-as-a-judge technique and DeepEval's[12] implementation of the G-Eval framework (Liu et al., 2023) to let an LLM, Qwen3-32B (Yang et al., 2025) in our case, evaluate the correctness of our generated dictionary articles. The scores range from 0.0 to 1.

4 Results

4.1 *Automated evaluation metrics: LLM-generated vs gold articles*

The results of the descriptive statistics of the first eight automated evaluation metrics comparing the 100 generated dictionary articles to their gold article counterparts are shown in Table 1. Metrics 1, 2, 5, 6, 7, and 8 are scores between 0 and 1 (which may be interpreted as percentages between 0 % and 100 %), whereas metrics 4 and 5 represent absolute numbers of errors. The two binary metrics (1 and 2) reach quite different scores: While the lemma in the record database has often not yet been normalised (this is another project our group is currently working on) and therefore matches the already normalised lemma in the dictionary article only in 51 % of the cases, the part of speech is mostly correctly extracted from the database and therefore agrees with the gold article in 98 % of the cases. Example IDs (metric 6) are represented similarly well, as they match in 96 % of the cases, while this rate is considerably lower (52 %) for the geographical information related to locations and regions. Sense Similarity (metric 5) may be considered the most important metric regarding semantic classification – it matches the gold article at a mean rate of 81 %.

TABLE 1 Descriptive statistics of the automated evaluation metrics 1-8 based on 100 lemmas/dictionary articles (mean, standard deviation = stdev, median, minimum = min., maximum = max.)

Metric	Mean	stdev	Median	min.	max.
1. Exact Match Lemma	**0.51**	0.50	1.00	0.00	1.00
2. Phrase Match POS	**0.98**	0.14	1.00	0.00	1.00
3. Sense Absolute Error	3.74	6.12	2.00	0.00	41.00
4. Sense Signed Error	-0.66	7.14	0.00	-36.00	41.00
5. Sense Similarity	**0.81**	0.07	0.81	0.59	0.99
6. Matched example IDs Proportion	**0.96**	0.15	1.00	0.00	1.00
7. Examples ANLS Score	**0.52**	0.42	0.50	0.00	1.00
8. Region Examples ANLS Score	**0.77**	0.34	0.94	0.00	1.00

[12] DeepEval: The LLM Evaluation Framework. Retrieved May 29, 2026, from https://deepeval.com/.

4.2 Automated evaluation metrics: LLM judges LLM

Figure 2 shows the results of the Dictionary Article Correctness score (metric 9) judged by the Qwen3-32B LLM: Most lemmas were judged as 0.3 and 0.4, resulting in a mean of 0.38 (stdev = 0.11) and a median of 0.40, while 0.0 was the minimum and 0.7 the maximum. The LLM's judgement was quite strict, as no single lemma was judged above 0.7.

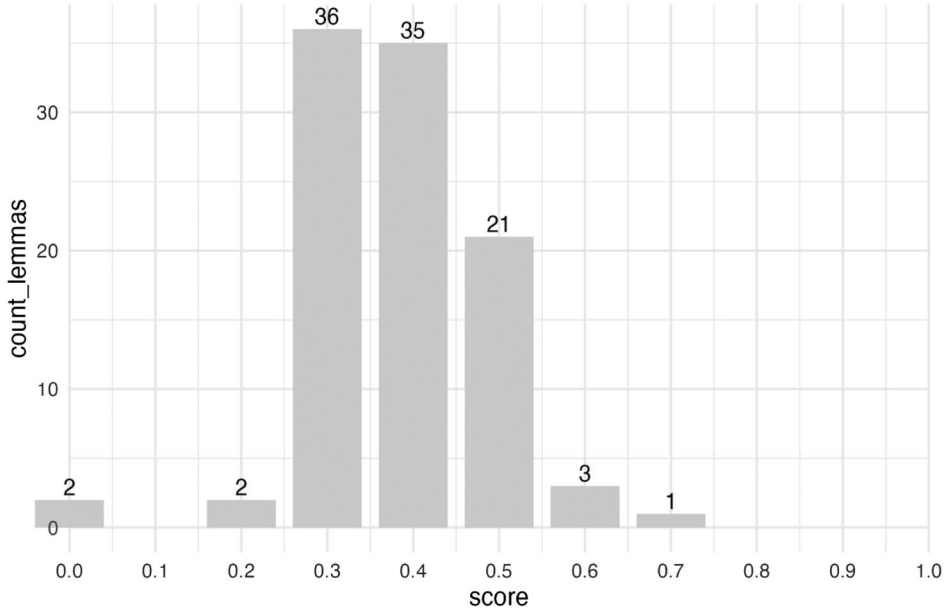


FIGURE 2 Number of lemmas per Dictionary Article Correctness score (0.0 to 1.0) provided by Qwen3-32B (descriptive statistics: mean = 0.38, stdev = 0.11, median = 0.40, min = 0.0, max = 0.7)

In addition to the Dictionary Article Correctness score, Qwen3-32B was also asked to justify its scores. For reasons of space, only the three lemmas that were also evaluated by human lexicographers can be discussed in detail. Table 2 gives an overview of the three lemmas, their basic senses and their automated evaluation metrics (1-8), the Dictionary Article Correctness score (9) as well as its justification by Qwen3-32B. If we compare e.g., Sense Similarity (5) and Dictionary Article Correctness (9), we see immediately that the LLM judgement is rather strict.

4.3 Human evaluation: LLM-generated vs gold articles

Now let us focus on the human evaluation of the AI-generated dictionary entries for *Kepenek*, *kommod*, and *gucken*: It shows a high degree of agreement among the three lexicographers with respect to structural quality. Across all lemmas, the entries are consistently described as very well structured, closely following the prompt specifications and established lexicographic conventions (e.g. clear subdivision into semantic categories, systematic use of region numbers, and absence of redundant examples). Minor differences concern the completeness of header information and the consistency with which example sentences are provided, but overall structural evaluations range from good to very good, indicating strong inter-annotator consensus in this domain.

Assessments of semantic classification reveal moderate agreement combined with noticeable divergence in detail. All three lexicographers agree that the semantic structure only partially corresponds to the WBÖ reference articles, with mismatches in the number and granularity of semantic categories and incomplete coverage of senses. However, individual judgements differ

TABLE 2 Automated evaluation metrics 1-8, Dictionary Article Correctness (9) and its justification (10) by Qwen3-32B

Metrics	Lemma *Kepenek*	Lemma *kommod*	Lemma *gucken*
	'coat; head'	'comfortable; sociable'	'to look; to play (cards; tag; hide and seek)'
1. Exact Match Lemma	0.00	0.00	1.00
2. Phrase Match POS	1.00	1.00	1.00
3. Sense Absolute Error	2	11	1
4. Sense Signed Error	-2	-11	1
5. Sense Similarity	0.91	0.83	0.71
6. Matched example IDs Proportion	1.00	1.00	1.00
7. Examples ANLS Score	0.00	0.33	0.46
8. Region Examples ANLS Score	0.96	0.65	0.66
9. Dictionary Article Correctness (Qwen3-32B)	0.00	0.30	0.40
10. Justification by Qwen3-32B (translated from German to English and shortened)	The lemmas do not match. The PoS is correct but several senses from the expected output (e.g., old garment or the second record for head) are completely missing. The grouping of the 'head' sense as subcategory is not expected. Names and codes of regions differ, and example sentences are mostly incomplete or missing.	Not all senses are there, e.g. senses like *spacious* or *easily accessible* are missing. The grouping is not complete. Regions are missing or show deviant codes, and example sentences are incomplete or missing.	The main sense is partially there, but important subcategories like playing (cards, hide-and-seek etc.) are completely missing. The grouping of the subcategories is not expected. The region codes and the regions of the senses do not match. Only some example sentences from the main sense were partially provided.

regarding the adequacy of semantic differentiation and coherence: while one evaluator rates the semantic categorisation as very good in some cases, others describe it as only moderately successful, pointing to over- or under-differentiation and missing senses. Despite these differences, there is shared recognition that category labels are generally comprehensible and that example sentences mostly fit their assigned categories.

The strongest consensus emerges in the evaluation of data fidelity, which is uniformly identified as the weakest aspect of the generated entries. All lexicographers note substantial problems in the assignment of attestations to semantic categories, the coverage of available corpus evidence (with less than half of attestations typically used), and especially the accuracy of regional information linked to examples. While evaluative severity varies slightly, data fidelity is consistently rated as merely sufficient to satisfactory. Overall, the human lexicographers rate the articles generally higher than Qwen3-32B (see Table 2) and show more agreement with the other automated metrics, particularly the Sense Similarity score (metric 5). This score might be the most useful of the automated scores, as it strongly differentiates high, medium and low-quality articles in a similar way as the human lexicographers.

4.4 Example analysis of one LLM-generated article: Similarities and differences between judgements by three human lexicographers and one LLM

To illustrate these general results, we take now a closer look at the LLM-generated article *Kepenek (Gepenek)* (see Figure 3) followed by its detailed evaluation by the three human lexicographers (see Table 3).

Gepenek

Subst (m, n?)

Bedeutung

I. Mantel, Überrock; auch weiter Männerüberrock, Reitmantel oder grober Mantel 6.3 Waldv., 6.5 Wien, 6.1 Mostv., 6.4 Weinv., 7.1 NBgl. – *Wos host du für an Gehbanek on?* 6.3a24 Waldkirchen an der Thaya

1. Kopf 3.2 mbair.ObStmk.; 3.4 MStmk. – *hǫuld her dein Gepaneg* 3.4c14 Kirchbach-Zerlach

FIGURE 3 LLM-generated article *Kepenek* (*Gepenek*) (layout reformatted to match the original WBÖ article)

The first column of Table 3 lists 28 different evaluation parameters that had already been used in the previous study by Stöckle et al. (2025): Parameters 1–12 assess the formal structure of the article, parameters 13–23 the semantic classification and parameters 24–28 the data fidelity. Columns 2–4 of Table 3 list the mostly verbal assessments of the three lexicographers (translated into English and slightly shortened).

The human evaluation by three lexicographers paints a largely positive but differentiated picture of the LLM–generated dictionary articles. In terms of structure, the outputs are consistently judged as highly organised: all three evaluators rate the overall structuring as "(very) structured," with most formal requirements (header, semantic subdivision, region numbering) implemented correctly. The overall structural quality is therefore assessed as "good" across evaluators, despite recurring minor issues such as incomplete grammatical information, missing example sentences in some categories, and insufficient or inconsistent inclusion of record IDs.

For the semantic classification, the results are more mixed but still clearly positive overall. The generated article generally captures at least the main semantic categories of the gold article, though often with deviations in the number or granularity of categories (under–differentiation in this case). Evaluators note that "some categories overlap" or that "the main categories are present," indicating partial but not full alignment. The internal coherence of categories ranges from "partly" to "good," and example sentences are usually judged as fitting "well"

TABLE 3 Evaluation of the LLM-generated *Kepenek (Gepenek)* article by three lexicographers based on 28 parameters

Evaluation parameters	Lexicographer 1	Lexicographer 2	Lexicographer 3
Header structure: basic information (lemma, PoS, grammar) provided?	Partially	Completely	Partially
Basic structure: article well-structured?	Very well	Very well	Very well
Regions correctly provided for each semantic category?	Rather correct	Rather correct	Rather correct
Example sentences provided for each semantic category?	Yes, 1 sentence on average	Yes, 1 sentence on average	Yes, 1 sentence on average
Regional information provided for each example sentence?	Always	Always	Always
Record IDs correctly provided for each semantic category?	Rather incorrect	Rather correct	Rather incorrect
Region numbers provided for each semantic category?	Always	Always	Always
Are semantic categories subdifferentiated?	Always	Always	Always
Subdifferentiation according to the schema given in the prompts?	Always	Always	Always
Are example sentences provided like in the original (e.g., Teuthonista)?	Always	Always	Always
Are example sentences provided more than once?	Never	Never	Never
General assessment of the article structure	Good	Satisfactory	Good
Number of semantic categories generated = number in the gold article?	Yes (main categories (MC)), but fewer subcategories (SC)	Fewer	Fewer
Number of semantic categories generated	2 (1 MC, 1 SC)	2	2 (1 MC, 1 SC)
To what extent do semantic categories correspond to those in the gold article?	MCs provided	MCs provided	MCs provided
Semantic coherence SCs and their corresponding MCs?	Always	Always	Partially
How well does the example sentence fit to the semantic category?	Very well	Well	Well

TABLE 3 (*CONTINUED*)

Evaluation parameters	Lexicographer 1	Lexicographer 2	Lexicographer 3
How well were the data assigned to semantic categories?	Very well	Well	Rather well
How well were the data subdifferentiated by the semantic categories?	Appropriately	Appropriately	Appropriately
Minimum number of records used to form a semantic category?	2	2	2
Coherence/fit of record IDs per subcategory?	Very high fit	Very high fit	Very high fit
How clearly were the semantic category labels formulated?	Clearly	Clearly	Clearly
General assessment of the semantic classification	Very good	Satisfactory	Satisfactory
Do all records provided exist?	Yes	Yes	Yes
Do all regions per semantic category correspond to the record ID provided?	Partially (more than 50%)	Partially (more than 50%)	Partially (more than 50%)
Were all record IDs assigned to a semantic category?	Fewer than 50%	Fewer than 50%	Fewer than 50%
Regional information correct for all example sentences?	Completely	Completely	Completely
General assessment of data fidelity	Good	Good	Satisfactory

or "reasonably well." As a result, the overall semantic evaluation lies mostly between "satisfactory" and "good," with one instance reaching "very good."

By contrast, data fidelity is the weakest dimension. While all cited evidence IDs exist in the source collection (a clear strength), their use and mapping are often problematic: fewer than half of the available data points are typically incorporated into the semantic categories, regional attributions are frequently incorrect or missing, and the correspondence between regions and evidence IDs is often rated as "predominantly incorrect." Consequently, the overall assessment of data fidelity falls between "good" and "satisfactory".

In sum, the human evaluation shows that the LLM performs strongly on formal structure, reasonably on semantic abstraction, but weakly on faithful and complete use of underlying data – a profile that contrasts with the reported poor results of the LLM-based evaluation (see metrics 9 and 10 in Table 2) and highlights the importance of expert human judgement in this task.

5 Discussion

The results of this study highlight a clear asymmetry in the performance of LLaMA 4 (Scout) in dialect lexicography: while the model reliably produces structurally well-formed dictionary articles, it continues to struggle with semantic completeness and faithful use of empirical data. This pattern emerges consistently across automated metrics, LLM-based evaluation, and

human expert judgement. Structural aspects such as article layout, hierarchical organisation of senses, and formal adherence to the WBÖ article schema are handled robustly, suggesting that LLMs can successfully internalise and reproduce explicit lexicographic conventions when these are strongly scaffolded through prompt design and structured output constraints. In this respect, the model behaves comparably to a technically competent but domain-novice lexicographer.

In contrast, semantic classification and data fidelity remain the main bottlenecks. Automated metrics indicate reasonably high Sense Similarity scores, yet both human lexicographers and the LLM-as-a-judge identify systematic problems: missing senses, deviations from established semantic groupings, insufficient coverage of attestations, and inaccurate or incomplete regional assignments. Importantly, the human evaluation reveals that these weaknesses are not uniformly distributed but depend on the lexical item and on the internal semantic complexity of the lemma. The moderate inter-annotator disagreement observed in semantic evaluation further underlines that semantic categorisation itself is a cognitively demanding task, even for experts, and that current automated metrics can only approximate this complexity. At the same time, the relatively strong alignment between human judgements and the Sense Similarity metric suggests that embedding-based semantic evaluation provides a promising proxy for human assessment, whereas holistic LLM-as-a-judge scores appear comparatively strict and less transparent.

Regarding the treatment of human-authored articles as a "gold standard", an important clarification is necessary: there is no single, definitive article against which all outputs must be measured. Lexicographic description inherently involves interpretive decisions, meaning that two lexicographers writing entries for the same lemma would almost certainly produce non-identical articles (as the inter-annotator divergence observed above illustrates). The gold article should therefore be understood not as an absolute ground truth, but rather as the best available reference, i.e., a benchmark that is itself the product of expert judgement.

From a methodological perspective, the combined evaluation approach adopted here proves essential. Purely automated metrics would overestimate the quality of the generated articles, particularly with respect to semantic adequacy, while human evaluation alone would not scale beyond a small sample. The results therefore argue for a hybrid evaluation paradigm in which automated metrics are used for large-scale screening and ranking, and targeted human review is applied where semantic precision and data fidelity are critical. More generally, the findings caution against viewing LLM-generated dictionary articles as near-final products in low-resource and historically grounded lexicographic settings. Instead, they support a workflow in which LLMs function as assistive tools for structuring, drafting, and pre-classifying material, with expert lexicographers retaining full responsibility for semantic interpretation and empirical validation. In this sense, LLMs already offer substantial efficiency gains, but only within a carefully controlled human-in-the-loop framework.

6 Conclusion and outlook

This study has shown that LLaMA 4 (Scout) can generate structurally coherent and formally well-aligned dictionary articles for a scientific dialect dictionary when guided by carefully designed prompts and structured output constraints. Both automated evaluation and human expert judgement indicate that the model reliably reproduces established lexicographic conventions and article layouts, confirming its potential as an effective assistive tool in lexicographic workflows. At the same time, the analyses clearly demonstrate that semantic completeness and data fidelity remain major challenges, particularly with respect to sense coverage, semantic granularity, and the accurate integration of regional and corpus-based evidence.

These limitations are especially salient in historically grounded, low-resource dialect contexts such as the WBÖ.

Looking ahead, the results suggest several promising directions for future work. Concretely, the workflow presented in Section 3 could be used to generate draft articles for larger groups of lemmas and automatically assess their quality using the evaluation metrics introduced in this study. Human lexicographers could then review, correct, and validate these drafts prior to publication, directing their attention especially to entries flagged by the metrics as potentially problematic. On the methodological side, further refinement of hybrid evaluation strategies is needed, combining scalable automated metrics with targeted human review in areas where semantic interpretation is critical. The assessment of further parameters would additionally broaden the evaluative scope of this paper's endeavour. This, amongst other, includes the impact of the LLM's training data on dictionary article generation. Only half of the lemmas, i.e., WBÖ articles written by human lexicographers have already been published online via LIÖ (cf. Introduction). The other half of the articles are not yet publicly available. The online published articles might have been included in the LLM's training data. It, thus, will be worth evaluating whether there are differences in LLM-generated articles depending on the publication status of the respective lemma. In addition, all lemmas were prompted twice to investigate the deterministic extent of the results generated. This is amongst the evaluative tasks scheduled next in our project.

From a modelling perspective, improvements may be achieved through enhanced data preparation methods and selection mechanisms, selective fine-tuning on curated lexicographic material, the application of an optimised prompting-strategy, the inclusion of new data models for structured output, and the implementation of additional LLMs. Since the project start in 2025, LLaMA (initially 3:70B, later 4) yielded promising results compared to other LLMs (cf. Stöckle et al., 2025), resulting in designing and applying the workflow presented in this paper. However, due to the astoundingly rapid development in the field of LLMs, it will be worth evaluating future models to both further enhance and automatise the workflow's tasks.

The automatic context retrieval, inference and evaluation workflow that was described provides a solid fundament for adjusting various parameters to improve the results. Next to testing other LLMs (open-weights and proprietary) this workflow can test changes in model parameters like temperature as well; with automated evaluation metrics showing improvement or decline. In conclusion, our findings support a view of current LLMs not as autonomous lexicographers, but as powerful human-in-the-loop assistants that can substantially increase efficiency while leaving semantic interpretation and empirical validation firmly in expert hands. Under these conditions, LLM-assisted approaches hold considerable promise for the sustainable development of large-scale dialect dictionaries in the digital age.

References

Bowers, J. & Stöckle, P. (2018). TEI and Bavarian dialect resources in Austria: updates from the DBÖ and WBÖ. In A.U. Frank, C. Ivanovic, F. Mambrini, M. Passarotti & C. Sporleder (Eds.), Proceedings of the Second Workshop on Corpus-Based Research in the Humanities (CRH-2), 25–26 January 2018 Vienna, Austria (pp. 45–54). Gerastree Proceedings. Retrieved April 9, 2026, from https://www.oeaw.ac.at/fileadmin/subsites/academiaecorpora/PDF/CRH2.pdf

De Schryver, G.-M. (2023). Generative AI and Lexicography: The Current State of Art Using ChatGPT. International Journal of Lexicography, 36(4), pp. 355–387. https://doi.org/10.1093/ijl/ecad021

Jakubíček, M. & Rundell, M. (2023). The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography? In M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček &

S. Krek (Eds.), eLex 2023. Electronic lexicography in the 21st century (pp. 518–532). Lexical Computing CZ.

Lew, R. (2023). ChatGPT as a COBUILD lexicographer. Humanities & Social Sciences Communications, 10(704), pp. 1–10. https://doi.org/10.1057/s41599-023-02119-6

Stöckle, P. (2021). Wörterbuch der bairischen Mundarten in Österreich (WBÖ). In A. N. Lenz & P. Stöckle (Eds.), *Germanistische Dialektlexikographie zu Beginn des 21. Jahrhunderts* (pp. 11–46). Steiner. https://doi.org/10.25162/9783515129206.

Stöckle, P., Elsner, D., Koppensteiner, W. & Korecky-Kröll, K. (2025). LLM-Assisted Dialect Lexicography: Challenges and Opportunities in Processing Historical Bavarian Dialects. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. A. Holdt, T. Munda & A. Salgado (Eds.), *Electronic lexicography in the 21st century (eLex 2025): Intelligent lexicography. Proceedings of the eLex 2025 conference. Bled, 18–20 November 2025* (pp. 453–475). Lexical Computing CZ s.r.o. Retrieved April 9, 2026, from https://elex.link/elex2025/wp-content/uploads/eLex2025-28-Stockle_etal.pdf

Websites:

Akram, M. K., Sturua, S., Havriushenko, N., Herreros, Q., Günther, M., Werk, M. & Xiao, H. (2026). jina-embeddings-v5-text: Task-Targeted Embedding Distillation. Retrieved May 29, 2026, from https://doi.org/10.48550/arXiv.2602.15547

Biten, A. L., Tito, R., Mafla, A., Gomez, L., Rusiñol, M., Valveny, E., Jawahar, C. V. & Karatzas, D. (2019). Scene Text Visual Question Answering. Retrieved May 29, 2026, from https://doi.org/10.48550/arXiv.1905.13648

DeepEval: The LLM Evaluation framework. Retrieved May 29, 2026, from https://deepeval.com/

Kepenek: WBÖ dictionary article. Retrieved May 29, 2026, from https://wboe.lapis-online.at/articles/Kepenek

Lexicography meets AI: Project website. Retrieved May 29, 2026, from https://www.oeaw.ac.at/acdh/research/linguistics/research/lexicography/lexicography-meets-ai

Liu, Y., Iter, D., Xu, Y., Wang, S., Xu, R. & Zhu, C. (2023). G-Eval: NLG Evaluation using GPT-4 with Better Human Alignment. Retrieved May 29, 2026, from https://doi.org/10.48550/arXiv.2303.16634

Pydantic: Python library. Retrieved May 29, 2026, from https://pydantic.dev

TOON: Token-Oriented Object Notation. Retrieved May 29, 2026, from https://toonformat.dev

WBÖ survey area: Regional classification. Retrieved May 29, 2026, from https://wboe.lapis-online.at/resources?link=materialien%2Fbearbeitungsgebiet%2F

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., Zheng, C., Liu, D., Zhou, F., Huang, F., Ge, H., Wei, H., Lin, H., Tang, J., Yang, J., Tu, J., Zhang, J., Yang, J., Yang, J., Zhou, J., Zhou, J., Lin, J., Dang, K., Bao, K., Yang, K., Yu, L., Deng, L., Li, M., Xue, M., Li, M., Zhang, P., Wang, P., Zhu, Q., Men, R., Gao, R., Liu, S., Luo, S., Li., T., Tang, T., Yin, W., Ren, X., Wang, X., Zhang, X., Ren, X., Fan, Y., Su, Y., Zhang, Y., Zhang, Y., Zhang, Y., Wan, Y., Liu, Y., Wang, Z., Cui, Z., Zhang, Z., Zhou, Z. & Qiu, Z. (2025). Qwen3 Technical Report. Retrieved May 29, 2026, from https://doi.org/10.48550/arXiv.2505.09388

Zhang, T., Kishore, V., Wu, F., Weinberger, K. Q. & Artzi, Y. (2020). BERTScore: Evaluating Text Generation with BERT. Retrieved May 29, 2026, from https://doi.org/10.48550/arXiv.1904.09675

Contact information

Katharina Korecky-Kröll
Austrian Academy of Sciences
katharina.korecky-kroell@oeaw.ac.at

Philipp Stöckle
Austrian Academy of Sciences
philipp.stoeckle@oeaw.ac.at

Daniel Elsner
Austrian Academy of Sciences
daniel.elsner@oeaw.ac.at

Wolfgang Koppensteiner
Austrian Academy of Sciences
wolfgang.koppensteiner@oeaw.ac.at

Elisabeth Eder
Austrian Academy of Sciences
elisabeth.eder@oeaw.ac.at

Iztok Kosem, Polona Gantar & Tjaša Arčon

DIVIDE AND DEFINE WITH LLMS

Generating definitions for the Digital Dictionary Database of Slovene

Abstract This paper presents an experiment on generating definitions for Slovene headwords using four Large Language Models (LLMs): two commercial (Gemini and GPT) and two open-source (Gemma and the Slovenian model GaMS). Each model was provided with contextual data, including semantic indicators, collocations, and examples. We tested a zero-shot approach and two few-shot approaches using sample definitions from the Digital Dictionary Database of Slovene (DDDS) and the New Dictionary of Standard Slovene. The results show that while all models produce high-quality definitions, Gemini performs best across all word classes, particularly when utilizing DDDS sample definitions. The second part of the study evaluated LLMs as judges of definition quality, testing Gemini, GPT, and Claude-Sonnet. Gemini again achieved the highest agreement with human lexicographers. Notably, even when LLMs selected a different definition than the experts, their choices were usually viable. These results demonstrate the considerable potential of LLMs in lexicography, and we plan to deploy this pipeline to generate definitions for a much larger set of Slovenian headwords.

Keywords LLM; definitions; dictionary database; evaluation; Slovene

1 Introduction

The Digital Dictionary Database of Slovene (DDDS; Gantar, 2000; Kosem et al., 2021) represents a fundamental shift in Slovenian lexicography, aiming to describe the modern language from scratch through a modular, data-driven approach. Unlike traditional dictionaries, the DDDS serves as a central repository where various lexicographic projects—such as the Thesaurus of Modern Slovene, the Collocations Dictionary of Modern Slovene, and the Comprehensive Slovenian-Hungarian Dictionary—interact. While these projects contribute specific microstructural layers (e.g., bilingual equivalents or collocation strings), they are anchored by a shared sense structure.

Currently, the semantic description of senses in the DDDS consists primarily of brief semantic indicators and grammatical or pragmatic labels. To evolve into a comprehensive monolingual resource, the database requires the addition of definitions. However, the manual drafting of definitions remains the most time-consuming and cognitively demanding aspect of the lexicographic workflow.

In this paper, we present a two-fold experiment on definition generation with Large Language Models (LLMs). After briefly summarizing related work, we first present the first part of the experiment in which definitions were generated with four LLMs, and then evaluated by two lexicographers. Next, we tested three LLMs in the LLM-as-a-judge task of evaluating generated definitions, comparing their performance with human evaluations. We conclude by summarizing the main findings and their relevance for our work, as well as outlining future plans.

2 Related work

Explaining what a word means is one of the fundamental tasks of a monolingual explanatory dictionary (Atkins & Rundell, 2008, p. 405), and at the same time one of the most challenging

lexicographical tasks, which has been and continues to be the subject of philosophical, semantic, and lexicographical debates. Explaining the meaning of words using other words of the same language has never been a trivial task. It requires lexicographers to make and adhere to clear decisions about what the definition will say, how it will say it, and for whom it is intended. There are multiple possible decisions here, as evidenced by the fact that different dictionaries define the same concepts in different ways (Atkins & Rundell, 2008, p. 406).

The rapid evolution of LLMs has sparked a global debate within the field. An increasing number of researchers have begun exploring the potential of LLMs to facilitate, or even completely take over, various parts of the lexicographic workflow (Rundell, 2023; Lew, 2023; de Schryver, 2023; Rundell & Jakubiček, 2023; Bartosz et al., 2024; McKean & Fitzgerald, 2024; Tiberius et al., 2024), with many studies showing that LLMs work particularly well in generating definition and examples being given the most attention (e.g. Lew, 2023; Ide et al., 2026; Rees & Lew, 2025). While English-centric models have shown high proficiency, their application to morphologically rich languages with smaller speaker bases, such as Slovene, remains a critical area of investigation.

Some studies have already been done on Slovene. For example, Gantar's (2024) tested LLMs (GPT) on generating definitions for senses of phraseological units, reporting adequate definitions in 60% of cases, with highly idiomatic cases proving particularly challenging. Kosem et al. (2024) performed a task on generating senses for headwords based on examples of collocations, reporting satisfactory results for sense division and definitions in approx. 70% of cases. However, definition generation was not the main focus of the experiment, and the authors postulated that it might be better to break the workflow into discrete micro-tasks—specifically separating sense division from definition drafting.

3 Definition generation task

The Digital Dictionary Database of Slovene currently contains nearly half a million independent lexical units (LU) that include single-word and multi-word (compounds and phraseological units) LUs. Approximately 22,000 LUs have been completed and include senses with semantic indicators, collocations, examples, synonyms and antonyms etc. Other LUs contain only automatically obtained content such as collocations (these LUs represent the majority) or are in one of the stages of lexicographic workflow. None of the completed senses in the DDDS currently include definitions, mainly on account of the fact that definitions were not considered an essential feature in the compilation of the resources such as the Comprehensive Slovenian-Hungarian Dictionary currently comprising the DDDS. Moreover, definition writing is one of the most time-consuming parts of lexicographic workflow.

As part of the Large Language Models for Digital Humanities (LLM4DH) project, we are customizing GaMS, a Slovenian Large Language Model, for lexicographic tasks. Our development framework involves testing and benchmarking this model against others across several key tasks: sense division, definition and semantic indicator generation, labelling, collocation validation, and synonym distribution. Based on previous findings and related literature, we test and evaluate the LLMs for each task independently. Ultimately, the main objective is to design an LLM-driven lexicographic workflow that also clearly identifies where human lexicographers' input is most critically needed. In the remainder of the paper, we focus on definition generation, which will be used for existing manual entries and on completely new entries with senses generated by LLMs.

3.1 Methodology

For the definition generation experiment, we needed three components:

- a sample of headwords
- examples of good dictionary definitions
- LLM models

Headwords were selected from the entries where sense division had already been finalized by lexicographers. We selected a heterogeneous sample of 42 headwords, carefully balanced to ensure representativeness across several variables, such as word class (10 nouns, 10 verbs, 12 adjectives, 10 adverbs), polysemy (most were polysemous, few were monosemous), corpus frequency, contextual information etc. All of the senses of each headword were included in the task, making it 114 senses in total. Each sense included semantic indicator, some also labels, and contextual information such as collocations and examples.

For the set of examples of good dictionary definitions, we selected 174 lexical units (LUs) (111 from the eSSKJ and 63 from the DSBS), together with the definitions of their individual senses, yielding a total of 344 definitions. For polysemous LUs, we selected either a definition for only one of the senses or, in some cases, definitions for several senses of the same LU. The sample entries were chosen with regard to part-of-speech representation, monosemy versus polysemy, and the inclusion of concrete, abstract, and metaphorical senses. We also included words with potentially offensive meanings, for example, cepec ('idiot'), which is defined as: *if we call someone a cepec, we consider them a foolish and thoughtless person*. In selecting the sample definitions, we included both classical dictionary definitions and full-sentence definitions, reflecting the defining strategies employed in the DSBS (cf. Gantar, 2015).

(a) full-sentence definitions from the DSBS

enosmerno ('one-way')

1. ki poteka zgolj od izhodišča proti cilju
2. če kdo vrednoti dejanja ali človeka enosmerno, pri tem ne upošteva vseh dejstev in objektivnih meril

1. *which proceeds only from the point of origin towards a destination*
2. *if someone evaluates actions or a person in a one-sided way, they do not take into account all relevant facts and objective criteria*

avtonomen ('autonomous')

1. avtonomno ozemlje ima politično samoupravo znotraj večje države
2. če je ustanova ali oseba avtonomna, je samostojna in deluje ter sprejema odločitve neodvisno in brez pomoči drugih

1. *an autonomous territory has political self-government within a larger state*
2. *if an institution or a person is autonomous, they are independent and act and make decisions on their own, without relying on others*

(b) classical dictionary definitions from the eSSKJ

kokošnjak ('henhouse')

1. prostor za bivanje, gojenje kokoši

2. stavba, prostor, kjer biva, se zadržuje več ljudi tesno skupaj, ali neurejen, hrupen prostor sploh
3. glasno, kaotično dogajanje ali skupina ljudi, ki tako dogajanje povzroča

1. *a place for keeping and raising hens*
2. *a building or space where many people live or stay in close proximity, or, more generally, a disorderly and noisy place*
3. *a loud and chaotic situation or a group of people that causes such a situation.*

betonirati ('to concrete')

1. delati, utrjevati kaj z vlivanjem, nanosom, vgradnjo betona
2. delati, da se česa ne da spreminjati

1. *to construct, reinforce, or stabilize something by pouring, applying, or incorporating concrete*
2. *to make something fixed and resistant to change*

Four state-of-the-art models were selected for the task, two commercial ones (GPT-5.1, Gemini-3-pro-preview) and two open-source ones (Gemma-3-27b, and the Slovenian model GaMS-27b-Instruct-Nemotron). Each model was tested using three distinct approaches:

- No-sample-definitions (Zero-shot): The model generated a definition based solely on the provided semantic indicator and contextual information, without stylistic guidance.
- eSSKJ-sample-definitions (Few-shot): The prompt included gold-standard definitions from the New Dictionary of Standard Slovene (eSSKJ) to provide a traditional stylistic benchmark.
- DSBS-sample-definitions (Few-shot): The prompt included custom-made definitions specifically designed to match the desired style and structural constraints of the DDDS.

Definitions were grouped according to word class. We then prepared system instructions and a prompt in Slovene and English, as we wanted to see in which language they work better. After several rounds of testing and adjustments we used the following Slovene version (English translation is provided here):

- System instructions: "You are a lexicographer writing definitions for a dictionary of Slovene, targeted at native speakers of Slovene. {Here are some sample good definitions for a selection of nouns/adjectives/verbs/adverbs: file.txt".}
- Prompt: "Write a dictionary definition for the given sense of noun/adjective/verb/adverb {word} based on provided contextual information. There is one sense only, so write only one definition. Do not provide any explanations, and do not repeat contextual information."

Using this method, we obtained 12 generated definitions per sense, 3 per model. The definitions were combined into a joint Excel document for evaluation, along with the information on the headword, sense number, model and type of sample definitions. Before the start of the evaluation, the definitions for each sense were put in a random order, and the information on the model and type of sample definitions was hidden.

Evaluation was conducted by two experienced lexicographers. The evaluation process was double-blind and two-tiered: individual ranking followed by a consensus-based selection of the "best" definition. First, each lexicographer selected three best definitions for each sense and ranked them according to their quality. The lexicographers were allowed to consult the contextual data provided to the model. The second step involved the discussion of the two lexicographers to reach an agreement on the best definition for each sense; of course, this was not needed in cases when both lexicographers picked the same definition as the best one. In some cases, two definitions were selected, e.g. when identical or nearly identical wording was used.

3.2 Results

Overall findings on the quality of LLM-generated definitions were highly encouraging. The evaluators noted a complete absence of "hallucinations"—a common concern with LLMs—and found that the models stayed strictly within the intended semantic boundaries of the sense. Very few instances of "Slovenisms" or non-target language interference were recorded. There were no cases of completely wrong definitions – one sense of "tihožitje" was discarded from further analysis as the definitions referred to the other sense of the word. Also, in many cases it was very difficult to pick top three definitions due to their similarity and quality.

Nonetheless, there were still some recurring issues identified, the main ones being:

- Syntactic Redundancy: A tendency toward "wordiness" and overly circular definitions.
- Structural Complexity: Definitions that were grammatically correct but too dense for a general-purpose dictionary.
- Stylistic Anomalies: The occasional use of rare vocabulary or the generation of non-typical collocations within the definition text itself.

Other problematic elements that were identified were synonyms (usually provided separately at the end), Latin names, explanations in brackets, excessive listing of items, and unnecessary extra information on the meaning. Also penalised was the omission of essential semantic information, and excessive length or syntactic complexity of the definition.

The agreement between the lexicographers in terms of selecting the best three definitions was relatively low; they selected the same definitions in the same order at only three senses out of 125, and the same definitions but in a different order at four senses. There was often an overlap between lexicographers at least one definition, and at 11 senses the lexicographers picked completely different definitions.

The agreement between the lexicographers about the best definition of each sense was much higher, with the lexicographers ranking the same definition as the best one in one third of the cases (41 out of 125). It should be pointed out that the total number of best definitions differs from the number of senses (114) as the lexicographers sometimes selected two definitions as the best ones. For 70 senses the jointly agreed best definition was initially picked as the best by at least one lexicographer (36 times by lexicographer 1 and 34 times by lexicographer 2). In 13 cases, the jointly agreed best definition was initially selected by at least one of the lexicographers as second or third best. In one case, the best definition was not among the initial selected top three candidates by any lexicographer (but was not the only best definition of the sense).

The data reveals that the Gemini3-pro-preview model was the most consistent performer (see Table 1), particularly when utilizing the DSBS-sample-definitions (51 or 41% of best definitions). This model-sample-definition combination was exceptionally successful for verbs, where it produced the preferred definition in nearly 70% of all senses. The second most successful combination was GPT-5.1+DSBS-sample-definitions (19 or 15% of best definitions), and the third was Gemini-3-pro-preview+no-sample-definitions (17 or 14% of best definitions). While Gemini-3-pro-preview showed dominance for nouns and adjectives (maintaining a 2:1 success ratio over GPT-5.1), the gap narrowed for adverbs, where GPT-5.1 and Gemini-3-pro-preview performed with similar efficacy.

TABLE 1 Numbers of best definitions selected by lexicographers (by model and by sample definitions)

	GaMS-27B			Gemini-3-pro-preview			Gemma-3-27B			GPT-5.1		
	DSBS	eSSKJ	no	DSBS	eSSKJ	no	DSBS	eSSKJ	no	DSBS	eSSKJ	no
nouns	1	0	0	11	1	9	1	0	0	4	2	3
verbs	2	0	1	15	3	2	1	0	0	4	1	0
adjectives	2	0	2	16	2	4	3	0	0	6	3	2
adverbs	0	0	1	9	2	2	0	0	0	5	2	3
TOTAL	**5**	**0**	**4**	**51**	**8**	**17**	**5**	**0**	**0**	**19**	**8**	**8**

The analysis of the top three definitions selected by the lexicographers (516 in total) shows a similar pattern than the results for the best definitions, with Gemini-3-pro-preview with all three sample definitions inputs leading the way (218 or 42%), followed by GPT-5.1 with all three inputs (146 or 28%). Of the remaining two models, Gams-27b-instruct-Nemotron (86 or 17%) performed slightly better than Gemma-3-27b (66 or 13%). The results per word-class are fairly consistent with the overall results, with one noticeable exception: for verbs, the Gams-27b+DSBS-sample-definitions is the third best performer (after Gemini-3-pro with DSBS and eSSKJ sample definitions).

3.3 *LLM as a judge*

The second part of the experiment was focussed on testing LLMs in evaluating the quality of definitions. We selected three advanced models, Gemini-2.5-pro, GPT-5.4 and Claude-Sonnet-4.5, and asked them to evaluate all the generated definitions using the provided lexicographic criteria. There were three criteria – content, structure and language – each marked from 1 to 5 (5 was the highest score). Detailed explanations of scores, which were included in the prompt (English worked better than Slovene in this case), are provided in Appendix 1. To ensure replicability and consistency as much as possible, the temperature was set to 0. In addition to lexicographic criteria, each model was provided with the same input data as in the definition generation task (indicators, collocations, examples etc.) and 16 sample DSBS or eSSKJ definitions from the original datasets (4 per word class). Also, we added the following definition guidelines:

> A classic dictionary definition contains the direct hypernym of the word being defined (N) and a certain number of distinguishing semantic features (RPL) that differentiate the hypernym, e.g., epika – ‘ literarno ustvarjanje (N), katerega izrazna oblika je zlasti roman, povest ali novela’ (RPL). This type of definition is useful for explaining concrete and

abstract nouns, nouns that lend themselves to systematic classification (animals, plants, etc.), and certain groups of verbs, e.g., verbs of movement.

A whole-sentence dictionary definition consists of a left and a right part of the explanation. The left part contains the basic description of the word and contextual elements, often including the basic structure of the word's typical usage in context (L), while the right part of the explanation contains a comment or clarification (D), e.g., mraz – 'če je zunaj ali v prostoru mraz, je temperatura zraka nizka'. This type of definition is useful for explaining pragmatic and metaphorical meanings, for most verb and adverbial meanings, and for meanings with typical collocation elements from the context, e.g., popoprati – 'če človek popopra dogodek, izjavo, ali vzdušje, ga s pripombo ali z dejanjem naredi bolj zanimivega in pikrega,' and for meanings typically used in specific syntactic constructions, e.g., with the verb 'biti': avtonomen – 'če je ustanova ali oseba avtonomna, je samostojna in deluje ter sprejema odločitve neodvisno in brez pomoči drugih'.

The analysis of combined model scores shows that on average definitions of noun and verb senses were evaluated the highest (38.24 and 37.93 of possible 45, respectively), followed by definitions of adverb and adjective senses (36.57 and 35.94 respectively). Of the three models, Claude-Sonnet-4.5 tended to award slightly lower scores on average (11.76) compared to the other two models (Gemini-2.5-pro 12.98 and GPT-5.4 12.56). The differences become more prominent if we look at the number of definitions awarded the highest score of 15: Gemini-2.5-pro awarded it to 450 definitions out of 1381, GPT-5.4 to 275, and Claude-Sonnet-4.5 to 120.

We then compared the model scores with the results of human evaluation. First, we compared the combined scores for the best definition selected by the two lexicographers with the combined scores for the best definition(s) selected by the models. As Table 2 shows, the best lexicographically selected definition received the highest score in 41 of 124 cases; the percentages for noun, verb and adverb definitions were around 40%, and much lower for adjectives (20%).

TABLE 2 Number of definitions per point difference (combined model score) between the score of the best definition by model and the score of the lexicographically selected best definitions

point difference	*nouns*	*verbs*	*adjectives*	*adverbs*
0	13	11	8	9
1	3	7	3	8
2	8	6	7	2
3	3	0	5	4
4	1	2	9	0
5 and more	4	3	8	0

Using the same comparison, we also analysed the scores of each model (Table 3). Gemini-2.5-pro showed the highest agreement with lexicographic selection, especially for adjectival senses. GPT-5.4 came second in all word classes, for adverbs it performed even slightly better than Gemini. Claude-Sonnet performed the worst of all three models, especially when considering only 0 point differences.

TABLE 3 Number of definitions per point difference (individual models) between the score of the best definition by model and the score of the lexicographically selected best definitions

point difference	nouns			verbs			adjectives			adverbs		
	GPT	**CS**	**Gem**	**GPT**	**CS**	**Gem**	**GPT**	**CS**	**Gem**	**GPT**	**CS**	**Gem**
0	18	15	22	16	13	20	14	6	22	17	8	17
1	10	7	1	7	9	3	15	10	6	3	8	3
2	1	9	5	6	4	4	5	11	6	3	4	1
3	2	1	3	0	3	0	4	8	4	0	3	2
4	1	0	1	0	0	1	1	2	1	0	0	0
5 and more	0	0	0	0	0	1	3	1	0	0	0	0

One of the aspects that was also examined was potential model bias. The explanation for the best performance of Gemini in this task could be that it may have scored its own definitions more favourably – this holds true as 52 of 71 definitions by Gemini were awarded all possible points. And given that a high percentage of lexicographically selected best definitions were produced by Gemini this might have led to better results. Nonetheless, one could argue that Gemini is generous with scoring in general, also to the definitions of the other models. For example, Gemini awarded all possible points to 21 definitions generated by GPT, which was even more than by GPT itself (17 definitions).

4 Conclusions and future plans

The experiment of definition generation has shown that LLMs can be very good at this task when provided with suitable context and sample definitions. Gemini-3-pro-preview outperformed other models in terms of providing the best definition for each sense, especially two open-source ones. However, it is important to note that all LLMs included in the study in the majority of cases produced fairly good definition candidates without hallucinations or incorrect information. This finding is particularly encouraging for the open-source Slovenian model GaMS, which performed especially well for verbal senses.

The findings of the LLM-as-a-judge task are equally encouraging, with LLMs, especially Gemini, showing high agreement with human evaluators. In addition, even when LLMs picked as the best different definitions from the human evaluators, the selected definitions were still very good and often among top three candidates of at least one human evaluator.

The fact that Gemini consistently outperforms other models is consistent with our findings that Gemini models are better tailored for Slovene language, observed across various language experiments run by our research group.

Based on these results, we have decided to combine definition generation and LLM evaluation as an automatic step in the entry generation workflow. We are currently using Gemini and GPT for definition generation, and Gemini for LLM-as-a-judge; adding lexicographic version of GaMS is planned in the near future. We have used this approach on a dataset of approximately 22,500 headwords with automatic collocations and examples, for which first the senses were generated by LLMs, and on all of the current manually completed entries (approx. 22,000) in the DSBS. Initial evaluation has shown very positive outcomes, and the improvements of shortcomings and further upgrades (i.e. identifying irrelevant/false senses) are already being tested.

Future research will expand definition generation testing to include other word classes, such as function words (e.g., prepositions and conjunctions), as well as multi-word lexical units like compounds and idioms. Testing these categories is essential, as function words lack concrete semantic referents, making their definitions highly abstract, while multi-word units require the model to grasp non-compositional meanings that cannot be derived from the individual words alone. Additionally, we will conduct a more rigorous, nuanced evaluation of definition generation for sensitive semantic categories, particularly derogatory and vulgar meanings.

References

Atkins, B. T. Sue & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press.

Bartosz, P., Wolfer, S. & Lew, R. (2024). A Learners' Dictionary versus ChatGPT in Receptive and Productive Lexical Tasks. *International Journal of Lexicography*, (37) 3, 322–336. https://doi.org/10.1093/ijl/ecae011.

de Schryver, G-M. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography*, (36) 4, 355–387.

Gantar, P. (2024). Formulisanje rečničkih definicija pomoću veštačke inteligencije na primeru slovenačkih frazeoloških jedinica. In S. Marjanović (ed.), *Leksikografski susreti 1: moderni rečnici u funkciji prosečnoga korisnika: stari problemi, savremeni pravci i novi izazovi: tematski zbornik radova s međunarodnoga naučnog skupa održanog od 27. do 29. maja 2024. na Filološkom fakultetu Univerziteta u Beogradu*, Beograd, Srbija (pp. 151-157).

Gantar, P. (2020). Dictionary of Modern Slovene: From Slovene Lexical Database to Digital Dictionary Database. *Rasprave Instituta Za Hrvatski Jezik i Jezikoslovlje*, 46(2), 589–602. https://doi.org/10.31724/rihjj.46.2.7.

Gantar, P. (2015). Leksikografski opis slovenščine v digitalnem okolju. Znanstvena založba Filozofske fakultete UL.

Ide, Y., Nohejl, A., Tanner, J., Yanaka, H., Lindsay, Ch. & Watanabe, T. (2026). Towards Automated Lexicography: Generating and Evaluating Definitions for Learner's Dictionaries. Computation and Language. Accessed May 5th 2026 at https://arxiv.org/abs/2601.01842v1

Jakubíček, M. & Rundell, M. (2023). The End of Lexicography? Can ChatGPT Outperform Current Tools for Post-Editing Lexicography? In M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček & S. Krek (Eds) *Electronic Lexicography in the 21st Century (eLex 2023): Proceedings of the eLex 2023 Conference*. Brno, Czech Republic (pp. 518–533). Lexical Computing CZ. https://elex.link/elex2023/wp-content/uploads/102.pdf.

Kosem, I., Gantar, P., Arhar Holdt, Š., Gapsa, M., Zgaga, K., Krek, S. (2024). AI in lexicography at the University of Ljubljana: case studies. In *Large Language Models and Lexicography, Book of Abstracts*, 8th October 2024, Cavtat, Croatia (pp. 29-32).

Kosem, I., Krek, S. & Gantar, P. (2021). Semantic data should no longer exist in isolation: the Digital Dictionary Database of Slovenian. In Z. Gavriilidou, L. Mitits & S. Kiosses (Eds.), *EURALEX XIX, Congress of the European Association for Lexicography, Lexicography for inclusion: 7–9 September 2021*, virtual: abstracts book (pp. 81–83). Democritus University of Thrace.

Lew, R. (2023). ChatGPT as a COBUILD Lexicographer. *Humanities and Social Sciences Communications*, 10, Article 704. https://doi.org/10.1057/s41599-023-02119-6.

McKean, E. & Fitzgerald, W. (2024) The ROI of AI in Lexicography. *Lexicography*, (11) 1, 7–27. https://utppublishing.com/doi/abs/10.1558/lexi.27569.

Rees, G. P. & Lew, R. (2024). 'The Effectiveness of OpenAI GPT-Generated Definitions Versus Definitions from an English Learners' Dictionary in a Lexically Orientated Reading Task.' *International Journal of Lexicography*, 37(1), 50-74.

Rundell, M. 2023. Automating the Creation of Dictionaries: Are We Nearly There? In *Proceedings of the 16th International Conference of the Asian Association for Lexicography: "Lexicography, Artificial Intelligence, and Dictionary Users"*, 22–24 June 2023, Seoul, South Korea (pp. 9–17). Yonsei University.

Tiberius, C., Heylen, K. de Does, J., Vanroy, B., Vandeghinste V. & van Doeselaar, J. (2024). LLMs and Evidence-based Lexicography. In *Large Language Models and Lexicography, Book of Abstracts*, 8th October 2024, Cavtat, Croatia (pp. 44–48).

Contact Information

Iztok Kosem
University of Ljubljana, Faculty of Arts & Faculty of Computing and Information Science/Jožef Stefan Institute
iztok.kosem@fri.uni-lj.si

Polona Gantar
University of Ljubljana, Faculty of Arts & Faculty of Computing and Information Science
apolonija.gantar@ff.uni-lj.si

Simon Krek
University of Ljubljana, Faculty of Arts & Faculty of Computing and Information Science/Jožef Stefan Institute
simon.krek@fri.uni-lj.si

Tjaša Arčon
University of Ljubljana, Faculty of Computing and Information Science
tjasa.arcon@fri.uni-lj.si

Acknowledgements

The research programs Language Resources and Technologies for Slovene (P6-0411) and Slovene language – basic, contrastive and applied research (P6-0215), the infrastructure Network of Research and Infrastructural Centers UL (I0-0022) and the project Large Language Models for Digital Humanities (LLM4DH; GC-0002) are funded by the Slovenian Research and Innovation Agency. We thank the reviewers for their constructive and valuable suggestions.

Appendix: grading system for the LLM-as-a-judge task

grade	*content*	*structure*	*language*
1	The definition is accurate for the given meaning, describes its essential characteristics, and does not contain any unnecessary additional information.	The structure is simple, containing at most one subordinate clause and one enumeration. The definition does not contain unnecessary additional elements, such as words in parentheses or additional synonyms separated by commas. The definition is neither too short—e.g., just a few words—nor too long—e.g., more than 30 words.	The definition is written in correct standard Slovenian; there are no grammatical errors, and punctuation is used appropriately. The text does not contain very rare or non-existent words, and typical word combinations are used. The definition does not contain metaphors or informal language. The definition does not contain vulgar or offensive language.

grade	*content*	*structure*	*language*
2	The definition is accurate for the given meaning, describes its essential characteristics, and contains some unnecessary additional information.	The structure is simple, containing at most one subordinate clause and a few enumerations. The definition does not contain unnecessary additional elements, such as words in parentheses or additional synonyms separated by commas. The definition is neither too short—e.g., just a few words—nor too long—e.g., more than 30 words.	The definition is written in correct standard Slovenian; there are no grammatical errors, and punctuation is used correctly. The text may contain one or two uncommon words or less typical phrases. The definition does not contain metaphors or informal language. The definition does not contain any vulgar or offensive language.
3	The definition is partially accurate for the given meaning; some essential semantic features are missing, and it also contains some unnecessary additional information	The structure is somewhat complex, containing several subordinate clauses and/or enumerations. It may also include redundant elements, such as words in parentheses or additional synonyms separated by commas. The definition is not too short—for example, just a few words—but may be somewhat longer, for example, more than 30 words.	The definition is written in standard Slovenian, though there may be the occasional linguistic error or misplaced punctuation. There may also be the occasional rare or non-existent word, or a less typical phrase. A metaphor or informal expression may be used. The definition does not contain vulgar or offensive language.
4	The definition is inadequate for the given meaning; it captures some of the essential semantic features, but may contain several unnecessary additional information.	The structure is complex, containing a series of subordinate clauses and/or excessive enumeration. It may also include redundant elements, such as words in parentheses or additional synonyms separated by commas. The definition may be too short—for example, just a few words—or somewhat longer—for example, more than 30 words.	The definition contains many linguistic errors, and several rare or non-existent words. There can be also many atypical phrases, metaphors, or informal expressions. The definition does not contain vulgar or offensive language.
5	The definition is incorrect for the given meaning; it does not capture the essential semantic features.	The structure is overly complex, containing a series of subordinate clauses and excessive enumerations. There are also redundant elements, such as words in parentheses or synonyms separated by commas. The definition is either too short—for example, just a few words—or too long—for example, more than 40 words.	The language of the definition is riddled with errors, inappropriate phrasing, non-existent words or phrases, metaphors, and informal expressions. An offensive or vulgar word is used.

František Kovařík, Marek Blahuš, Miloš Jakubíček & Vojtěch Kovář

AGREE TO DISAGRE

Intra-Annotator Agreement in Semi-Automatic Vocabulary Building

Abstract Intra-annotator agreement is the rate at which an annotator makes the same choices in the same situation. This paper examines the annotation process in a semi-automatic dictionary-making project, Czech Dictionary Express. In the vocabulary-building process, its annotators went through 100,000 headwords from a corpus frequency wordlist, marking each as incorrect or (partially) correct. The aim of the Dictionary Express projects is a high dictionary-making speed, so the annotators weren't given the context of the headwords.

In this paper, we examine the shift in annotators' opinions on headwords before and after context was provided. We argue against providing no context of the headwords, because, unlike the (partially) accepted words, which can still be cut from the vocabulary at later stages, the rejected words cannot be put back in, and this is the case of some correct words that have not been recognised by the annotators without context. We show that the context-based annotation process doesn't take significantly longer. We also take a brief look at a study about using LLMs for the same annotation process.

Keywords Inter-annotator agreement; intra-annotator agreement; dictionary drafting; post-editing lexicography; corpus annotation; semi-automatic dictionary making; dictionary express o Czech

1 Introduction

The process of creating any dictionary at some point incorporates vocabulary selection – a stage of thousands of macro- and micro-level choices about the composition of the dictionary wordlist and the specific words that get to be in the dictionary and receive the often resource-expensive lexicographic treatment. The semi-automatic approach to this problem, described in this paper using the Dictionary Express methodology (formerly "rapid dictionaries from scratch" (Blahuš et al., 2023)), consists of automatically assembling frequency wordlists and subsequently annotating them by native speakers. This annotation is as simple as possible, requiring only an answer on whether the word is suitable for recording in a lexicographic resource; the rest of the dictionary entry for the accepted words is created in later tasks (more in Section 2).

One problem with manual annotation is whether a human annotator's introspection can be trusted for long-term consistency – especially since projects like Dictionary Express use native speakers who are not fully trained linguists (although problems with word acceptance can arise as frequently inside as outside the linguistic community). The rate by which an annotator agrees with their own past decisions, with reference to the inter-annotator agreement (IAA; "agreement among annotators"), is sometimes called the **intra-annotator agreement** (IaAA).

In this paper, we examine data from two separate annotation processes in the Czech Dictionary Express project (see Section 2), carried out by a group of annotators twice, with a one-year gap between the two experiments. We ask whether and how certain techniques and interface changes can influence annotators' micro-choices about accepting a word in their native language, and whether these techniques and changes could improve or harm the selection quality of the

final vocabulary (see Section 3). In the final part, we apply statistical measures to the intra-annotator data and mention a similar experiment conducted on LLMs (see Section 4).

2 Semi-automatic Vocabulary Selection

Dictionary Express is a group of projects that aims to create general-purpose, descriptive general-language dictionaries for different (often small) languages using a semi-automatic approach. DE dictionaries often include forms of both proper names (place names, people names, abbreviations, organisation names, etc.) and forms, senses, thesaurus synonyms and English translations of common words, and more. The methodology uses data gathered from large web corpora, manually annotated by a group of native speakers. The projects are divided into smaller, simple tasks, each creating chunks of dictionary entries for the whole vocabulary, with the first task – the headword selection – centred around creating the vocabulary itself (Kovařík, 2023).

The annotation team usually consists of 10 or fewer annotators, most of whom are not university-level linguists. This is often due to the lack of expert linguists or resources in small-language native-speaker communities. Although Czech lacks neither, the Czech Dictionary Express (CDE) project simulates this environment, allowing researchers to experiment with the tools (Kovařík et al., 2024).

At the beginning of each project, a list of general-language words (single words only; no multiwords are included at this stage) is gathered and alphabetised (so similar words are close to each other) for different frequency ranges. For CDE, the wordlist was gathered from the Czech Web 2012+17+19 corpus, which contains over 11,7 billion words (*CsTenTen – Czech Corpus from the Web*, 2024).

After the initial annotation workshops, the DE annotators are given the task of *headword selection* – they go through batches of the general-language wordlist and assign each *headword* (usually lemma-POS pairs) one of seven flags: *I don't know the word*, *not Czech* (not in the language of the dictionary), *non-standard*, *not a lemma* (also for wrong lemmas, typos etc.), *wrong POS*, and finally *proper name* and *ok* (for correct common words).

Until recently, this has been done without providing the typical context of the headwords. The annotators would only see the wordlist and a simple diagram of the flag priority (e.g. if a word can be marked *non-standard* and *not a lemma*, it should be marked *non-standard* since this flag has a higher priority). The context from corpus concordances would be provided at a later stage, in the revision task, in which another group of annotators would go through a smaller list of headwords that have been marked with **different** flags, or flags **of a specific problem** (*non-standard*, *not a lemma*, *wrong POS*), but **not the "rejection" flags** (*I don't know the word*, *not Czech*).

The project coordinators assumed that combining a context-free annotation with a context-based revision would lead to a quicker vocabulary-building process (we argue against this in Subsection 3.4). It was not considered desirable to provide the initial annotators with context, since each word has to be marked by at least two different annotators, whereas revisions are done on a much smaller part of the candidate vocabulary by a single annotator. There have been concerns, however, about the desired objectivity of the vocabulary, and especially about the proper Czech words unanimously rejected by the annotators working under time pressure and without context because they simply didn't recognise a lemma-POS form of a commonly used word. Where the human factor comes into play, errors are always to be expected.

Headwords rejected in the annotation phase were not taken to the revision nor the vocabulary, thus "lost forever" to the project. The problem of incorrectly rejected headwords is a type II (false negative) error. It leads to correct words omitted in the dictionary, and presents a much bigger problem in the semi-automatic methodology than the opposing type I (false positive) error: Where accepted incorrect words can be discarded in the later stages of the projects, the rejected correct words can never be put back into play (except on rare occasions in which the lexicographers/coordinators notice the missing word themselves).

To assess this problem, another interface feature has been added to the headword annotation task recently, and serves as a part of our experiment: If an annotator would assign the headword a different flag than *ok* (common word) or *proper name*, they would get a message with an additional question on whether they really want to do that, accompanied by the common context of the word in given language (from concordances of the word in Czech Web corpus), as seen in Figure 1. In this paper, we refer to this feature simply as ***The Message***.

Applying this approach may yield more vocabulary headwords, but it may also slow down the annotation process. In the next section, we discuss whether the headwords not omitted thanks to *The Message* are worth the delay, how many there could be, and whether they could serve as proper candidates to the recorded vocabulary.

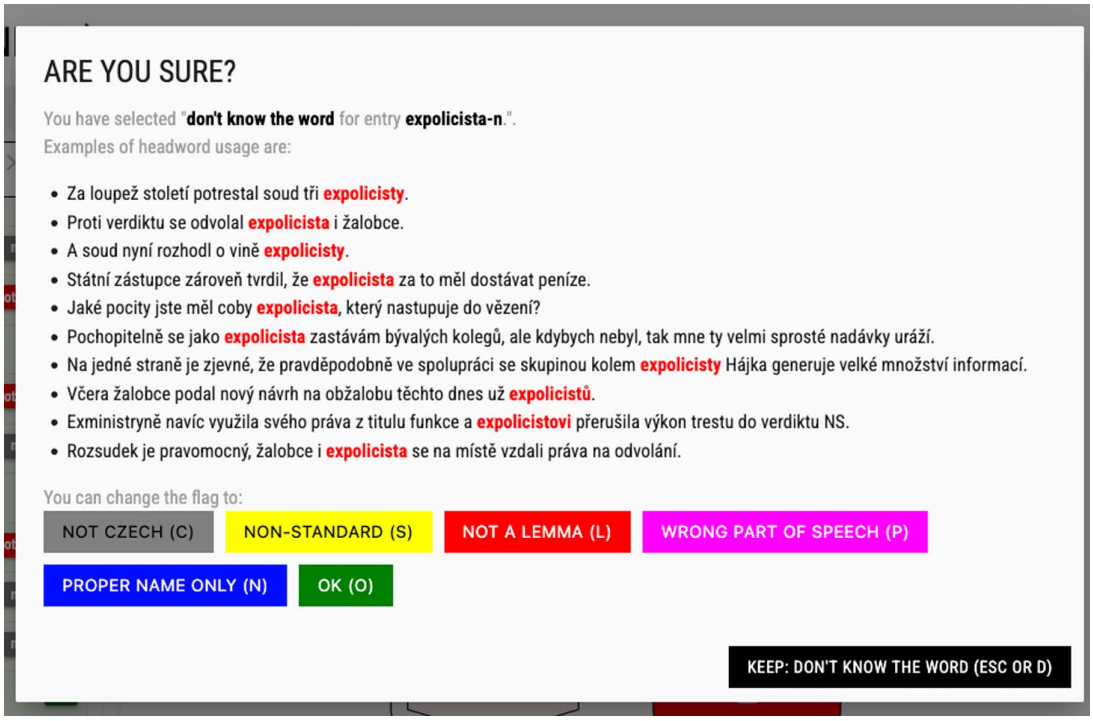


FIGURE 1 The Message appearing after marking the noun "expolicista" as I don't know in the Lexonomy (Měchura, 2017) interface

3 IaAA

In this experiment, three Czech native speakers have each been given 14 batches of 1,000 headwords (14,000 headwords in total) from different document frequency ranks: 3 batches from 15,000 most frequent headwords, 4 batches from the next 35,000 (ranks 15,001–50,000), 3 batches from the next 30,000 (ranks 50,001–80,000), 3 batches from the next 20,000 (ranks 80,000–100,000), and finally 1 batch from the very low frequency ranks

(ranks 100,001–150,000). They annotated each batch twice, with a one-year gap between annotations. The first annotation didn't include *The Message* with context (Figure 1). In the second annotation, two of the three annotators (annotator 1 and 2) have been warned by *The Message* when selecting a flag other than *ok* or *proper name*; the third annotator (annotator 3) has never been given *The Message* except for the last batch. (Since contributing to a separate experiment, the last batch of the ranks 100,001–150,000 received a special treatment, and has been annotated twice with a gap of several weeks, first time without and second time with *The Message*.)

3.1 Quantitative analysis: First vs second annotation

In this subsection, we present the data gathered in the experiment, with the focus on the difference between the numbers of "rejection" flags (*I don't know* and *not Czech*), which do not lead to putting the word into dictionary and can possibly disqualify the word from the subsequent revision process with context, and (partial) "acceptance" flags (*non-standard*, *not a lemma*, *wrong POS*, *proper name* and *ok*).

Table 1 presents the intra-annotator statistics of each annotator. *Words annotated* is the number of words annotated in the experiment (the third annotator saw slightly fewer words because of a software error). *Acc (1st annotation)* and *acc (2nd annotation)* are the amounts of words accepted; same goes with *rej (1st annotation)* and *rej (2nd annotation)*, which have been rejected the first or second time.

The columns correspond to the frequency ranks of the words.

It is important to understand that not every annotator works the same way. First off, we see that the words in the most frequent ranks get accepted more often in the first annotation, and that annotators 1 and 3 tend to accept fewer words than annotator 2 (all the annotators have been given the same words). We can also see that the annotator 3, who has not been given any headword context for the first or second time (except the last batch), tended to accept words more often in the second run, and that this shift is pretty close to that of annotator 2, who has been given *The Message* in the second run.

Annotator 1, however, given the context of the initially rejected headwords (*2nd annotation*), has accepted almost every headword candidate. By giving the context, the *ok*- or *proper name*-flag rate has risen by almost 500 out of 3,000 in the third rank, and by more than 900 out of 3,000 in the fourth rank. The rate of the other (partially) accepted flags has also increased. If most of these headwords are proper words suitable for dictionary recording, this would be a great help for the semi-automatic drafting. We look into some of them in the next subsection.

The acceptance rate has increased for all annotators in the second run, i.e., for those given *The Message* as well as for those who were not. We should thus remember that *The Message* is not the only factor at play, and that the annotators have a year of annotation experience between the two experiments. However, we argue that it can be a powerful tool for the annotators in the next subsection, because it is not always simple to recognise the word or lemmatisation/tagging error by looking at a lemma-POS pair without any context.

3.2 More on the acceptance shift

A word that presented a problem for the annotators in the early stages of the project was the adjective *rybařený*. Although the word is known for some Czech speakers as a rare slangism

TABLE 1 Word counts of accepted and rejected words in the first and second annotation

frequency rank	1-15,000	15,000-50,000	50,000-80,000	80,000-100000	100,000-150,000
annotator 1					
words annotated	3,000	4,000	3,000	3,000	1,000
acc (1st annotation)	2,907	3,478	2,251	1,587	803
acc (2nd annotation)	2,982	3,984	2,966	2,945	977
rej (1st annotation)	93	522	749	1,413	197
rej (2nd annotation)	18	16	34	55	23
annotator 2					
words annotated	3,000	4,000	3,000	3,000	1,000
acc (1st annotation)	2,908	3,733	2,541	2,349	591
acc (2nd annotation)	2,947	3,796	2,668	2,430	759
rej (1st annotation)	92	267	459	651	409
rej (2nd annotation)	53	204	332	570	241
annotator 3					
words annotated	2,999	3,998	2,994	2,998	1,000
acc (1st annotation)	2,918	3,382	2,226	1,820	479
acc (2nd annotation)	2,927	3,588	2,401	1,978	616
rej (1st annotation)	81	616	768	1,178	521
rej (2nd annotation)	72	410	593	1,020	384

(in some forms of the climber slang, e.g. *rybařená hora* as "fished mountain" – a mountain that someone is trying to climb), it should be regarded as more or less an occasionalism, an infrequent word with unclear sense. And yet, it appeared amongst the most frequently used headwords.

While looking at the context, it was clear that it was a mistake made by the lemmatiser. he correct lemma should be the noun *rybaření* (*fishing*), which indeed is a common and frequently used Czech word. Given the word without its context, the annotators rejected the word, marking it *I don't know*.

The adjective *rybařený* belongs to the *rej→acc* words, i.e. words that have been rejected in the first run (and lost forever), but (partially) accepted in the second run (have the potential to be saved by their context). While looking into the words that the annotators 1 and 2 rejected in the first run, but accepted after looking at the context, we mostly see:

- abbreviations and foreign company names, e.g. *GSM*, *ICT*, or even *International* (supposedly as a part of a longer company/organisation name);
- incorrectly lemmatised words (e.g. the nouns *plavka*, or city name teplice without a capital letter) – these can still occur in the dictionary if their "correct" counterpart (*plavky* – "swimsuit", *Teplice*) is frequent enough, which, however, is not the case for many (e.g. mentioned noun *rybaření*);
- words hard to recognise without context, such as *transparent* (noun) – "banner", *pozměňovací* (adjective) – "amendatory", *disperze* (noun) and *disperzní* (adjective) – "dispersion" and "dispersive", *disponibilní* (adjective) – "disposable", *dohrávka* (noun) – "postponed

match", *vleknout* (verb) – "to drag (on)", *vylhaný* (adjective) – "lied out", *brejkový* (adjective; connected to break dance), *brzdovka* (noun; sometimes instead of *brzdová kapalina* – "break fluid"), *buchar* (noun) – "power hammer", *bující* (adjective) – "proliferating", *leitmotiv* (noun), *pučista* (noun) – "putschist", *psychedelie* and *psyché* (nouns) – "psychedelia" and "psyche", *pudit* (verb) – "to force out", etc.

TABLE 2 Intra-annotator agreement (how many times did the annotator select **the exact same flag** as before) and Cohen's kappa (weighted on "rejected", "partially accepted" and "accepted").

frequency rank	1-15,000	15,000-50,000	50,000-80,000	80,000-100000	100,000-150,000	All
annotator 1						
agreement	2808	3260	2085	1422	679	10254
disagreement	192	740	915	1578	321	3746
Cohen's Kappa	0.572	0.37	0.324	0.177	0.479	0.349
annotator 2						
agreement	2774	3381	2348	2132	662	11297
disagreement	226	619	652	868	338	2703
Cohen's Kappa	0.75	0.683	0.708	0.656	0.656	0.715
annotator 3						
agreement	2827	3195	2330	2150	696	11198
disagreement	172	803	664	848	304	2791
Cohen's Kappa	0.758	0.635	0.737	0.708	0.669	0.742

Note that the words from the last group are not only terms used in a specific specialist text types (*disperze* or *pozměňovací*), but also common words (*vylhaný*) or their variants (*vleknout* as a less used variant of *vléci*).

3.3 Agreement and Cohen's kappa

For evaluating the data, we cannot use the final vocabulary as reference "correct words", since there might have been words omitted because of a false negative error (e.g. the noun *rybaření* "disguised" as adjective *rybařený*). We use Cohen's Kappa to evaluate the intra-annotator agreement.

Table 2 presents the kappa and agreement statistics of the experiment for each frequency rank, and for all words together. For each annotator, the first line presents the words for which the annotators selected **the exact same flag**, the second the words where the two flags were different. For the kappa statistics, we are using Cohen's Kappa quadratically weighted for 3 values:

1. Rejection flags: *I don't know* and *not Czech*.
2. Partial acceptance flags: *non-standard*, *not a lemma* and *wrong word*.
3. Acceptance flags: *ok* and *proper name*.

In the overall score, annotators 2 and 3 get a kappa of over 0,7 (substantial agreement (Viera & Garrett, 2005)), whereas annotator 1 gets a kappa of 0,349, which isn't as high as we could expect (possibly because of *The Message*).

We can also see that, even though agreement on the exact same flag decreases with frequency, it is not the case with the weighted kappa between "rejected", "partially accepted", and "accepted".

3.4 The Demands of Context

One of the advantages of semi-automatic dictionary making is its speed. The Dictionary Express projects were previously called "rapid dictionary-making projects". One of the concerns of providing the headword annotators with context was that it could slow down the vocabulary-building process.

Thanks to good language resources for Czech (corpus, lemmatiser and POS-tagger), the usual headword annotation is faster compared to other "medium-sized" languages. A batch of 1,000 headwords takes about 1,5–2 hours on average for one annotator to annotate, without looking into the context.

Our results show that annotators 1 and 2, given the headwords without context, finished a batch in approximately 1,63 and 1,80 hours. While working with *The Message* (i.e. given word context if they would reject a word), the first annotator finished a batch in approximately 2,58 hours on average, whereas the second one even faster than before, in 1,52 hours.

While keeping in mind that both annotators had a year's worth of experience more at the second run and could have gotten faster at their work, the work didn't slow down significantly while providing context.

On the other hand, adding context to the annotations might slow down the revision process, because it yields more words with the partial acceptance flags. However, based on semi-automatic dictionary coordinators' experience, the revision task takes up only a small amount of time compared to other tasks (headword annotation, sense disambiguation, etc.).

4 Large Language Models as annotators

In her bachelor's thesis, Černá [1] (2025) has tested whether ChatGPT-4o or other large language models could substitute or supplement the work of the annotators. Part of her experiment included intra-model testing. ChatGPT-4o has been given words from the 100,000 document frequency wordlist 3–6 times. Consensus on a final flag was reached if a flag was assigned two more times than any other (i.e., in 3:1, 3:1:1, 4:2, 4:1:1, or 3:1:1:1 ratio). If no flag had the most "votes" in the 6th run, no flag was assigned.

Table 3 was taken from Černá's thesis. The second column is the number of occurrences, i.e. the number of annotated headwords. The fourth column shows the percentage of these words in the experiment.

Most of the words (77.40 %) reached a consensus on a flag in the first three runs or four runs. On the same dataset (100,000 most frequent headwords), annotator 3, who didn't receive *The Message*, had an agreement of 80.85 %; the other annotators had agreements of 81.81 and 73.65 %.

[1] While not being a direct part of the DE team, Černá was collaborating with the project and has made several important contributions, particularly about the use of LLMs.

Although we don't expect human annotators not to reach consensus after the 5 runs as LLMs did, the "intra-annotator agreement" of GPT is apparently close to the IaAA of native speakers. Keep in mind, however, that this experiment was done on a fairly resourced language – Czech – and the consensus of the large language models might be less accurate for smaller languages.

TABLE 3 Consensus in different runs of ChatGPT-4o annotation. (Taken from Černá, 2025.)

	Number of occurrences	Percentage
3 runs (3:0)	377	75.40 %
4 runs (3:1)	10	2.00 %
5 runs (3:1:1)	26	5.20 %
6 runs (4:2, 3:1:1:1)	9	1.80 %
6 runs, weak consensus (3:2:1, 2:1:1:1:1)	15	3.00 %

5 Conclusion

This paper has presented the statistics of agreement in the Czech wordlist annotation.

The vocabulary creation for Dictionary Express resources does not end with a single annotation of each headword by a single annotator, without context. Instead, it is backed up by several opinions from different annotators and the work of an "inspector", another, more skilled annotator who sees the context of use of each word, as well as the previous annotations, and thus has a better objective overview of the headword. This leads to greater objectivity in the language vocabulary selection, done with minimal resources (number of annotators, time, etc.). Human error is expected at some rate in manual annotation.

However, given that only the headwords accepted in the annotation get to see the revision process, we argue that the annotation process should not go without context. Another approach would be to include fully rejected headwords in the revision process, for a "third opinion". We have shown that, given the context, annotators changed their minds about rejecting some words, incorrectly lemmatised, POS-tagged, or just unrecognised in their lexicographic form. We have also shown that the time spent annotating doesn't significantly change when the annotators are given the context. The demands on time in the annotation and revision process would rise slightly, but provide more correct headwords.

References

Blahuš, M., Cukr, M., Herman, O., Jakubíček, M., Kovář, V., Kraus, J., Medved, M., Ohlídalová, V. & Suchomel, V. (2023). Rapid Ukrainian-English Dictionary Creation Using Post-Edited Corpus Data. In C. Tiberius, M. Medved & M. Měchura (Eds.), *Electronic Lexicography in the 21st Century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 Conference* (pp. 613–637). Lexical Computing CZ s.r.o. https://elex.link/elex2023/wp-content/uploads/114.pdf

Černá, S. (2025). *Czech Headword Annotation: Native Speakers Vs Large Language Models* [Bachelor's thesis, Masarykova univerzita, Filozofická fakulta, Brno]. https://is.muni.cz/th/jzl31/

CsTenTen: *Czech Corpus from the Web* (2024). https://www.sketchengine.eu/cstenten-czech-corpus/

Kovařík, F. (2023). Semi-Automatic Dictionary Creation for Czech. In A. Horák, P. Rychlý & A. Rambousek (Eds.), *Recent Advances in Slavonic Natural Language Processing (RASLAN 2023)* (pp. 93–100). NLP Consulting, s.r.o. https://nlp.fi.muni.cz/raslan/raslan23.pdf

Kovařík, F., Kovář, V. & Blahuš, M. (2024). On Rapid Annotation of Czech Headwords: Analysing the First Tasks of Czech Dictionary Express. In K. Š. Despot, A. Ostroški Anić & I. Brač (Eds.), *Lexicography and Semantics: Proceedings of the XXI EURALEX International Congress* (pp. 357–366). Institut za hrvatski jezik. https://euralex.jezik.hr/wp-content/uploads/2021/09/Euralex-XXI-proceedings_1st.pdf

Měchura, M. (2017). Introducing Lexonomy: An Open-Source Dictionary Writings and Publishing System. In M. Jakubíček, I. Kosem & C. Tiberius (Eds.), *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017 Conference* (pp. 662–679). Lexical Computing.

Viera, A. J. & Garrett, J. M. (2005). Understanding Interobserver Agreement: The Kappa Statistic. *Family Medicine*, 37(5), 360–363.

Contact Information

František Kovařík
Lexical Computing/Faculty of Arts, Masaryk University, Brno, Czech Republic
frantisek.kovarik@sketchengine.eu

Marek Blahuš
Lexical Computing
marek.blahus@sketchengine.eu

Miloš Jakubíček
Lexical Computing/Faculty of Informatics, Masaryk University, Brno, Czech Republic
milos.jakubicek@sketchengine.eu

Vojtěch Kovář
Lexical Computing/Faculty of Informatics, Masaryk University, Brno, Czech Republic
vojtech.kovar@sketchengine.eu

Barbara Lewandowska-Tomaszczyk

THE INTEGRATION OF BRITISH AND AMERICAN SLANG IN CONTEMPORARY POLISH AND THE ROLE OF LLM-DRIVEN DYNAMIC LEXICOGRAPHY

Abstract This paper investigates the accelerating influx of English slang into the Polish language, driven by digital globalization and social media usage. While historical Anglicisms often focused on technical or professional domains, modern borrowing increasingly penetrates the affective realm—informal expressions of identity, emotion, and subcultural belonging. Central to this exploratory research is the evaluation of Large Language Model (LLM) assistance as a supplementary tool in dynamic lexicography, where manual editorial processes traditionally struggle to capture short-lived online neologisms. This paper proposes a framework utilizing LLMs for initial candidate extraction and semantic mapping. Using a small-scale pilot survey of Polish speakers, we evaluate LLM outputs, discuss the systemic risks of automated lexicography (such as hallucinated forms and semantic overinterpretation), and outline the vital role of human validation in tracking modern slang.

Keywords American English; borrowings; British English; Dynamic Lexicography; LLMs; Polish; semantic shifts; slang; Pilot Study; limitations

1 Introduction

For over thirty years, the Polish language has experienced an accelerating influx of English loanwords. While historical borrowings were largely confined to specialized, technical, or professional domains (Zabawa, 2017), contemporary linguistic adoption looks fundamentally different. Driven by digital globalization and the ubiquity of social media, modern borrowing has penetrated the everyday, affective realm of communication. Polish youth increasingly rely on English slang as informal expressions of identity, emotion, and subcultural belonging. One of its signs is Poland's *Youth Word of the Year* contest, whose 2025 anniversary edition featured finalists mostly drawn from English slang—such as *67*, *brainrot*, *GOAT*, and *slay*—following the triumphant win of *sigma* a year earlier (https://www.polskieradio.pl/395/7789/Artykul/3593234,poland%E2%80%99s-youth-language-contest-returns-for-its-anniversary-edition).

This rapid, digitally driven evolution poses an acute challenge to traditional lexicography. The conventional editorial process, bound by manual compilation and prolonged publishing cycles, is inherently static. As a result, it struggles to capture the lifecycle of modern slang, which can emerge, peak, and fall into obsolescence within months.

To address this methodological lag, recent attention has turned to computational assistance. Rather than seeking to replace the human lexicographer, this framework positions advanced AI as an assisting tool to empower workflows (cf. Lew, 2011; Lewandowska-Tomaszczyk & Pawlowski, 2025). By utilizing LLMs to perform automated candidate extraction and initial semantic mapping across social media data, computational pipelines might bridge the gap between static dictionaries and living language. However, incorporating LLMs requires scrutinising significant linguistic risks, including false etymologies, hallucinations, and an inability to decode irony.

This paper presents an exploratory pilot study evaluating LLM-prompted slang data against a small participant survey, aiming to map both the potential benefits and the limitations of LLM-assisted dynamic lexicography.

2 Traditional Lexicography versus Dynamic Lexicography

Traditional lexicography relies on highly nuanced, verified data compiled over extensive update cycles (e.g., Cierpich-Kozieł et al., 2025). Conversely, dynamic lexicography attempts real-time updates by monitoring fluid digital streams such as social media, chat logs, and forums (cf. Pęzik, 2020). Modern Transformer-based LLMs (e.g., Vaswani et al., 2017; Lewandowska-Tomaszczyk & Liebeskind, 2024) leverage self-attention mechanisms to detect emerging linguistic contexts on a massive scale.

The current survey of relevant literature brings about a number of different approaches to identify new forms in language with the use of LLMs e.g., Mei et al. (2024), in their paper on slang comprehension, introduce two instruments – SLANG and FOCUS. SLANG is a benchmark designed to autonomously integrate novel data and assess LLMs' ability to comprehend emerging concepts, while FOCUS is an approach that uses causal inference to enhance LLMs to understand new phrases (including metaphoricity) and their colloquial context. They are designed to autonomously integrate novel data and assess LLMs' ability to comprehend emerging concepts.

There are also other most recent attempts at the LLM application in lexicography such as e.g., Curcuglioniti et al. (2026), who present Slang4Pepper, a strategy for constructing a structured glossary of neologisms and slang. The authors use a mixed methods approach including a target-specific corpus of teen-adapted vocabulary. The proposed pipeline is quite demanding, as it includes multi-source collection from diverse online platforms, layered corpus annotation with expert-led labelling (for LLM training), followed by user-based validation. It is meant more precisely for the design of culturally aware dialogue systems and translation for social robots targeting adolescent users.

Other crowd-sourced platforms like *Urban Dictionary* or the Polish *Miejski.pl*, reflect this dynamic shift, although they lack structural control. LLMs offer scale and rapid pattern recognition, yet their outputs cannot be accepted unreflexively. Lexicographic claims regarding currency, register, stability, and distribution require rigorous validation. Therefore, the present exploratory study aims at confronting LLM-generated data (via Gemini Flash) with the data from a preliminary pilot survey of Polish speakers to assess their alignment and divergence.

3 LLM-Based Slang Detection and Contextual Nuance

An LLM was prompted to categorize the semantic adaptations of popular English loanwords operating in Polish (e.g., *cringe, rel, delulu, cheugy*). The model successfully highlighted common structural patterns, such as:

- Neological Hybridization: (e.g., Polish *scamować* 'to scam').
- Contextual Polysemy Detection.
- Social Register Mapping.

However, the model's interpretation of certain items revealed the necessity for strict caution. For example, when evaluating the word *peak*, the LLM flagged the negative British slang usage

(where *peak* denotes an unfortunate or tough situation, reversing its standard polarity). The LLM suggested that Polish speakers "misinterpreted" the term by using it exclusively in a positive sense (*pik możliwości*—'the peak of possibilities'). This interpretation does not correspond to actual facts.

In fact, the word *pik* (and its variants) in Polish has been possibly carrying too much heavy lifting in the Polish language for the negative English slang meaning to push its way in.

The word *pik* came to Polish from two different directions: the French route (in the sense of the downward trajectory from the French *piquer*) is already firmly established in Polish as the concept of 'crashing, diving, or plummeting straight down', associated with the action of picking/piquing with a negative downward trend (Doroszewski, 1958-1968). The late 20th-century English route resulted in Polish in the noun *pik* (from the English *peak*, see Żmigrodzki) in the sense of the upward trajectory was imported to mean the exact opposite: the absolute maximum, the literal highest point on a chart, or a surge of energy (Żmigrodzki, 2007).

This underscores a core limitation: LLMs frequently overinterpret semantic shifts or misattribute lack of adoption as a cognitive error, highlighting the need for corpus and speaker verification.

4 Analysing Transmission Channels and Varieties of Influence

Lexicography requires a precise accounting of linguistic provenance, transmission channels, and speaker exposure. While older loanwords were often traced to specific geographic varieties (e.g., British vs. American English vectors), digital globalization has blurred these distinctions.

4.1 Algorithmic and Digital Transmission

Many contemporary English borrowings are introduced through global digital culture and algorithm-driven media platforms (such as TikTok and YouTube). TikTok, while originating from the Chinese tech company ByteDance, serves as a primary global conduit for predominantly US-centric internet culture, memes, and social commentary.

Social & Conceptual Extensions: The term *Karen* entered Polish discourse directly via global digital networks. While rooted in specific US social dynamics (e.g., the Black Lives Matter movement), it has been functionally mapped by Polish users onto local cultural archetypes, sometimes analogized to the native pejorative *Grażyna*.

Subcultural Lexemes: Terms like *beef* (disagreement/conflict) and *spill the tea* (gossip) flow through these global digital channels, rapidly entering youth vernacular regardless of geographic proximity to the US.

And yet, a more complete lexicographic description would require a deeper understanding of the origin of such words and phrases as *beef* or *spill the tea*. The addition of the etymological notes would enrich the description, signalling that both terms underwent significant semantic drift through subcultural borrowing (https://www.etymonline.com):

> *Beef* drifted from British criminal code → American complaint → hip-hop feud.
> *Spill the tea* drifted from Black queer truth-telling ("T") → digital internet gossip ("tea").

4.2 Educational and Contact-Driven Transmission

Other lexical influences enter Polish via formal English as a Foreign Language (EFL) instruction, specialized professional contact, or bilingual exposure.

Phonetic Indicators: Lexemes like *naura* - a blend of the Polish *nara* which itself is a short form of the Polish farewell *na razie* 'see you'/'bye) and English phonetic elements, particularly [au] – a pseudo-anglicism, presents an interesting case, offering clues into speaker exposure models (Wilczek, 2021).

Functional Differences: inba (from English forum shorthand *inb4/in before)*, comes from the Internet slang used on imageboards to predict a reaction or event before it happens, while in Polish it is used as a noun meaning a drama, scandal, chaotic party, or wild confrontation, describing the event itself, e.g., *Co za inba!* 'What a crazy/chaotic situation!' (Stankiewicz, 2018).

In the context of the present research, **Source-Vectors** refer to the specific geographic and cultural origins (American vs. British) of the English loanwords and the linguistic backgrounds of the individuals who introduce them into Polish discourse. These vectors determine not only the vocabulary used but also the phonetic and social nuances of the slang. The American vector is heavily driven by global digital culture, social media platforms (TikTok, YouTube), and specific US-based social movements. While the American vector dominates social media, the British vector in this study is often associated with formal educational and professional contexts. However, the origin of particular terms must be scrutinized through reliable data sources and etymological dictionaries.

5 Explanatory Pilot Survey

To ground the LLM's outputs, an exploratory pilot survey was conducted with a small, localized sample of 20 Polish subjects (7 tested directly, 13 online, followed by individual discussions).

5.1 Participant Demographics

The sample was distributed unevenly across age categories:

- Ages 13–15 (n = 4)
- Ages 18–24 (n = 10)
- Ages 35–40 (n = 6)

Participants were presented with a list of 50 English slang items active in Polish digital spaces (see Appendix)—generated by Gemini Flash and confronted with three sources: popular lists (e.g., Angloville.pl), a current dictionary of Anglicisms (Cierpich-Kozieł et al., 2025), and tweets on X. The participants were asked to provide definitions, usage contexts, and their approximate frequency of use. The LLM's automated structural categorization (Table 1) served as a comparative baseline.

TABLE 1 LLM structural categorization of slang loans

Category	Examples	Context/Meaning in Polish
Personality/Memes	*Sigma*, *Giga chad*, *Alpha male*	*Sigma* denotes an independent/admirable individual; *Giga chad* is used casually and positively.
Social Interactions	*Rizz*, *Beef*, *Sus*, *Side-eye*	*Rizz* refers to romantic charisma; *Sus* (from *Among Us*) denotes suspicious behavior.
Digital/Gaming	*NPC*, *GG*, *POV*, *Brain rot*	*NPC* refers to individuals perceived as lacking agency; *GG* (*good game*) denotes group success.
Relationships	*Ship*, *Friend zone*, *Crush*, *Delulu*	*Delulu* signifies delusional thinking; *Ship* denotes pairing characters or people.
Generational	*Boomer*, *Karen*, *Grażyna*	Highlights trans-generational conflict and entitlement stereotypes.

5.2 Observations on Adaptations and Demographics

The exploratory data illustrated several structural mechanisms of Polish borrowing integration (see also Mańczak-Wohlfeld, 1995, 2006, 2008, 2010):

1. *Grammatical Integration:* Morphological processing of English roots via Polish verbal suffixes (e.g., *rage'ować*, *missclicknąć*).
2. *Phonetic/Semantic Drift:* Shifts where terms assume specialized meanings, such as *randomowy* drifting from 'unpredictable' to mean 'strange' or weird.
3. *Derivation & Brevity: Suffixation* (*romantyziara* with the English root and the Polish ending) and extensive clipping (*rel* from *relatable*)
4. *Methodological Note on Economy and Clipping Tools*: The initial pilot data highlighted widespread usage of clipped forms, such as *bro* (from *brother*) and *rel* (from *relatable*). However, a key distinction must be made regarding their origin. While these items demonstrate brevity within Polish discourse, they are highly prominent pre-clipped forms in native English internet slang. Polish speakers possibly borrowed these terms directly as pre-packaged, short-form units rather than executing native morphological truncation. Distinguishing direct borrowings of pre-clipped words from genuine intra-Polish truncation requires extensive diachronic corpus validation.

The pilot observations suggest that the youngest demographic (13–15) reports the highest active usage of these borrowed terms. In contrast, older demographics (18–24 and 23–40) report a steep decline in active, unironic usage, shifting instead toward highly contextual, satirical applications, or replacing affective slang with professional English jargon.

Attitudes toward these borrowings also varied: individuals with higher self-reported English proficiency frequently viewed loanwords as expressive *enrichments*, whereas those with lower proficiency tended to view them as intrusive.

6 Methodological Risks of LLM-Driven Lexicography

While LLMs can efficiently assist in initial hypothesis generation, candidate extraction, and draft clustering, they introduce severe systemic risks if left unverified:

- Hallucination and False Etymologies: LLMs are prone to generating non-existent hybrid forms or mapping false semantic paths based on statistical probability rather than authentic linguistic evidence.
- Inability to Decode Irony: Slang operates heavily on subversion, dark humour, and reverse-polarity patterns. LLMs struggle to differentiate between a stable semantic shift and a fleeting, deeply ironic usage context.
- Subcultural Exclusion: Fast-changing, highly insular youth communities or closed digital networks (e.g., private Discord servers, niche forums) are largely inaccessible to the static training data of commercial LLMs, leading to a structural documentation lag.
- Unreliable Frequency Judgments: LLMs lack real-time empirical corpus measures and cannot inherently determine whether a word is broadly currency-backed or merely an isolated internet meme.

7 Proposed LLM-Assisted Lexicographic Pipeline

To safely leverage computational benefits while mitigating the risks detailed in Section 6, we outline a structured, five-phase lexicographic pipeline where the LLM functions strictly as a digital assistant subject to human validation.

[Phase 1: Harvesting] → [Phase 2: Morphological Tagging] → [Phase 3: Pragmatic Mapping] → [Phase 4: Lexicographic Entry Generation] → [Phase 5: Human Verification]

- **Phase 1: Multi-Channel Data Harvesting:** Automated scrapers tap diverse digital ecosystems (social media platforms, video transcriptions) to isolate potential neologism candidates within specific "community bubbles".
- **Phase 2: Morphological & Phonetic Detection:** Algorithmic filtering flags signs of structural polonization, isolating instances of native affixation (e.g., *-ara*, *-ówa*) or verbal inflection (e.g., *-ować*).
- **Phase 3: Semantic & Pragmatic Mapping:** The model generates initial hypotheses regarding domain shifts (e.g., *randomowy*) and notes contextual triggers (e.g., ironic usage signs).
- **Phase 4: Lexicographic Entry Generation:** Automated parsing arranges the extracted data into a standardized draft entry format (Table 2).

TABLE 2 Standardized draft entry format

Entry Component	Pipeline Action	Example
Headword	Standardize Polish orthography.	*Essa (abbreviated as Ez)*
Etymology	Trace source root and channel.	From English *easy* via digital/gaming culture
Definition	Draft nuanced Polish equivalents.	Expression of success, triumph, or relaxation
Register	Propose sociolinguistic markers.	Informal youth slang
Demographics	Flag primary target user groups.	Primary currency: early teenagers

- **Phase 5: Human Verification (Critical Core):** Because of the severe limitations of automated models, all linguistic, etymological, and semantic claims generated by the LLM must be cross-examined. Human lexicographers must validate candidates against authentic, large-scale corpus data, usage contexts, and speaker verification surveys before any entry is formally published.

8 Conclusions and Future Directions

8.1 Limitations of the Present Study

A critical evaluation of the empirical and computational methods reveals several limitations that prevent generalizable claims. The major ones are our questionnaire samples size and representativeness as far as the number of subjects are concerned. The empirical basis of this study is extremely small, comprising only 20 participants distributed unevenly across distinct age brackets. Consequently, these findings are strictly *exploratory and provisional*. They cannot support broader conclusions regarding nationwide Polish youth sociolects, regional distribution, register stability, or definitive frequency statistics. Confirming these trends requires a vastly expanded, balanced corpus and a representative sociological sample.

8.2 Opportunities

Yet, this exploratory study demonstrates that while LLMs present powerful opportunities for accelerating data acquisition, candidate extraction, and initial mapping of fast-evolving English slang in Polish, they cannot (yet) operate autonomously. Their outputs remain highly prone to semantic overinterpretation or problems to navigate localized subcultural contexts.

The empirical findings gathered from our 20-participant sample are strictly provisional and serve primarily to highlight the necessity of human verification within any dynamic lexicographic framework. Human-administered questionnaires and robust empirical corpus linguistics remain the indispensable foundation for authenticating slang integration, formal drift, and eventual trend decay. Future research must scale up these empirical methodologies, testing the proposed pipeline against larger, more representative speaker corpora and examining its adaptability to other language contact environments.

As a final conclusion, we proposed an adequate prompting structure and, in the discussion above, a dedicated LLM pipeline for managing LLM-generated English borrowings in Polish. The proposed Chain-of-Thought (CoT) prompting structure might enable the LLM to systematically analyze slang candidates through specific lenses such as Origin, Adaptation, Pragmatics, and Sociolinguistics. As presented above in Table 2. the recommended Lexicographic LLM-Driven Pipeline for tracking dynamic English-Polish slang neologisms, the system might automate the transition from a media meme to a (proposed) structured dictionary entry.

One of the most powerful aspects of Dynamic Lexicography is the ability to track the life cycle of the loan on the one hand and its "cringe threshold" on the other. By analysing the sentiment value of comments responding to certain words, the LLM identifies the Trend Decay, when a loan word is no longer frequently used starting from a serious, frequent integrated use, via ironic/satirical senses, used to mock the trend, ending finally as obsolete ("cringe", as called by the younger users), i.e. in the 'death phase'. Interestingly enough, the word *cringe*—originally an English verb that shifted into a noun and adjective in American youth slang to describe embarrassment at someone's awkward behavior—is enjoying a remarkably broad lifespan in Poland across the age groups considered in the Questionnaire.

On the other hand, some other slang loans, scrutinized in the survey answers, and exemplified with such items as e.g., *sigma*, *skibidibi*, *delulu*, etc., are considered obsolete or situationally inadequate by more mature (over 24) Polish users with concurrent most widespread usage by the Polish population between ca 13–15-year-old.

The paper concludes with a suggestion that the integration of AI in lexicography to document the fluid nature of modern slang in loanwords will be more effective than manual lexicographic search. And yet, the study indicates a necessity to critically evaluate the LLM-generated data and subject it to – linguistic and lexicographic – scrutiny. By using a combination of traditional lexicographic tools and LLMs we argue that we can be moving from a rather traditional lexicography to descriptive real-time mapping, with the LLM acting as a digital assistant to a linguist – lexicographer, though (still) with the need of human control.

In this framework, human questionnaires serve as a vital verification tool for the LLM-driven Dynamic Lexicography pipeline, ensuring more precise identification of British and American slang borrowings, as well as tracking their formal, semantic, and variations or decay within Polish. The suitability of other LLM models, also for tracking slang elements in other languages, will need to be scrutinized in further studies.

Appendix

Questionnaire

The questionnaire was administered to the Questionnaire subjects in Polish and English, followed by individual face-to-face interviews.

Part 1 English Slang Use

(1) How often do you and your Polish friends use English slang words when you're speaking Polish (list of 50 forms)? Can you give me some examples of words you use?

(2) Are there certain situations or topics where you're more likely to use English slang (a selection of 50 words) ? For instance, when you're talking about games, music, or something else entirely?

(3) Do you notice any differences in who uses English slang more? For example, do boys use it more than girls, or do older students use it more than younger ones?

(4) Are there any English slang words that you or your friends use ironically or as a joke?

(5) Have you ever used an English slang word and had your Polish friends not understand It? What happened then?

(6) Do you think using English slang makes you sound more "cool" or modern among your friends? Why or why not?

Part 2 Sources of learning

(7) Where do you and your friends typically pick up new English slang words? Is it from movies, TV shows, social media (like TikTok or YouTube), music, or something else?

(8) Are there any specific English-speaking YouTubers, TikTokers, or musicians that you and your friends follow who use a lot of slang?

(9) Do you ever intentionally try to learn new English slang, or does it just "stick" with you naturally?

Part 3 Attitudes and Perception

(10) What do you think your teachers or parents would say if they heard you using English slang in Polish? Do you think they'd understand it?

(11) Do you ever feel like English slang is replacing some Polish words or phrases?

(12) In your opinion, is it generally seen as a positive or negative thing to use English slang when speaking Polish among your peers?

(13) Do you think English slang is enriching the Polish language, or is it making it less "pure"?

(14) Are there any Polish equivalents for some of the English slang words you use? If so, why do you choose the English one instead?

Part 4 Bilingual experience

(15) Do you ever find yourself mixing English slang into your Polish conversations without even realizing it?

(16) Do you feel there's a difference in how you express yourself when you're speaking purely Polish versus when you incorporate English slang?

50 candidate words and phrases (Gemini Flash generated):

Rizz, Slay, Brainrot, Skibidi, Delulu, Aura, Lowkey, Highkey, Fr/Fr fr, Side-eye, NPC, Sigma, Brat, Womp womp, Yapping, Cringe, GOAT, Ghosting, Simp, Vibe, Chill, Flex, Sus, Glow up, Drill, Tuff, Bruh, Bro, Bruv, Rel, Very demure, Mewing, Looksmaxxing, Gigachad, Gatekeep, Fanum tax, Ragebait, Situationship, Clout, Crush, Baddie, Salty, ESSA, Glamour, Twin, Innit, Freaky, 6 7, Red flag, Touch grass

References

Angloville.pl. Retrieved February 2, 2026, from https://angloville.pl/materialy/anglicyzmy-poznaj-najpopularniejsze-zapozyczenia-z-jezyka-angielskiego/

Cierpich-Kozieł, A., Mańczak-Wohlfeld, E. & Witalisz, A. (2025). *Nowy słownik zapozyczeń angielskich w polszczyźnie*. https://doi.org/10.12797/9788383682693

Curcuglioniti, D., Pigureddu, L., & Gena, C. (2026). Slang4Pepper entered the chat: Building a slang-aware glossary for Social Robots interacting with Teenagers. In *Joint Proceedings of the ACM UMAP Workshops 2026*. https://ceur-ws.org/Vol-4206/SOCIALIZE-02.pdf

Doroszewski, W. (1958-1968). *Słownik Języka Polskiego*. PWN.

Lew, R. (2011). Online dictionaries of English. In P. A. Fuertes-Olivera & H. Bergenholtz (Eds.), *e-Lexicography: The Internet, Digital Initiatives and Lexicography* (pp. 230–250). Continuum.

Lewandowska-Tomaszczyk, B. & Pawłowski, G. (2025). Testing Chat GPT on terminology generation, defining, translation, and ontology creation in German, English and Polish. *Research in Language*, 23, 320-340.

Lewandowska-Tomaszczyk, B. & Wilson, P.A. (2025). *English language borrowings as neologisms and context markers in Polish slang discourse of teenagers and young adults* [Conference presentation]. Polish Cognitive Linguistics Association conference, Cracow.

Mańczak-Wohlfeld, E. (1995). *Tendencje rozwojowe współczesnych zapożyczeń angielskich w języku polskim*. Universitas.

Mańczak-Wohlfeld, E. (2006). *Angielsko-polskie kontakty językowe*. Wydawnictwo Uniwersytetu Jagiellońskiego.

Mańczak-Wohlfeld, E. (2008). Morfologia zapożyczeń angielskich w językach europejskich. *Studia Linguistica Universitatis Iagellonicae Cracoviensis*, 125. 113–120.

Mańczak-Wohlfeld, E. (Ed.). (2010). *Słownik zapożyczeń angielskich w polszczyźnie*. Wydawnictwo Naukowe PWN.

Mańczak-Wohlfeld, E. & Witalisz, A. (2019). Anglicisms in The National Corpus of Polish: Assets and Limitations of Corpus Tools. *Studies in Polish Linguistics*, 14(4), 171–190.

Mei, L., Liu, Sh., Wang, Y., Bi, B. & Cheng, X. (2024). SLANG: New Concept Comprehension of Large Language Models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 12558–12575). Association for Computational Linguistics. https://aclanthology.org/2024.emnlp-main.698/

Miejski.pl – Miejski słownik slangu i mowy potocznej. Retrieved March 20, 2026.

NOIZZ. Retrieved from https://noizz.pl/rozrywka/co-to-znaczy-ze-ktos-jest-januszem-grazyna-albo-seba/fgfqv88

Pęzik, P. (2020). Budowa i zastosowania korpusu monitorującego MoncoPL. *Forum Lingwistyczne*, *7*(7), 133–150. https://doi.org/10.31261/FL.2020.07.11

Stankiewicz, P. (2018). https://piotrstankiewicz.pl/anglicyzmy-drazniace-i-nie/

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L., & Polosukhin, I. (2017). Attention Is All You Need. In *31st Conference on Neural Information Processing Systems (NIPS 2017)* (pp. 1–15).

Urban Dictionary. Retrieved March 20, 2026, from https://www.urbandictionary.com/

Youth Word of the Year. Retrieved January 29, 2026, from https://www.polskieradio.pl/395/7789/Artykul/3593234,poland%E2%80%99s-youth-language-contest-returns-for-its-anniversary-edition; https://knews.media/2025/11/13/youth-word-of-the-year-2025-surprising-proposals-in-the-final/

Wilczek, A. (2021). (Non)pandemic Youth Speak. On Lexical-Semantic Tendencies in the Polish Language of the Youth. *Roczniki Humanistyczne*, 69, 155–174.

Witalisz, A. (2016). *Przewodnik po anglicyzmach w języku polskim*. Wydawnictwo JAK.

Zabawa, M. (2017). *English semantic loans, loan translations, and loan renditions in informal Polish of computer users*. Uniwersytet Śląski.

Żmigrodzki, P. (2007). *Wielki słownik języka polskiego*. Instytut Języka Polskiego PAN. http://www.wsjp.pl

Software tool:

Google. (2026). *Gemini* (Gemini 3.5 Flash version) [Large Language Model]. https://gemini.google.com/

Contact information

Barbara Lewandowska-Tomaszczyk
Department of Language and Communication, University of Applied Sciences in Konin, Poland
barbara.lewandowska-tomaszczyk@konin.edu.pl

Acknowledgements

The present study was conducted within the COST Action CA22126 *European Network On Lexical Innovation* (ENEOLI).

Tinatin Margalitadze, Giorgi Meladze, Ketevan Mchedlishvili, Tamar Laluashvili & Giorgi Okropiridze

EFFECTIVENESS OF AI FOR THE LEXICOGRAPHY OF LOW-RESOURCE LANGUAGES

The Dictionary of Georgian Neologisms

Abstract The purpose of this paper is to present the results of a study investigating the effectiveness of AI in the compilation of the *Dictionary of Georgian Neologisms*, an ongoing project at the Centre for Lexicography and Language Technologies at Ilia State University. The data for the study were selected from a dataset of 1,700 neologisms identified in previous research that developed a semi-automatic method for detecting neologisms in Modern Georgian (Laluashvili & Margalitadze, 2025). Three major categories of Georgian neologisms were selected: (1) borrowings, including direct loans, morphologically adapted loanwords, and loan translations; (2) word formation, with a special focus on internally created Georgian formations; and (3) semantic evolution. The experiments conducted in this study employed ChatGPT 5.2, Gemini 3.0 Thinking, and Grok 4.1 Thinking in combination to mitigate model-specific bias in handling Georgian linguistic material. These models were applied to the primary task of evaluating the effectiveness of LLMs in identifying the meanings of neologisms and generating lexicographic definitions. The findings demonstrate that LLMs can serve as powerful lexicographic tools capable of effectively processing Georgian linguistic data. However, their performance and accuracy are highly dependent on the prompts designed and employed by lexicographers.

Keywords Dictionary of neologisms; Georgian language; AI; prompts; lexicographic tasks

1 Introduction

The study of neologisms, which originated in France in the 19th century and expanded from the 1970s with the development of lexicological and lexicographic approaches (Salgado, 2026, p. 606), has gained increasing prominence in the 21st century. In today's rapidly changing world, characterized by major technological advances and profound transformations across virtually every sphere of society, the identification, analysis, and documentation of new words have become essential.

1.1 Definition of Neologism

In its broadest sense, a neologism is any new word or expression that enters a language. However, defining what constitutes a "new" lexical unit is more complex than it may initially appear. Some linguists avoid providing a rigid definition, arguing that the concept is largely self-explanatory (Kerremans, 2015, p. 29). Others, however, emphasize the importance of a precise definition. Picone (1996, p. 3), for example, defines a neologism as "any new word, morpheme or locution and any new meaning for a pre-existent word, morpheme or locution that appears in a language." This definition highlights that lexical innovation is not limited to entirely new lexical creations. Existing words may acquire new meanings or functions, and such semantic innovations are likewise regarded as neologisms (Waszink, 2020, p. 31).

Cabré (1998, p. 205) establishes four primary criteria to evaluate whether a lexical unit is a genuine neologism:

1. Diachrony: The unit is a neologism if it has emerged within a recent, defined historical period;
2. Lexicography: The unit is a neologism if it is not yet recorded in established dictionaries;
3. Systematic Instability: The lexical unit shows formal instability (morphological, graphic, phonetic) or semantic fluidity;
4. Psychology: The lexical unit is perceived by speakers as a new.

These criteria for identifying neologisms align with the opinions of other linguists in which novelty is the primary characteristic of a neologism, marked by chronological appearance, formal or semantic instability, and speaker perception (Link, 2021, p. 32). In Georgian linguistics, Gamkrelidze (2008, p. 389) supports these ideas, defining neologisms as lexical units "which have appeared in the language no earlier than the last decade and are perceived by speakers as new."

According to Trap-Jensen, there is "no any generally accepted period of time for which a word has to have been in use before it ceases to be considered new" (2020, p. 12). Therefore, he sets 1990 as the cut-off point for his project and defines words and multiword expressions as new if their earliest recorded use dates from 1990 onwards. The Algemeen Nederlands Woordenboek (ANW) considers a word as a neologism if it was coined in or after the year 2000 (Waszink, 2020, p. 34). More recent studies use the year 2005 as a threshold (Kilim, Soojin & Hae-Yun, 2025).

Since no dictionary of neologisms has previously been published in Georgia, the dictionary compiled within our project represents the first attempt of its kind. Moreover, the latest (third, revised and supplemented) edition of Chabashvili's Dictionary of Foreign Words was published in 1989 and has not been updated since. Therefore, for the purposes of this project, we define as neologisms lexical units and meanings that have entered the Georgian language over the past 20 years (Laluashvili & Margalitadze, 2025).

1.2 AI in Lexicography

In his paper, de Schryver (2023, p. 358) states that "lexicographers have always been at the forefront of dreaming up ways to harness the latest technology in order to compile better dictionaries". Digital and corpus revolutions created new possibilities for traditional lexicography and expanded the toolkit available for lexicographers (Lew, 2024, p. 3). Since the launch of ChatGPT in 2022, the new age of successful application of generative AI in lexicography, also known as the "post-editing lexicography" model, has dawned (de Schryver, 2023, p. 355; Marković & Stanković, 2025, p. 79). Some authors expressed optimism about integrating generative AI into the lexicographic workflow (de Schryver, 2023, p. 361), while others expressed doubts about its limitations (Jakubíček & Rundell, 2023) and debated whether reliance on this tool would render lexicography, dictionaries, and lexicographers redundant. Research indicates that in-the-flesh lexicographers are still crucial for the lexicographic process, particularly at the beginning for crafting optimal prompts and later at the so-called vetting stage to evaluate and refine generated output (de Schryver, 2023, p. 361). Despite these controversies, the emergence of this software tool represents a significant development that cannot be overlooked.

A growing body of literature has emerged on the application of AI in lexicography, with AI being used across all macrostructural and microstructural components of dictionary entries (de Schryver, 2023; Lew, 2024; Trap-Jensen, 2024; de Schryver, 2025). While early studies

focused primarily on the use of ChatGPT for the compilation of English dictionaries (de Schryver, 2023; Lew, 2023), more recent research has explored the applicability of other models to bilingual and multilingual dictionaries, less-resourced languages, and neology (Marković & Stanković, 2025; de Schryver, 2025; Georgiou, 2025).

The success rate of testing LLMs for different lexicographic tasks varies. As Trap-Jensen (2024, p. 30) states, there is broader agreement among researchers regarding LLMs, particularly ChatGPT's capabilities, to generate fluent, acceptable dictionary definitions, and it is argued that ChatGPT performs well as a "defining lexicographer". It is also argued that the results can be further enhanced by refining prompts, customizing models, or feeding them with corpus-based illustrative examples (Barrett 2023, as cited by de Schryver 2023, p. 362; McKean and Fitzgerald 2023, as cited by de Schryver 2023, p. 369; de Schryver 2025).

A closely related task to the successful generation of definitions is the identification of senses and word-sense disambiguation, which have proved considerably more challenging for LLMs (Jakubíček & Rundell, 2023, as cited by de Schryver, 2023, p. 370; Trap-Jensen, 2024, p. 32). Several studies show that LLMs tend to split a single meaning into multiple artificial senses, merge distinct senses, omit an important sense, or invent senses not supported by actual usage evidence. They therefore function more as sense "splitters" than lexicographic "lumpers" (Jakubíček & Rundell, 2023; Marković & Stanković, 2025).

These challenges are further intensified when defining neologisms using generative AI tools. While human speakers can adapt to ongoing language change, large language models may encounter difficulties due to their dependence on training data, which may not adequately reflect recent linguistic developments (Zheng, Ritter & Xu, 2024, p. 13885). As Henisch (2025, p. 76) states: "hallucinations are more likely in niche domains, rapidly developing domains, or newly emerging domains, as not all current data can be present during training of the model".

Despite the advantages of using AI ChatBots, such as user-friendliness, speed, and the potential to lighten a lexicographer's workload, LLMs still have serious methodological and linguistic limitations that warrant consideration. One of the central problems is the lack of reliability and replicability (Trap-Jensen, 2024, p. 35). LLMs do not disclose the sources on which their outputs are based, and their probabilistic nature means that the same prompt may produce different outputs on different occasions (de Schryver, 2023, p. 358; Trap-Jensen, 2024, pp. 33, 35). These factors may constitute serious limitations for lexicographic work. Another important issue concerns the application of LLMs beyond English. Previous studies have highlighted the English-centredness of LLMs, including the fact that more than 90% of the training data used for ChatGPT was in English and that prompts in other languages were often processed through English-based patterns (de Schryver, 2023, p. 370; Trap-Jensen, 2024, p. 35). Despite a growing body of scientific papers primarily focused on applying LLMs to English lexicography (de Schryver, 2023; Lew, 2024; Heinisch, 2025; de Schryver, 2025), research on applying GenAI tools to Georgian is relatively new (Lomidze, 2025; Chighladze, 2025; Mchedlishvili, 2025). Given the rapid development of AI and the ongoing transformation of lexicography in response to these developments, there is a growing need to explore how generative AI tools can be effectively integrated into Georgian lexicographic practice, particularly in the compilation of a dictionary of neologisms.

2 Data and Methodology

The data for this study were selected from 1,700 neologisms identified in previous research that developed a semi-automatic method for detecting neologisms in Modern Georgian (Laluashvili & Margalitadze, 2025). This approach is based on the exclusion method. A study corpus was compiled for the project, comprising textual material from Georgian-language online newspapers, magazines, news websites, media platforms, and the websites of NGOs and governmental agencies. The corpus contains over 100 million tokens and covers the last 20 years of the development of the Georgian language. Two Georgian-language lemmatisation tools based on finite-state technology were applied in the study. The first lemmatiser, created at Ilia State University and containing the macrostructures of existing academic dictionaries, was used as an exclusion list. The second lemmatiser, developed for the Georgian National Corpus was used to reduce the number of out-of-vocabulary (OOV) lexical items (Laluashvili & Margalitadze, 2025). The neologisms collected in the above-mentioned study were classified according to semantic and formal criteria. The main semantic categories distinguished include politics, technology and the internet, media, medicine, education, and society and social life, among others. Regarding formal classification, the most substantial group of Georgian neologisms consists of borrowings, followed by word-formation and semantic evolution (Laluashvili, 2025).

For further analysis, the three most numerous categories of Georgian neologisms were selected: 1) borrowings, including direct loans, such as ბრაუზერი / *brauzeri* (“browser”), morphologically adapted loanwords, e.g., და-ლაიქ-ება / *da-laik-eba* (“to like”), and loan translations, such as სოციალური ქსელი / *sotsialuri kseli* (“social network”); 2) word-formation, e.g., გა-სანდო-ება / *ga-sando-eba* (“to make a device trusted”); and 3) semantic evolution, e.g., გვერდი / *gverdi* (“page”) – denoting a social media page. To evaluate the effectiveness of the AI-assisted lexicographic workflow, three freely available, publicly accessible LLM web interfaces – ChatGPT 5.2, Gemini 3.0 Thinking, and Grok 4.1 Thinking – were used under their default settings. The use of multiple models was intended to mitigate model-specific bias in processing Georgian. No API access was used, and no additional settings were manually adjusted. Given the rapid evolution of these models and the frequent updates to web-based LLM systems, the model names should be understood only as interface-level identifiers.

The above-mentioned models were used to perform the main task of systematically evaluating the effectiveness of LLMs in identifying the meanings of neologisms, distinguishing senses in cases of polysemy, and generating lexicographic definitions. In the context of the present research, ‘effectiveness’ is understood not as benchmark accuracy in a quantitative sense, but as the practical adequacy of the generated outputs, assessed during the validation stage through qualitative review by lexicographers. For in-depth evaluation, we randomly selected 25 neologisms from each assigned category, resulting in a sample of 75 borrowings representing three subclasses – direct loans, morphologically adapted loanwords, and loan translations – as well as 25 items from each of the categories of word formation and semantic change.

All prompts were formulated and submitted in Georgian, with responses generated across multiple independent sessions between September 2025 and February 2026. The outputs were preserved in the original chat histories and subsequently exported to an Excel spreadsheet. The archived dataset includes the lexical unit, the model, the prompt type, the generated output, and the assigned qualitative adequacy category. These archived outputs served as the basis for the qualitative analysis. For the majority of tasks, two prompts were employed: a so-called basic prompt and a prompt enriched with corpus-based examples (see Table 1). The total number of interactions with ChatBots per-model and per-prompt equaled to 750.

TABLE 1 The prompts used in the study

Basic Prompt	Context-Enriched Prompt
You are a professional lexicographer, working on the dictionary of Georgian neologisms. Create a lexicographic definition for the Georgian Neologism [*ambasadoriali*] and, if applicable, distinguish the polysemous meanings of the word in question	You are a professional lexicographer, working on the dictionary of Georgian neologisms. Create a lexicographic definition for the Georgian Neologism [*ambasadoriali*] and, if applicable, distinguish the polysemous meanings of the word in question based on the following illustrative examples: . . . [corpus-based contexts]

Outputs were assessed manually through an expert qualitative analysis. Each lexicographer working on the dictionary of neologisms compared the AI-generated entries with the entries they had compiled themselves, analyzed the data, and classified the lexical units into the designated categories. The results were then presented to and discussed with the lexicographic team. This team review served as an expert validation stage, ensuring consensus among the lexicographers regarding the categorization of AI-generated outputs. The assessment focused on the overall acceptability and adequacy of the output and used three evaluation categories:

1. Fully Adequate – all polysemous senses were correctly identified, and the drafted lexicographic definitions were contextually and grammatically adequate;
2. Partially Adequate – at least one polysemous sense was correctly identified, but other senses were omitted, insufficiently specified, or hallucinated, and the definitions lacked contextual or grammatical accuracy;
3. Inadequate – senses were not identified, LLMs hallucinated, and the drafted lexicographic definitions were not usable.

Mixed cases, in which one meaning was correctly identified, but another was missing or hallucinated, were consistently classified as Partially Adequate. Therefore, for the selected sample, borderline cases were not attested.

3 Case studies: crafting lexicographic definitions and distinguishing polysemous senses

3.1 Direct loans, morphologically adapted loanwords and loan translations

AI models handle lexicographic tasks well for direct loans, even with a basic prompt, when the borrowed word has the same meaning(s) in Georgian as in English. They handle particularly well loanwords from the fields of technology, media, and politics. For instance,

(1) in the case of the neologism დეტოქსიკაცია / *det'oksik'atsia* ("detoxification"), using a basic prompt, all three models identified two main senses and generated 2 lexicographic definitions: (1) (medicine) – "the removal of poisons, drugs, alcohol, narcotics, or other harmful substances from the body through medical methods"; (2) (social media, internet) – "taking a temporary break from digital devices, social media, or the internet to reduce stress, addiction, and information overload". Subsequent corpus-based concordance analysis attested that this direct borrowing in Georgian possesses two distinct meanings, mirroring the polysemous extension of its English counterpart.

(2) Turning to the analysis of another borrowing, namely ამბასადორიალი / *ambasadoriali* ("conference of ambassadors"), this neologism in Georgian refers to the annual forum of ambassadors. However, since the English term "ambassador" may also denote brand representatives and brand ambassadors, all LLMs, using a basic prompt, demonstrated visible bias towards English. In addition to the primary meaning related to the conference of ambassadors, ChatGPT 5.2 and Gemini Thinking 3.0 identified and crafted a definition for the meeting of the representatives of luxury brands. By contrast, Grok 4.1 Thinking produced a hallucination, relating the Georgian neologism to the name of the hotel Ambasadoriali in Tbilisi.

A similar pattern is observed in morphologically adapted loanwords, particularly in the domains of technology and the internet. For instance, terms such as დალაიქება / *dalaikeba* ('to like'), დატესტვა / *dat'est'va* ("to test"), დაბლოკვა / *dablok'va* ("to block") yielded well-generated lexicographic definitions with a basic prompt. These results suggest that all models perform well when borrowed words from English retain similar meanings in Georgian. In such cases of direct lexical correspondence, using basic prompts produces satisfactory answers.

A comparable trend is evident in loan translations as observed in direct borrowings. If artificial intelligence correctly links a Georgian calque to its English equivalent, it provides an accurate explanation. For example,

(3) below is given the definition of მედიალანდშაფტი / *medialandshapt'i* ("media landscape") generated by Gemini 3.0 Thinking: "A general overview of the media outlets, platforms, technologies, and audiences in a specific country, region, or globally. The media landscape describes what types of media outlets (television, radio, press, online publications, social networks) exist in the market, how their shares are distributed, and how the public consumes information".

Table 2 presents a summary of the performance of all three models using the basic prompt in the case of borrowings.

TABLE 2 Evaluation results for the three LLMs using the basic prompt for borrowings

Prompt type	ChatGPT 5.2	Gemini Thinking 3.0	Grok Thinking 4.1
Basic prompt	**Direct loans**: Fully Adequate: 56% (13/25) Partially Adequate: 36% (9/25) Inadequate: 12% (3/25) **Morphologically adapted loanwords**: Fully Adequate: 92% (23/25) Partially Adequate: 8% (2/25) Inadequate: 0% (0/25) **Loan translation**: Fully Adequate: 100% (25/25)	**Direct loans**: Fully Adequate: 64% (16/25) Partially Adequate: 24% (6/25) Inadequate: 12% (3/25) **Morphologically adapted loanwords**: Fully Adequate: 96% (24/25) Partially Adequate: 4% (1/25) Inadequate: 0% (0/25) **Loan translation**: Fully Adequate: 100% (25/25)	**Direct loans**: Fully Adequate: 64% (16/25) Partially Adequate: 24% (6/25) Inadequate: 12% (3/25) **Morphologically adapted loanwords**: Fully Adequate: 96% (24/25) Partially Adequate: 4% (1/25) Inadequate: 0% (0/25) **Loan translation**: Fully Adequate: 100% (25/25)

The next phase involved testing AI models with context-enriched prompts. All example sentences were carefully selected by lexicographers, participants of the present study from two Georgian corpora: *The Corpus of Georgian Neologisms*[1] (Laluashvili & Margalitadze, 2025) and *Araneum Georgianum,* a Georgian corpus, developed by V. Benko within the *Aranea Project* at the Slovak Academy of Sciences.[2]

The study produced interesting and promising results when context-enriched prompts were used. Even when satisfactory results were obtained with a basic prompt, the definitions became more targeted and refined with context-enriched prompts. With regard to resolving instances of semantic divergence in Georgian (e.g. the case of *ambasadoriali*), context-enriched prompts proved indispensable and improved considerably the output of the models. Figure 3 summarises the performance of all three models using the context-enriched prompt in the case of borrowings.

TABLE 3 Evaluation results for the three LLMs using the context-enriched prompt for borrowings

Prompt type	ChatGPT 5.2	Gemini Thinking 3.0	Grok Thinking 4.1
Context-enriched prompt	**Direct loans**: Fully Adequate: 100% (25/25) **Morphologically adapted loanwords**: Fully Adequate: 100% (25/25) **Loan translation**: Fully Adequate: 100% (25/25)	**Direct loans**: Fully Adequate: 100% (25/25) **Morphologically adapted loanwords**: Fully Adequate: 100% (25/25) **Loan translation**: Fully Adequate: 100% (25/25)	**Direct loans:** Fully Adequate: 100% (25/25) **Morphologically adapted loanwords:** Fully Adequate: 100% (25/25) **Loan translation**: Fully Adequate: 100% (25/25)

During the research process, we observed several instances in which the meaning of an English loanword in Georgian diverged from its English equivalent, and the models were still able to effectively identify and describe this divergence using the basic prompt.

(4) This was the case with *ჯოისტიკი / joist'ik'i* ("joystick"). It possesses two meanings in Georgian, the second of which is incorrectly used instead of "a game controller", "a gamepad". The AI provided the following comment on this issue: "Even though today [in the Georgian language], *joystick* is often erroneously called any game controller (gamepad), its technical precision must also be maintained in the dictionary. If previously we called any gaming device a joystick, today users are increasingly using *gamepad* to denote a specific type of controller".

(5) The same applies to *არტჰაუსი / art'hausi* ("art house"). In English, the term is monosemous, referring to "a cinema which specialises in showing films that are artistic or experimental rather than merely entertaining,". In Georgian, however, this neologism has developed another meaning: "Cinematic genre/style". Despite the fact that in English *art house* is used only in one meaning, AI did not rely on a direct analogy with English and also correctly defined the "cinematic genre (art film, art house film)" as one of the meanings in the Georgian language.

[1] https://neologism.iliauni.edu.ge/

[2] http://unesco.uniba.sk/aranea/run.cgi/first?corpname=AranGeor&reload=1

The analysis of such examples suggests that access to substantial Georgian illustrative material enables AI models to satisfactorily handle even cases of semantic divergence in Georgian borrowings. Such examples are more frequently encountered when using Gemini, as this model has access to Georgian material and does not always follow the English analogy.

The error analysis enabled us to discover some reasons for AI mistakes and hallucinations in borrowings.

(6) An interesting case to consider is the term ბასლაინი / *baslaini* ("bus lane"), which exemplifies how language models can generate inaccurate information (hallucination), particularly in the case of ChatGPT. Initially, ChatGPT was tasked to define the word without example sentences. Since ბასლაინი / *baslaini* is a direct borrowing from English, it was assumed that AI would easily deliver an accurate definition. However, it produced a hallucination by providing a musical definition, interpreting the term as bassline, "a low-pitched musical part that is played by instruments like the bass guitar in a musical work." The source of this confusion might be the Georgian version of the word itself. The term was adopted into Georgian from the English bus lane. Although the correct phonetic transliteration based on the English pronunciation [leɪn] should be ბასლეინი / *basl**e**ini*, the word has become established in common usage with the incorrect spelling ბასლაინი / *basl**a**ini*, probably due to a semantic association with the English word line. Ultimately, this error was corrected by providing corpus-based materials, which enabled the model to offer the accurate definition "a specially designated lane on a roadway intended for the exclusive use of public transport (buses)."

AI may also have difficulties in recognizing the meanings of some Georgian neologisms because many English-sounding Georgian neologisms, or those containing English components, are in fact borrowed from Russian rather than directly from English. The Russian language continues to influence Georgian even after the collapse of the Soviet Union. This may be partly explained by the fact that a considerable segment of the Georgian population is more fluent in Russian than in English.

(7) This is the case with the Georgian neologism პიარი / *piari* (PR). This neologism has two meanings in Georgian: 1. public relations, PR; 2. publicity, promotion of smth or smb. The second meaning seems to have developed under the influence of the corresponding Russian word.

3.2 Georgian word-formation

Another class of neologisms belonged to word-formation using native Georgian roots. In this case, the performance of all LLMs was markedly inferior, and the level of hallucinations was higher in comparison to borrowings. Furthermore, different models produced outputs with varying degrees of correctness.

(8) Let us consider the political and educational neologism ბარიერგადალახული / *bariergadalakhuli* (*"Political parties or students* who met the minimum threshold scores"). This term consists of two roots: "barrier + something that was overcome". When utilizing the basic prompt, ChatGPT 5.2 produced a hallucination by erroneously linking the Georgian word *barieri* ("barrier") to the primary meaning of the English word *barrier* – "obstacle, difficulty." Consequently, it generated three senses and lexicographic definitions, none of which are attested in Modern Georgian.

On the contrary, the output generated with a basic prompt by Grok 4.1 Thinking was strikingly accurate. This once again highlights the varying levels of bias in models towards English and the different amounts and sources of training data for the Georgian language. Grok identified one of the main senses and even specified within the generated lexicographic definition that both the word and its meaning are new to Georgian. The generated output translated into English, is as follows: a political party (or group/coalition of parties) that crossed the 5% (or equivalent) electoral threshold in the 2024 parliamentary elections and received mandates in parliament (often used in the context of opposition parties that boycott the elections or consider them illegitimate, but still crossed the threshold). This output might indicate that Grok has access to the social media platform X, where many election-related texts, including this particular neologism, were posted in Georgian. Although Gemini 3.0 Thinking identified the correct sense, it also generated two additional senses and definitions that were clear instances of hallucination. One of the meanings was related to overcoming mental problems, and another to overcoming physical hindrances.

While some models handled the task better than the others, the prompt enriched with selected corpus-based examples had a significant effect on the generated output. Based on the contextual analysis, all 3 models identified two main senses alongside crafting their lexicographic definitions: (1) referring to a political party that has accumulated the minimum percentage of votes established by electoral legislation in elections and which is entitled to participate in the distribution of mandates in parliament; (2) referring to a student, teacher, or test taker who has passed a specified minimum score threshold or standard and is considered to have successfully passed the test stage. It should be mentioned that none of the models managed to identify the latter sense when utilizing only a basic prompt.

(9) A similar case is the neologism *დახარვეზება / dakharvezeba* (“to find smth. faulty, defective or inadequate”), which is entirely a Georgian formation. It is derived from the noun *ხარვეზი / kharvezi* (“a defect, a fault”), by adding the prefix *da-* and the suffix *-eba* (da-kharvez-eba). With the basic prompt, the models were misled by the primary meaning of the noun “a defect, a fault” and defined the neologism as “the process by which a specialist checks whether an item has technical or visual defects, and records these defects.” It should be noted that *დახარვეზება / dakharvezeba* is not used in Georgian with this meaning and denotes the following “to find smth. faulty, defective or inadequate.” For example: to find documents, a project, etc., faulty, inadequate and reject them. After being provided with corpus examples, all three models were able to correctly interpret the Georgian neologism.

In contrast to the aforementioned cases, there were examples of internal creations in Georgian, where the models were unable to correctly interpret Georgian neologisms even after the corpus material had been provided.

(10) Such is the case with a banking neologism *გასანდოება / gasandoeba* (“to make a device trusted, to add a device as trusted in order to log in without OTP”). This neologism is also based on Georgian roots, the adjective *სანდო / sando* (“trustworthy, reliable”) and the suffix and prefix *ga–eba* (ga-sando-eba). In the case of this neologism as well, the adjective *სანდო / sando* (“trustworthy”) misled the AI models, which resulted in the misinterpretation of the derived word “the temporary transfer of property, money or a right, without changing the ultimate ownership.”

There were a few, albeit different instances, when Georgian internal creations were understood, described and defined by the models with a high degree of accuracy. Our analysis

suggests that in such cases, the LLMs correctly associated Georgian words with their English equivalents and interpreted Georgian neologisms correctly.

(11) This was the case with the word *მოწყვლადი / mots'q'vladi* ("vulnerable"), which was paired with English *vulnerable*. This neologism consists of only Georgian roots: *მოწყვლა / mots'q'vla* ("to wound, to hurt") and the suffix *-ადი / adi* (mots'q'vl-adi). To our surprise, all three models processed this neologism effectively, even providing a correct etymological explanation. Although after providing additional corpus examples, containing collocations such as "vulnerable taxons," "vulnerable ecosystem," "vulnerable sectors of economics," the models reinterpreted the definition and split it into several senses. Splitting senses is a characteristic feature of the AI models (Jakubíček & Rundell, 2023). At the end, it is up to lexicographers to make final decisions about the number of senses, as well as splitting or lumping them.

The qualitative analysis revealed that Grok Thinking 4.1 handles this particular category of Georgian neologisms slightly better than the other two models, even when provided with a basic prompt. At the same time, this category of neologisms clearly demonstrates the importance of enriching prompts with corpus-based examples in order to identify correct meanings and generate usable lexicographic definitions. Table 4 summarises the evaluation results of the three LLMs across two prompt types for Georgian word formations.

TABLE 4 Evaluation results for the three LLMs across two prompt types for Georgian word formations

Prompt type	ChatGPT 5.2	Gemini Thinking 3.0	Grok Thinking 4.1
Basic prompt	Fully Adequate: 56% (14/25) Partially Adequate: 28% (7/25) Inadequate: 16% (4/25)	Fully Adequate: 64% (16/25) Partially Adequate: 24% (6/25) Inadequate:12% (3/25)	Fully Adequate: 88% (22/25) Partially Adequate: 8% (2/25) Inadequate: 4% (1/25)
Context-enriched prompt	Fully Adequate: 96% (24/25) Inadequate: 4% (1/25)	Fully Adequate: 96% (24/25) Inadequate: 4% (1/25)	Fully Adequate: 100% (25/25)

3.3 Semantic evolution

The results of the semantic evolution revealed that all LLMs successfully identified new senses and generated definitions in Georgian when English words had analogous new meanings, i.e. when new senses of Georgian words were semantic borrowings from their English equivalents.

(12) An illustrative example is *გვერდი / gverdi* ("a page") referring to a social media page in both languages. However, in this case, different models yielded diverging outputs.

(13) For instance, let us discuss the Georgian technology- and Internet-related neologism *გაზიარება / gaziareba* ("to share"). This term was coined by analogy with its English counterpart, as a result of the semantic evolution of its primary sense, which is related to sharing a portion of something with another person or others. In this case, ChatGPT 5.2 and Gemini 3.0 Thinking with a basic prompt lumped the new meaning in one lexicographic definition and sense: (1) making digital content (text, image, video, link, etc.)

available to others via the Internet and social media. However, Grok 4.1 Thinking distinguished two additional technology-related senses: (1) sharing internet connection; (2) sharing screen (documents, presentation) during the video conference through such platforms as Zoom. In this case, Grok demonstrated a tendency towards splitting technology-related meanings. The cumulative analysis of all the meanings and definitions provided by different models suggests that the final determination of sense distinction within a dictionary remains a matter of lexicographer's discretion.

As these examples show, the models perform effectively when addressing semantic borrowings from English, where the Georgian word has developed a new sense by analogy with its English equivalent. However, when a semantic change in Georgian has no analogue in English, such neologisms can be considerably more challenging for LLMs to analyze.

(14) From this point of view, it is interesting to analyze the word *არაკონკურენტული / arak'onk'urent'uli* ("not competitive"). Beyond the standard meaning "without competition, uncompetitive," this word has developed a new sense in the Georgian language, namely "incapable of competing, non-competitive; unable to compete with someone or something." For example, "Products made with slave labour are not competitive with those produced according to European standards". The LLMs failed to recognize this meaning. Even after providing example sentences with this meaning, none of the models was able to explain this new sense of the word.

This also applies to cases of the terminologisation of common Georgian words.

(15) For instance, when using a basic prompt, all three models failed to provide an accurate definition for the terminological neologism *ადევნება / adevneba* ("stalking") that was the result of the terminologisation of a common Georgian word. All three models provided the meaning of the general language "to follow back". Later, we specified the basic prompt indicating that the neologism belongs to the legal domain, and the resulting definition was slightly more accurate, specifically highlighting that this action "causes a person to have a well-founded fear for their own safety." Only after providing the specific article from the Georgian legislation along with additional corpus-based examples and instructing the model to base its lexicographic definition strictly on the provided information, did the generated definition become satisfactory. At this stage, the output was well-aligned with the law accepted in the Georgian legislation system. This case once again highlights and supports the idea of existing cultural, legislative, and linguistic bias towards English (Trap-Jensen, 2024; Pensec & Williams, 2025).

The analysis presented in this chapter clearly demonstrates that the provision of corpus materials is essential for generating appropriate dictionary definitions when using large language models. While LLMs possess the capacity to identify semantic shifts, they require a suitable environment that provides sufficient information, such as illustrative examples. This enables them to incorporate the relevant information and correct initially inaccurate definitions, as they can identify patterns in the provided data and generate more accurate conclusions.

Table 5 summarises the evaluation rate for 3 LLMs across two prompt types for semantic evaluation.

TABLE 5 Evaluation results for the three LLMs across two prompt types for semantic evolution

Prompt type	ChatGPT 5.2	Gemini Thinking 3.0	Grok Thinking 4.1
Basic prompt	Fully Adequate: 88% (22/25) Partially Adequate: 8% (2/25) Inadequate: 4% (1/25)	Fully Adequate: 88% (22/25) Partially Adequate: 8% (2/25) Inadequate: 4% (1/25)	Fully Adequate: 88% (22/25) Partially Adequate: 8% (2/25) Inadequate: 4% (1/25)
Context-enriched prompt	Fully Adequate: 92% (23/25) Inadequate: 8% (2/25)	Fully Adequate: 92% (23/25) Inadequate: 8% (2/25)	Fully Adequate: 92% (23/25) Inadequate: 8% (2/25)

4 Conclusion

The case studies discussed above have revealed that LLMs serve as powerful lexicographic tools capable of effectively processing Georgian linguistic data. However, much depends on the prompts utilized by lexicographers, which must be tailored to the specific task at hand.

In most cases, LLMs perform particularly well with direct borrowings, such as *ვებსაიტი / vebsait'i* ("website"), *გრინვოშინგი / grinvoshingi* ("greenwashing"). These concepts are widely recognised and disseminated in global databases and are monosemous. Consequently, LLMs grasp their meanings without additional context and even generate highly relevant illustrative examples. A similar behaviour is observed in morphologically adapted loanwords, and loan translations where the meaning of a Georgian neologism coincides with its English equivalent. All models perform effectively when terms borrowed from English (whether direct, morphologically adapted or calque) retain their original meanings in Georgian. In such cases, using basic prompts yields satisfactory results.

However, such borrowings do not represent the full spectrum of Modern Georgian neologisms. There are instances where a borrowed term has developed a divergent meaning in Georgian (e.g. *ამბასადორიალი / ambasadoriali* "conference of ambassadors"), cases where the borrowing is polysemous in both English and Georgian with different senses, occasions where a borrowing is established in Georgian incorrectly (e.g. *ბასლაინი / baslaini* "bus lane") or is borrowed from Russian and not directly from English (e.g. *პიარი / piari* "PR"). Therefore, it is important, even in the case of borrowings, to develop a more sophisticated prompt enriched with corpus examples and to analyze the AI output in detail.

Another class of neologisms belonged to word-formation using native Georgian roots. In this case, all LLMs performed much worse, and the level of hallucination was higher in comparison to borrowings. Furthermore, the degree of accuracy varied considerably between different models. In such instances, the importance of correctly formulated and context-enriched prompts was even higher. All three models interpreted these native formations more adequately once provided with such prompts. Nevertheless, there were cases of Georgian internal creations, when the applied models were unable to provide an adequate definition and hallucinated even after providing corpus examples. A similar pattern was observed in the cases of semantic evolution that were not based on semantic borrowing from English but arising entirely from Georgian internal creation. It is worth noting that other studies have likewise confirmed that the performance of AI models varies depending on the linguistic origin of words (Zheng, Ritter & Xu, 2024, p. 13885). Similar findings have also been reported in recent research on different types of Greek neologisms (Georgiou, 2025, pp. 1–2).

Regarding the lexicographic task of distinguishing polysemous meanings of a word, AI proved to be efficient in solving this task, depending on the information provided to applied models. As the study has shown, the performance of AI models depends on the corpus examples that lexicographers carefully select from corpora. Prompts enriched with such corpus examples substantially improve the quality of the outputs; however, the extent of this improvement varies considerably across models. Consequently, a cumulative analysis of all the meanings and definitions provided by various models leaves it to the lexicographer to decide how nuanced the sense distinction should be for any given polysemous Georgian neologism.

Concerning the models applied in the study, Gemini Thinking 3.0 and Grok Thinking 4.1 were slightly better than ChatGPT 5.2 for borrowings. Grok Thinking 4.1 performed much better than other two models for Georgian word-formation. All three models yielded the similar results for semantic evolution. With the context-enriched prompts all AI models improved considerably their output.

Based on the findings of this study, we conclude that, given the status of Georgian as a low-resource language and the fact that neologisms constitute a specific segment of its vocabulary, lexicographers should first analyze the target word, examine illustrative sentences, and identify its core senses. Only at a later stage should AI be employed as a supportive tool for addressing different lexicographic tasks. Furthermore, it is advisable to apply two or three AI models in parallel to mitigate model-specific bias when processing Georgian data.

References

Barrett, G. (2023). Defin-O-Bots: Challenging A.I. to Create Usable Dictionary Content. In the *24th Biennial Conference of the Dictionary Society of North America* (DSNA 24) (pp. 10-11).

Cabre, T. M. (1998). *Terminology: Theory, Methods and Applications*. John Benjamins Publishing.

Chabashvili, M. (1989). *A Dictionary of Foreign Words* (in Georgian). Ganatleba.

de Schryver, G.-M. (2025). Customised GPTs for Lexicography. *Lexikos*, 35(2), 397–422. https://doi.org/10.5788/35-2-2104

Gamkrelidze, T. (2008). *Theoretical Course in Linguistics* (in Georgian). Tbilisi University Publishing.

Georgiou, G. P. (2025). Can AI mimic the human ability to define neologisms? *arXiv*. https://doi.org/10.48550/arXiv.2502.14900

Google Gemini. (2026). *Gemini Response to the authors*. Retrieved February-March 2026, from https://gemini.google.com

Heinisch, B. (2025). Next-gen Terminology: Transforming Terminology Work with Large Language Models. *Across Languages and Cultures*, 26 (S), 64–80. https://doi.org/10.1556/084.2025.01061

Jakubíček, M. & Rundell, M. (2023). The End of Lexicography? Can ChatGPT Outperform Current Tools for Post-Editing Lexicography? In M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas & M. Jakubíček (Eds.), *Proceedings of the eLex 2023 Conference: Electronic Lexicography in the 21st Century* (pp. 508–523). Lexical Computing.

Kerremans, D. (2015). *A Web of New Words: A Corpus-Based Study of the Conventionalization Process of English Neologisms* (Vol. 15). Peter Lang Verlag.

Kilim, N., Soojin, L. & Hae-Yun, J. (2025). Detection and Description of Neologisms in Korean Lexicography: Methodological Issues in Corpus Balance, Word Unit Bias and LLM Assistance. *Lexikos*, 35(1), 414–438. https://doi.org/10.5788/35-1-2045

Laluashvili, T. & Margalitadze, T. (2025). Semi-Automatic Detection of New Words in Modern Georgian. *Lexikos*, 35(1), 399–413. https://doi.org/10.5788/35-1-2044

Laluashvili, T. (2025). Semantic and Formal Classification of Modern Georgian Neologisms. In T. Margalitadze (Ed.), *Proceedings of the II International Conference Lexicography in the XXI Century* (pp. 36–48). Ilia State University.

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, 11, 426. https://doi.org/10.1057/s41599-024-02889-7

Link, S. V. (2021). *What makes a Neologism a Success Story? An Empirical Study of the Diffusion of Recent English Blends and German Compounds* [Doctoral dissertation, Ludwig-Maximilians-Universität].

Lomidze, K. (2025). AI in Lexicography for Low-Resource Languages: The Case of the Georgian Language. In T. Margalitadze (Ed.), *Proceedings of the II International Conference Lexicography in the XXI Century* (pp. 49–66). Ilia State University.

Marković, A. & Stanković, R. (2025). So close but still far: Case study on application of LLMs in idioms identification, definition, and generation of illustrative examples. In *Electronic Lexicography in the 21st Century (eLex 2025): Intelligent Lexicography. Proceedings of the eLex 2025 Conference* (pp. 79–94). University of Ljubljana.

Mchedlishvili, K. (2025). Harnessing AI for Creating Definitions of Georgian Military Terminological Neologisms in the context of the Russia-Ukraine War. In T. Margalitadze (Ed.), *Proceedings of the II International Conference Lexicography in the XXI Century* (pp. 80–96). Ilia State University.

McKean, E. & Fitzgerald, W. (2023). The ROI of AI in Lexicography. In P*roceedings of the 16th International Conference of the Asian Association for Lexicography: Lexicography, Artificial Intelligence, and Dictionary Users* (pp. 10-20). Yonsei University. https://doi.org/10.1558/lexi.27569

OpenAI ChatGPT. (2026). *ChatGPT Response to the authors*. Retrieved February-March 2026, from https://chatgpt.com

Pensec, E. & Williams, G. (2025). Where Generative AI Fails: Explaining international standards through an explanatory dictionary. In T. Margalitadze (Ed.), *Proceedings of the II International Conference Lexicography in the XXI Century* (pp. 188–199). Ilia State University.

Picone, M. D. (1996). *Anglicisms, Neologisms and Dynamic French* (Vol. 18). John Benjamins Publishing Company.

Salgado, A. (2026). Neologisms in Lexicography: From Detection to Inclusion Criteria. In H. Nesi & P. Milin (Eds.), *International Encyclopedia of Language and Linguistics*, 3(10) (pp. 605–609). Elsevier.

Tchighladze, K. (2025). Integrating AI into the Term Creation Process: Case Study of English and Georgian Terms of Emerging Technologies. In T. Margalitadze (Ed.), *Book of abstracts of the II International Conference Lexicography in the XXI Century* (pp. 149–151). Ilia State University.

Trap-Jensen, L. (2020). Language-Internal Neologisms and Anglicisms. *Dictionaries: Journal of the Dictionary Society of North America*, 41(1), 11–25. https://dx.doi.org/10.1353/dic.2020.0002

Trap-Jensen, L. (2024). The Best of Two Worlds: Exploring the Synergy between Human Expertise and AI in Lexicography. In T. Margalitadze (Ed.), *Proceedings of the I International Conference Lexicography in the XXI Century* (pp. 26–39). Ilia State University.

Waszink, V. (2020). Neologisms in an Online Portal: The Dutch Neologismenwoordenboek (NW). *Dictionaries: Journal of the Dictionary Society of North America*, 41(1), 27–44. https://dx.doi.org/10.1353/dic.2020.0003.

X Grok. (2026). *Grok Response to the authors*. Retrieved February-March 2026, from https://grok.com/

Zheng, J., Ritter, A. & Xu, W. (2024). NEO-BENCH: Evaluating Robustness of Large Language Models with Neologisms. In L.-W. Ku, A. Martins & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 13885–13906). Association for Computational Linguistics.

Contact information

Tinatin Margalitadze
Centre for Lexicography and Language Technologies, Ilia State University
tinatin.margalitadze@iliauni.edu.ge

Giorgi Meladze
Centre for Lexicography and Language Technologies, Ilia State University
giorgi.meladze.4@iliauni.edu.ge

Ketevan Mchedlishvili
Centre for Lexicography and Language Technologies, Ilia State University
ketevani.mchedlishvili.3@iliauni.edu.ge

Tamar Laluashvili
Centre for Lexicography and Language Technologies, Ilia State University
tamar.laluashvili.1@iliauni.edu.ge

Giorgi Okropiridze
Centre for Lexicography and Language Technologies, Ilia State University
giorgi.okropiridze.4@iliauni.edu.ge

Acknowledgements

The study was conducted as part of a three-year project funded by the Shota Rustaveli National Science Foundation of Georgia, which focused on the comprehensive study of neologisms in Modern Georgian (FR-23-4304).

Lydia Risberg, Kristina Koppel, Margit Langemets,
Hanna Maask, Esta Prangel, Maria Tuulik & Silver Vapper

ARE BEARS MORE OFFENSIVE THAN DOGS WHEN DRUNK?

Using LLMs to Systematize the Labelling of Estonian Synonyms

Abstract This paper investigates the potential of large language models (LLMs) as assistants with corpus analysis in lexicography, focusing on the assignment of register labels in the *EKI Combined Dictionary* (CombiDic). Inconsistencies in register labels become evident across synonym sets in CombiDic, which is the result of different lexicographers working on words at different times. We examine how LLMs from Anthropic, Google, and OpenAI handle the challenge of distinguishing between offensive and colloquial usage, which pose a challenge in lexicographic practice. Using a evaluation dataset of 297 words evaluated by five native Estonian-speaking annotators and three LLMs queried via API, we find that LLMs are quite stable tools for register categorisation and that corpus context significantly improves their performance. Depending on the LLM, 79.8–88.6% of suggestions were deemed acceptable for lexicographic use, although since LLMs may be trained to detect offensive language, they tended to select the OFFENSIVE label more often than human annotators. Nevertheless, Gemini 3.1 Pro performed best overall. The experiment already enabled corrections to CombiDic, demonstrating practical value. While full systematisation has not yet been achieved – the broader goal is labels grounded in usage data – LLMs represent a promising and time-efficient aid in semi-automatic lexicographic workflows.

Keywords Large language models; synonyms; offensive language; register labels; human evaluation dataset; Estonian

1 Introduction

The Estonian general language dictionary *EKI Combined Dictionary* (CombiDic, 2026; via the language portal *Sõnaveeb*, Koppel et al., 2019), compiled at the Estonian Language Institute (EKI), has been continuously edited and updated since 2019, including the addition of synonyms. In 2019, dictionaries that had been stored in separate databases were consolidated into a single database, Ekilex (Tavast et al., 2020, 2023). Within Ekilex, teams of lexicographers who had previously worked on different dictionaries in separate lexical databases began working collaboratively with all types of data (e.g. meanings, collocations, translation equivalents, etc.) in a unified database environment. One of the new layers of information introduced and developed in CombiDic was synonyms. Grouping words with similar meanings into synonym sets (which consists of both full and partial synonyms) in CombiDic has helped uncover inconsistencies in sense division (Leemets et al., 2024, p. 153) and register labels – synonymous words have been worked on at different times and by different lexicographers, many of synonyms had not been reviewed together and/or analysed on the basis of corpus data to reflect current usage. For example, regarding the question in the title of this paper, in January 2026, the word *joomakaru* '(literally:) drinking bear' was labelled as OFFENSIVE in CombiDic, and *joomakoer* '(literally:) drinking dog' as COLLOQUIAL (both mean 'alcoholic') – is this also how they are actually used, or have different lexicographers been working on them at different times?

This raises a broader issue: differing register labels for synonyms have led to confusion among both native Estonian speakers and language learners, highlighting the need for greater consistency across synonym sets. However, the broader goal extends beyond mere uniformity: rather than simply achieving uniform presentation, register labels should be grounded in

corpus data and reflect authentic usage. Some synonyms carry a clearly stronger connotation than others in use; no two words are perfect synonyms – subtle differences in register, connotation, or usage always remain (Cruse, 1986). In CombiDic, full and partial synonyms are distinguished (see Figure 1, where full synonyms are listed under "Synonyms" and partial synonyms under "Similar meaning"). While true full synonymy is rare in natural language, in Ekilex's data model full synonyms share the exact same definition, whereas partial synonyms have similar but not identical meanings (Leemets et al., 2024). Throughout the paper, for brevity, the terms *synonyms* and *synonym sets* refer to both types of synonyms: full and partial.



FIGURE 1 A screenshot of *Sõnaveeb* showing how in CombiDic the word *taignapea* ('a fool', literally 'dough-head') is labelled as RARE and OFFENSIVE, while its synonym *tainapea* (an alternative form for the same meaning; the variation concerns parallel word forms *taigen* and *tainas* 'dough') carries no label at all, and the synonym *tainas* ('a fool') is labelled as INFORMAL (colloquial). These words have been worked on by different lexicographers at different times

Synonym sets in CombiDic vary considerably in size: while some contain only a few members, many include more than ten words, and some clusters exceed one hundred. Bringing these to a more consistent basis manually is therefore extremely time-consuming, as it would require several lexicographers working exclusively on this task. Manual corpus analysis – even with professional corpus query tools like Sketch Engine (Kilgarriff et al., 2014) – compounds this challenge further, since frequent words across large synonym sets must each be examined individually, and results can be somewhat subjective owing to the differing linguistic intuitions of the linguists (Risberg et al., 2025a, p. 613).

In this paper, we describe an experiment in which we investigated LLMs as assistants in corpus analysis, with the goal of optimising dictionary work while preserving the quality and reliability of linguistic information that dictionary users expect (Müller-Spitzer & Koplenig, 2014). Since we were primarily creating a human evaluation dataset for LLMs, this paper does not describe a finished workflow but rather investigates how LLMs from Anthropic, Google,

and OpenAI – previously shown to deal with Estonian[1] and distinguish registers effectively (Risberg et al., 2025b, 2026a) – handle the specific challenge of assigning register labels to Estonian synonyms.

More specifically, we focus on the relationship between offensive and colloquial language, as lexicographers sometimes find it difficult to distinguish between pejorative and colloquial usage – the interpretation of usage contexts is also more subjective than with, for example, strongly marked vulgar words (which lexicographers have less difficulty labelling). We are thus deliberately working with ambiguous material in pursuit of practical lexicographic benefit, as such choices must constantly be made in dictionary work. To address this, we examine to what extent and in what ways LLMs are helpful in identifying registers based on provided corpus data.

2 Experiment design

Since we are working within the CombiDic framework, our understanding of offensive and colloquial language, and their respective usage labels, follows the definitions established there (see Section "Method and evaluation dataset"). More broadly, we follow what is set out in the EKI Handbook:[2] "From a linguistic perspective, words do not inherently carry offensive meaning – they acquire connotation and register only in context. A word becomes offensive when it has been repeatedly used to label or demean people." (Paet, 2022)

Drawing on the approach adopted in our previous studies (Risberg et al., 2025a, 2025b, 2026a), *register* is understood as a scale of situational language use ranging from informal to formal, shaped by social and functional contexts. Rather than fixed categories, registers are seen as a continuum that can be recognized through linguistic markers but not clearly delimited to them (Biber & Conrad, 2009; Biber & Egbert, 2023).

The experiment, conducted between January and May 2026, investigated the potential of LLMs for assigning register labels to synonyms in CombiDic. LLM outputs were compared against the judgements of five human annotators, creating a evaluation dataset to assess which LLM would be most suitable for labelling task.[3]

2.1 Data

Our initial plan was to examine 250 synonym sets (as of January 2026) containing at least one word (its sense) marked with an OFFENSIVE label. On closer inspection, however, it became apparent that the sets contained many recurring words, as partial synonyms carrying an OFFENSIVE label were retrieved from the database as separate sets even when they had already appeared in another word's set. We therefore shifted from a set-based to a word-based approach.

We conducted a series of pilot experiments (which, for brevity, will not be described here), on the basis of which we decided to exclude from the final experiment words that (1) are too rare

1 In early June 2026, EKI launched a public website ranking LLMs on Estonian language and culture benchmarks. Retrieved July 24, 2026, from https://moodupuu.eki.ee/

2 An online publication on contemporary Estonian, including spelling rules and language advice. Retrieved July 24, 2026, from https://teatmik.eki.ee/teatmik/ (only in Estonian)

3 All resources related to the experiment (e.g. prompts, data) are available on GitHub. Retrieved July 24, 2026, from https://github.com/keeleinstituut/EKKD-III1/tree/main/registrid/katse4

(0–99 hits) in the *Estonian National Corpus* 2023 (ENC; Koppel et al., 2023) or too frequent (one million hits or more), (2) have too many senses – a word was allowed to have up to 3 senses[4] in CombiDic, (3) multi-word expressions, as they are harder to retrieve from the corpus than single-word entries (compound words in Estonian are written as one word and were therefore included). Our study thus focused on average, everyday words, with all extremes and complexities excluded. Optimal retrieval of corpus information for such extreme or complex cases remains a separate research question, and for very rare words LLM assistance may not even be necessary.

Our final sample included 25 synonym sets[5], each containing at least 4 words with at least 1 labelled OFFENSIVE, totalling 329 words. After removing duplicate words with the same meaning, 304 words remained. We later excluded 7 words because the human annotators did not reach a consensus on their labels. The final dataset of the analysis comprised 297 words.

2.2 Method and evaluation dataset

The evaluation dataset was created by 5 human annotators, all native speakers of Estonian with varying levels of experience in CombiDic compilation. They did not see each other's responses and were not permitted to consult CombiDic as a reference. The LLMs were accessed via API and likewise had no access to the labels assigned to word senses in CombiDic. We did not vary parameters such as temperature in this experiment, but future research could also compare model performance across different parameter settings. We selected the most recent models available as of early April 2026 from providers previously shown to perform well on Estonian register tasks (Risberg et al., 2025b, 2026a); specifically, we included Claude Opus 4.6 (Anthropic, 2026), Gemini 3.1 Pro (Google, Gemini, 2026), and GPT-5.4 (OpenAI, 2026). For brevity, the LLMs are hereinafter referred to by name only, without specifying the version.

Both humans and LLMs were given a word, its specific meaning as represented in the synonym set, and 10 sentences drawn from the ENC 2023, and were asked to make a concrete decision: based on the sentences provided, would they assign that meaning an OFFENSIVE label, a COLLOQUIAL label, some other label, or consider it neutral – or were the sentences misleading (e.g. used in the wrong sense), making it impossible to reach a decision. Thus, the analysis was carried out on the same grounds. Additionally, the LLMs were assigned the role of an Estonian dictionary compiler and each of their choices was required to be accompanied by a justification of the labelling choice (up to 5 sentences).

This choice between OFFENSIVE and COLLOQUIAL label was necessitated by the specifics of CombiDic, as more than one label of the same kind is not assigned to a single word sense (e.g. both OFFENSIVE and COLLOQUIAL simultaneously), although a register label and a frequency label can co-occur (e.g. OFFENSIVE and RARE). Admittedly, even within informal language there is a spectrum ranging from mildly to strongly offensive words, but the editorial decision in CombiDic has been that if a word already carries an OFFENSIVE label, it would not appear in neutral text anyway, and a COLLOQUIAL label is therefore not added on top of it.

[4] For example, one of the pilot experiment words was the Estonian verb *võtma [kellegi kallal]* 'to bicker', whose synonym set meaning was the 21st sense out of 21, occurring in 6.7 million sentences in the corpus – retrieving that specific sense was practically impossible.

[5] This represents 10% of the originally planned sample. However, the reduced scope was also informed by practical constraints, including API context window limitations, financial and computational costs, and time. Even at this scale, the results are sufficient to draw conclusions about whether further work is warranted and which LLM is best suited for the task.

The prompt included examples, with the instructions specifying when to select which label accompanied by sample words. The instructions stated that the OFFENSIVE label should be selected when the usage of the word in the given sense is disparaging, denigrating, derogatory, abusive, or insulting towards a person or a thing (for example, *ajuhälvik* 'dimwit', *debiilik* 'retard', *inimrämps* 'human scum'), and that the COLLOQUIAL label should be selected when the usage of the word in the given sense belongs to a register that is more informal than formal or neutral language use (for example, *bensukas*, a colloquial shortening of *bensiinijaam* 'petrol station', *bemm*, a colloquial adaptation of the car brand BMW, *piff* 'babe'). A word sense was considered neutral when its usage was not marked in any particular way and it would be appropriate in, for example, a neutral newspaper article, an opinion piece, or a news report.

The decision to use 10 corpus sentences per word was informed by pilot experiments in which 25 sentences showed no notable improvement in annotation quality, and we therefore retained the smaller number – especially since using more sentences would have increased the time required for human annotation. The 10 sentences per word were extracted from the ENC 2023[6] and disambiguated by sense using Claude Haiku 4.5 (Anthropic, 2025). The disambiguation pipeline proceeded in several stages. For each headword, a candidate pool was first retrieved from ENC 2023 via Sketch Engine API. For polysemous lemmas the corpus query was narrowed by a list of included or excluded collocates targeting the intended sense. The retrieved sentences were then reranked against dictionary definition using multilingual sentence embeddings and cosine-similarity pre-filter (Wang et al., 2024). The remaining candidates were passed to previously mentioned Claude Haiku model in batches, where each sentence was labelled as expressing the target sense with certainty of three levels. The final 10 sentences per word were selected from the most certain set of sentences according to cosine similarity or from the middle certainty set if there were not enough candidates as a fallback. Thus, we can be confident that the sentences were not hallucinated but originated from the ENC. Manual sense disambiguation of the same volume would have taken a human annotator several weeks, whereas the automated process was completed in 3–4 hours – a considerable saving in time and effort, though not without errors.

We queried each LLM three times at 24-hour intervals to assess response stability, treating the majority response (at least 2 out of 3) as the model's final answer. To determine whether and to what extent corpus sentences improve LLM performance, we also queried the same words once without providing the 10 corpus sentences, as a point of comparison and to evaluate the zero-shot capabilities of LLM models in this type of task.

3 Results

First, as noted above, 7 words were excluded from the final analysis because the 5 human annotators failed to reach consensus on the appropriate label (*totakas* 'dimwit', *togu* 'dullard', *ilmasammas* '(ironically:) pillar of society', *äpard* 'screw-up', *mooramees* 'Moor', *pugeja* 'bootlicker', *mats* 'yokel'). The common difficulty was that 10 sentences were insufficient to determine whether the word sense was OFFENSIVE or COLLOQUIAL. This was particularly evident with *mooramees* 'Moor' – an outdated and offensive ethnic term – which annotators were unfamiliar with and could not identify as OFFENSIVE from the sentences alone, as none of the provided examples used it in an overtly offensive manner.

[6] ENC contains a diverse range of text types (see Koppel & Kallas, 2022, p. 209). Most of the texts have been web-crawled and represent various genres (including blogs and forums), but the corpus also includes contemporary and older literary works, and academic texts. Therefore, the corpus can be considered sufficiently diverse to capture the offensive and colloquial language use examined in this study.

A second observation, which will not be discussed in detail in this paper, concerns the authenticity of corpus sentences retrieved by Claude Haiku 4.5: for approximately 50 out of 297 words (about 17%), 1 or 2 of the 10 retrieved sentences were dictionary definitions[7] or metalinguistic commentary rather than authentic instances of the word in use. Notably, although Claude Haiku was not assigned the role of a dictionary compiler during sentence retrieval – unlike the LLMs used in the main experiment – it still surfaced dictionary content from the corpus. This relates to a broader issue of circular analysis, whereby LLMs draw on dictionary resources contained within the corpus (in this case, EKI dictionaries), a phenomenon we have encountered in LLM justifications in previous experiments and discussed elsewhere (Risberg et al., 2026b; LLMs' prescriptivism has been studied by others, e.g. Kostadinova & Gardner, 2026; Smith et al., 2026). We will need to analyse the sample sentences more carefully in the future.

Third, while we did ask the LLMs to justify their choices – as this encourages more deliberate reasoning (Henriksson et al., 2024; Yang et al., 2025) – the justifications were not our focus for now, and we only reviewed them selectively in cases where the choices raised questions. Further analysis of the justifications is therefore needed, both in terms of the quality of the Estonian produced and in terms of what changed or remained consistent across the three runs. The confidence levels expressed by the LLMs have also not been examined in detail, even though we asked them to indicate this in the prompt – as expressing confidence likewise encourages more deliberate reasoning (Razghandi et al., 2025). We will analyse this in the future.

In what follows, we focus on LLM response stability; the overlap between LLM and human labelling choices; comparison with CombiDic; and overlap and agreement in synonym sets.

3.1 Stability of LLM answers

Given the widely reported variability of LLM responses (e.g. Piper 2026: 6), we examined the stability of answers across 3 runs (focusing exclusively on responses given with corpus context; see Table 1). The LLMs were accessed via API rather than through chat interfaces, and the task involved categorisation rather than open-ended text generation – both factors likely contributing to high stability, as studies have shown that LLMs perform better on categorisation tasks (Henriksson et al., 2024, p. 310; Davies 2025).

TABLE 1 LLMs' stability across 3 runs with corpus context and Fleiss' κ

LLM	Same response every time	Percentage	Fleiss' κ across 3 runs
Claude	**295 words**	**99.3%**	**0.99**
Gemini	274 words	92.3%	0.92
GPT	278 words	93.6%	0.93

[7] CombiDic is not included in ENC 2023; however, earlier dictionaries whose content has been incorporated into CombiDic in a revised form are included. Therefore, none of the 10 sentences contained the exact content currently available in the continuously updated CombiDic. Since CombiDic is freely available online (although its database, Ekilex, is not), we cannot be certain that LLMs have not encountered its contents during training. At the same time, this information constitutes such a small portion of the overall training data that it is unlikely to have had a substantial impact on the models.

As can be seen from Table 1, LLMs are highly stable (92.3–99.3%). The most stable model was Claude, which changed its response for only 2 words.[8] However, there is no statistically significant difference across the 3 runs within any of the LLMs.[9] At least two of the responses overlapped across all 3 runs for all the models, so in the following analyses, the majority of the 3 responses is considered as the LLM response. Stability alone, however, is insufficient for a lexicographer – the more pertinent question is whether the LLMs also make choices that are suitable for lexicographic use.

3.2 Agreement between LLMs and human annotators

We examined the degree of overlap between LLM responses and human choices. Here, overlap refers to how closely the LLM response matched the majority response of the 5 human annotators (the human evaluation dataset). As can be seen in Table 2, LLMs were not assigning labels randomly: had they been, the expected overlap would have been approximately 20%, given the five possible response options. However, the difference between human and LLM responses is statistically significant. Of the three models, Gemini produced the best results, overlapping with human judgement in 70.7% of the responses when provided with the 10 corpus sentences. Not surprisingly, all LLMs showed greater overlap with human responses when they were shown the corpus context.

A more surprising finding is that Gemini's overlap with human choices without corpus context exceeded that of Claude and GPT even when the latter were provided with corpus sentences (see Table 2). Specifically, Claude with corpus context and Gemini without assigned the same label to 252 words (84.5% overlap). However, their difference in overlap with human responses does not reach statistical significance, meaning that Gemini's performance based on training data alone is not statistically closer to human judgement than Claude's or GPT's when supported by corpus evidence.

TABLE 2 Overlap between LLM responses and human choices

LLM	Overlap *with* corpus context	Overlap *without* corpus context
Claude	63.3%	55.9%
Gemini	**70.7%**	**64.3%**
GPT	60.9%	52.9%

There was substantial agreement in responses between LLMs, both with and without corpus context (Fleiss' $\kappa = 0.76$ and $\kappa = 0.71$, respectively). Nevertheless, the models' responses were statistically significantly different from each other in all cases, with or without corpus context (with Gemini performing statistically significantly better than the others). However, inter-annotator agreement among humans was moderate (Fleiss' $\kappa = 0.5$) and differed statistically

[8] Estonian verb *haukuma* 'to speak shamelessly and sharply to someone or to nag at them' (2 times COLLOQUIAL, 1 time OFFENSIVE) and *klähvima* 'to repeatedly say something sharply, spitefully, or in a nagging manner' (2 times OFFENSIVE, 1 time COLLOQUIAL).

[9] The choice of statistical tests in this study has depended on the data structure. We used chi-square test for comparing independent groups of categorical observations, McNemar's test for paired comparisons of categorical observations, Spearman's rank correlation for evaluating monotonic relationship between the variables and bootstrap resampling for comparing nested subsets for which standard statistical tests were not applicable. For all the statistical analyses we used $\alpha = 0.05$ as the level of significance.

significantly from every LLM variant's agreement[10] – a finding particularly relevant for assessing the reliability of LLMs in lexicographic workflows. In our previous experiment (Risberg et al., 2025b), where LLMs distinguished informal from neutral language, LLM–human overlap was approximately 75–76%, while human inter-annotator agreement was considerably lower (Fleiss' $\kappa = 0.20$), reflecting the inherent subjectivity of informal/colloquial language. The present experiment yielded better inter-annotator agreement among humans, though this does not necessarily indicate that distinguishing offensive from colloquial language is a less subjective task. These results raise a provocative question: should humans serve as the (gold) standard or a benchmark[11] for LLMs – or could it be the other way around?

This question gains further weight given that the 60.9–70.7% LLM–human overlap does not tell the full story, as differing suggestions may also hold value for lexicographers. In cases where human opinions diverged – even where a majority response could be determined – LLM choices sometimes offered a useful alternative perspective. For instance, while the human majority labelled *püstihull* 'stark raving mad' as neutral, all LLMs classified it as COLLOQUIAL, which seems defensible given that the word would be unlikely to appear in a neutral newspaper article. Conversely, in cases like *tohlakas* 'stupid', where human opinions were split between COLLOQUIAL and OFFENSIVE, the unanimous LLM verdict of OFFENSIVE might warrant adding a note to the definition in CombiDic.

To account for this, we assessed whether the models' responses that differed from those of the human annotators could nevertheless be considered acceptable. Combining exact matches with cases in which human variation indicated that either label could be legitimate, we found that, with corpus context, 88.6% of Gemini's, 81.5% of Claude's, and 79.8% of GPT's choices would be suitable for inclusion in CombiDic. Without corpus context, the figures were 82.2%, 73.4%, and 70.7%, respectively, with statistically significant differences between the two conditions for all LLMs, confirming that corpus context improves their performance. This evaluation approach is also consistent with previous work arguing that annotation should be viewed as a means of capturing the diversity of human judgment, which is essential both for studying complex cultural phenomena and for evaluating LLMs. In this view, a deterministic (plausibility-based) annotation approach focuses on excluding clearly implausible responses instead of seeking consensus or a single correct answer, making it well suited to tasks where the goal is to assess whether a model can produce an acceptable interpretation (cf. Piper, 2026, p. 5).

3.3 Labelling choices

Overlap and inter-annotator agreement might have been even higher had the task permitted the simultaneous selection of two labels (e.g. COLLOQUIAL and OFFENSIVE) which was not possible due to CombiDic's compiling conventions. Another feature characteristic of CombiDic is that labels are assigned sparingly, rather than in excess. Examining the divergent responses of LLMs and human annotators, we noticed that LLMs tended to default to the OFFENSIVE label whenever the sentences contained even a hint of that connotation (e.g. *jõhkard* 'brute', *julmur* 'cruel person', *suli* 'crook'), whereas in CombiDic one would not label so many words as

[10] This result is consistent with some other studies, which have also found that inter-coder agreement tends to be lower than expected in complex annotation tasks and that, in some cases, LLMs have demonstrated consistency comparable to or exceeding that of human expert coders (see Zupic, 2026, p. 20).

[11] Our original plan was to create a benchmark dataset and the methodology for evaluating the models. However, it later became clear that the human annotators' responses did not converge sufficiently to establish a reliable gold standard. This is partly because, in CombiDic, one must choose between the COLLOQUIAL and OFFENSIVE labels rather than assigning both simultaneously.

offensive, given that they can also be used neutrally or with only milder stylistic colouring (colloquial). We therefore examined further the frequency of each selected label by humans and LLMs (see Table 3).

TABLE 3 The percentage of label chosen by humans and LLMs

Choice	Human annotators (with corpus)	Claude		Gemini		GPT	
		with corpus	without	with corpus	without	with corpus	without
neutral	**39.1%**	21.6%	16.8%	30.6%	25.9%	21.6%	15.5%
COLLOQUIAL	25.9%	24.2%	17.5%	16.2%	15.5%	16.8%	14.8%
OFFENSIVE	30.0%	**51.9%**	**61.3%**	**51.2%**	**56.2%**	**56.9%**	**55.6%**
some other label	1.0%	2.4%	4.4%	2.0%	2.4%	4.7%	14.1%
wrong sentences	4.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%

As can be seen from Table 3, human annotators most frequently selected the neutral option (39.1%) and chose the COLLOQUIAL and OFFENSIVE labels at similar rates (25.9% and 30.0%, respectively), whereas all LLMs – both with and without corpus context – selected the OFFENSIVE label in the majority of cases (51.2–61.3%), and neutral and COLLOQUIAL considerably less often. Looking more closely, when humans selected the OFFENSIVE label, LLMs agreed in most cases (92.1–97.8%), for example on words such as *ahvinägu* 'monkey face', *ilalõug* 'slobber-jaw', and *ajukääbik* 'brain dwarf'. Notably, when humans selected the neutral label, the LLMs (with and without corpus context) statistically significantly differed from each other in their agreement with human annotators. Gemini's choices aligned with them in 75.0% of cases with corpus context and 64.7% without. For Claude, the same agreement with humans was 53.4% with corpus, 43.1% without; and for GPT 54.3% and 38.8%, respectively. For other labels, this difference between models is not statistically significant. Furthermore, Gemini selected the neutral option more frequently than the other two LLMs, which likely accounts in part for its higher overall overlap with human responses (see Table 2). With GPT, it became apparent that certain vulgar words, such as *perselakkuja* 'arse-licker' and *hoorapoeg* 'son of a bitch', appear to have been added to a content blacklist, as without corpus context GPT declined to make a choice for those words at all.

None of the LLMs selected "wrong sentences", whereas human annotators did so for 12 words (4.0%). Closer inspection revealed that despite explicit prompt instructions specifying when an option should be chosen, LLMs were creative. This suggests that LLM decision-making does not always align with the intended interpretation of the instructions. A more thorough analysis of the justifications in future work will yield further insight into this. For now, one illustrative case is the word *muld*, for which all LLMs selected "some other label" rather than "wrong sentences" – the option chosen by human annotators, since the 10 corpus sentences represented the word's literal meaning 'fertile soil' rather than its intended synonym set sense 'old person'. Reading through the justifications revealed that the LLMs had in fact recognised the mismatch but interpreted it differently: Gemini characterised the language in the sentences as archaic and poetic and assigned a corresponding label, while Claude noted the absence of suitable corpus evidence for the intended sense but still didn't select "wrong sentences".

We were particularly interested in cases where one annotator considered a word neutral and another OFFENSIVE and therefore examined this variation across three dimensions: (1) within an LLM's own 3 runs, (2) among humans, and (3) between LLMs and humans. Within-LLM variation was rare: Claude had none, while GPT and Gemini each showed this disagreement on 3 words. Such divergences occurred most frequently among humans (52 words). Comparing

humans and LLMs, there were circumstances where human annotators chose "neutral" while LLMs chose OFFENSIVE (when shown the corpus context: Gemini for 19 words, Claude for 28, GPT for 35), but no reverse cases existed. There was only one word that varied like this for a LLM, specifically Gemini, and also caused disagreement between humans: *neeger* 'the N-word', an internationally sensitive term. Among the five human annotators, 3 selected OFFENSIVE, 1 selected COLLOQUIAL, and 1 selected neutral. Notably, two annotators commented that the word had not been used offensively in the provided sentences, and one explicitly stated that their OFFENSIVE label was based on general knowledge rather than the examples. This points to a broader problem: the 10 sentences retrieved by Claude Haiku were formally correct in sense but unrepresentative of actual usage, being largely neutral in character. The issue thus arose from both the limited number of sentences and the requirement to base decisions solely on the provided sentences. In actual dictionary work, such constraints as those imposed in this experiment usually do not exist.

The inconsistency referenced in the paper's title – that bears are more offensive than dogs when drunk – was resolved as a result of this experiment. In January 2026, *joomakaru* 'drinking bear' was labelled OFFENSIVE and *joomakoer* 'drinking dog' COLLOQUIAL (both mean 'alcoholic'). As it turned out, the relevant cases did not make it into the final sample due to being too rare (11 and 6 sentences respectively in the ENC 2023), so we analysed them separately. Analysis of corpus sentences led both to being COLLOQUIAL companions. Interestingly, and unlike what CombiDic showed in January, some annotators noted that "the word *koer* 'dog' seems somehow more derogatory than *karu* 'bear'", reflecting that *koer* can be used derogatorily to refer to a person, whereas *karu* carries a more positive connotation. Both annotators who commented nonetheless selected COLLOQUIAL based on corpus evidence. Claude and Gemini likewise found *joomakaru* COLLOQUIAL, while *joomakoer* was seen more clearly as OFFENSIVE in corpus sentences "because *joomakoer* never appears in sentences with a warm, playful, or affectionate tone." GPT considered both words OFFENSIVE – consistent with GPT's broader tendency to assign OFFENSIVE more readily.

Overall, depending on the model, at least 79.8–88.6% of LLM suggestions would be acceptable to a lexicographer (this means that the initial consensus opinion had been formed from the decisions of 2–3 annotators, but the other 3–2 had chosen a different label; additionally, some had added comments expressing their uncertainty), though human validation of LLM output remains necessary. The approximately 11–20% of LLM choices that a lexicographer would not accept (cf. section "Agreement between LLMs and human annotators") appear to stem from two tendencies: first, LLMs assigned the OFFENSIVE label in more than half of all cases, compared to only nearly one third among human annotators (Table 3); second, LLMs always attempted to make a choice, whereas the human majority response for 12 words (4.0%) was that the sentences did not represent the intended sense and that no decision could be reached.[12] These findings suggest that future prompts might need to include more precise instructions regarding label selection and commands that would compel LLMs to follow instructions more closely.

3.4 Labelling choices compared to CombiDic

As seen in Table 4, human choices showed the greatest mutual overlap with CombiDic (62.0%), as humans similarly selected neutral and options other than OFFENSIVE more frequently.

[12] We examined whether word frequency and the human majority response of "wrong sentences" are related. They are not – Spearman's correlation coefficient is 0.08, but the correlation is not statistically significant ($p = 0.12$).

Both human annotators' and LLMs' responses differed statistically significantly from CombiDic labels (as in February 2026). In any case, the experiment proved useful for lexicographers, as words where human or LLM choices did not match the existing CombiDic label are worth reviewing – to determine whether a label should be added or a clarification included in the definition, such as "sometimes used affectionately" (e.g. for *tobuke*, a diminutive form of *tobu* 'stupid').

TABLE 4 Overlap of labelling choices with labels in CombiDic (February 2026)

	Overlap with corpus	Overlap without corpus
Human annotators	**62.0%**	–
Claude	42.8%	37.7%
Gemini	51.9%	45.5%
GPT	42.1%	34.7%

Many words in our sample were labelled neutral in CombiDic (178 words, 59.9%), fewer carried the COLLOQUIAL label (64 words, 21.6%) or the OFFENSIVE label (45 words, 15.2%), with only a small number of other labels (RARE, OLD-FASHIONED, VULGAR combined: 10 words, 3.4%). Thus, CombiDic contains an even higher proportion of neutral word senses than human annotators selected. Human annotators assigned the COLLOQUIAL label at a rate like CombiDic (77 words, 25.9%) but marked fewer words as neutral (116 words, 39.1%) and more as OFFENSIVE (89 words, 30.0%). The predominance of neutral entries reflects a deliberate decision by the compilers of CombiDic to leave word senses unlabelled where possible, avoiding unnecessary stigmatisation. Based on corpus sentences, however, more labels could potentially be added – though this raises a more philosophical lexicographic question: do we want to label every word sense, especially when their usage is ambivalent?

Nevertheless, this indicates that the LLMs showed greater mutual agreement precisely because they all tended to default to the OFFENSIVE label. Human choices, meanwhile, overlapped more with the labels already present in CombiDic (62.0%), as the annotators might have been aware of them from previous work. Individual human annotators' overlap with the human majority response ranged from 71.4% to 89.9%. The lowest overlap was found for the annotator with the least experience in CombiDic editing, while the highest belonged to a member of the synonym working group.

Another possibility is that human annotators showed considerable variation on our sample words, because many words appear mild in form yet are used to portray others negatively. LLMs, by contrast, tended to select OFFENSIVE, possibly because many given definitions contained negative connotations. One further aspect to consider is that Eicher & Irgolič (2024; Aasa et al., 2026 on newer models) found that some LLMs tend to select the first object in a list. This could partially explain why the LLMs in our experiment chose the OFFENSIVE label, which was the first in the list, more often than human annotators did. However, a more likely explanation is that LLMs may be trained to be cautious with derogatory language, which may have led them to favour the OFFENSIVE label. Offensive and harmful content and hate speech recognition are addressed during LLMs' pre-training and fine-tuning (the ability and capacity to detect such language has been studied, e.g. Zhang et al., 2025; Proskurina et al., 2026). While both factors may have contributed to the observed pattern, the latter explanation seems more plausible given our experimental setup; future work examining the LLMs' justifications may provide a clearer answer.

3.5 Synonym sets

We examined how well LLM choices (both with and without corpus context) aligned with human choices at the synonym set level. Great overlap with human choices across all LLMs (average of at least 80.0%) was found in 6 synonym sets (out of 25 sets, 24.0%), including one case with 100% overlap (synonyms expressing clearly offensive ways to insult someone, such as *ahvinägu* 'monkey face'). The overall worst overlap (average below 13.0%) was in the set expressing personal failure or incompetence, such as *hädavares* 'hapless wretch', *käpard* 'bungler', *kobakäpp* 'clumsy oaf'. Closer inspection revealed why this set showed such low results. For the word *atu* in the sense 'a person who fails at everything', all human annotators selected "wrong sentences", while LLM choices varied considerably: with corpus context, models selected COLLOQUIAL, "some other label", and neutral, whereas without corpus context all LLMs selected OFFENSIVE – upon examining the justifications, it becomes apparent that this is because the provided definition itself carries a disparaging connotation. For *kobakäpp*, human annotators selected neutral (3 times) and COLLOQUIAL (twice), while all LLMs selected OFFENSIVE both with and without corpus context. For *käpard*, human opinions were divided across neutral, COLLOQUIAL, and OFFENSIVE, while all LLMs again selected OFFENSIVE.

Second, we investigated agreement among human annotators and stability within LLMs using Fleiss' κ. Humans showed the greatest mutual agreement on 3 synonym sets: synonyms for a farmer (mostly neutral), synonyms for a bureaucrat (mostly neutral), and synonyms referring to Black people (mostly racial slurs which were labelled OFFENSIVE). LLMs showed the greatest mutual consistency on 2 synonym sets, one of which overlapped with the humans – neutral synonyms for a bureaucrat – while the other concerned words for cruel people (e.g. *julmur* and *jõhkard* 'a cruel and brutal person'). It thus appears that making consistent labelling choices across synonym sets through collaboration between LLMs and human annotators (lexicographers) requires a bit clearer guidelines in the future.

4 Concluding discussion

Are LLMs the solution to resolving register label inconsistencies among synonyms? Our experiment showed that LLMs queried via API are stable tools for categorisation tasks (response stability above 90.0%). Of the three models, Gemini 3.1 Pro overlapped most closely with human annotators' choices, with up to 88.6% of its labelling choices usable in CombiDic when borderline cases are included. All models performed statistically significantly better when provided with corpus context, and as few as 10 corpus sentences may suffice – a finding that holds for human annotators as well, several of whom commented that they would have reached a different decision without corpus context. A further argument in favour of LLM use is that when LLM label choices diverged from the human majority response, human opinions on those same words also tended to diverge, suggesting that LLM uncertainty reflects genuine linguistic ambiguity rather than model error.

LLMs also brought meaningful time savings, for example in retrieving sense-disambiguated sentences from the corpus. Thus, although LLM-generated responses must be verified before use in actual lexicographic work, LLMs offer a promising tool for processing large amounts of data. That said, API context windows impose practical limitations, and piloting and running experiments incur financial and computational costs. The experiment also focused on average, moderately polysemous words – rare, highly frequent, and highly polysemous cases remain additional analytical challenges.

Can we say, then, that we used LLMs to systematise the labelling of Estonian synonyms? Not in a fully comprehensive sense, but several significant steps have been taken, and inconsistencies in CombiDic have already been corrected based on this experiment. Regarding the *taignapea* synonym set from Figure 1, both human annotators and all LLMs chose the OFFENSIVE label for all three words, which had previously carried inconsistent labels. Second, alongside other analyses, the example of *joomakaru* and *joomakoer* ('alcoholic') above illustrates Gemini's utility for borderline labelling decisions, as it appears to approach corpus sentences with less bias than GPT.

However, can this approach be applied to register inconsistencies beyond offensive language? While the present experiment was designed around the challenging OFFENSIVE–COLLOQUIAL boundary, the underlying methodology – LLM-assisted corpus analysis evaluated against human annotation – is not specific to this label type. The same workflow could in principle be extended to other register distinctions (e.g. OLD-FASHIONED, FOLK), provided that appropriate annotation guidelines and corpus resources are available. The focus on offensive language was motivated by the practical problem of inconsistency in CombiDic, not by a theoretical constraint of the method.

Dictionaries are widely regarded as authoritative references (Müller-Spitzer & Koplenig, 2014; Kosem et al., 2019; Arias-Arias et al., 2025), as confirmed by a 2023 user experience survey of CombiDic (Langemets et al., 2024). Usage labels are generally sought-after information, but it remains an open question whether labelling all word senses is needed, given that some exhibit genuinely ambivalent usage. Some users treat labelled words as ones to be avoided under any circumstances – expecting a decisive, even if subjective, answer from lexicographers. LLMs may help reduce this subjectivity, grounding decisions in usage data rather than individual judgement. For borderline cases, however, where either label would be acceptable, no tool – human or automated – can fully resolve the ambiguity, since language does not operate within fixed boundaries and there are no watertight categories (Rundell, 2002). In such cases, CombiDic compilers might consider assigning both OFFENSIVE and COLLOQUIAL labels or including a note such as "sometimes offensively" in the definition.

The longer-term goal initiated by this experiment is a human-in-the-loop workflow, in which an LLM proposes label candidates for synonym sets that a lexicographer then validates, enabling the semi-automatic review of CombiDic based on corpus data. The aim is to ensure that when synonyms carry different labels, this reflects genuine differences in usage rather than simply being an artefact of different lexicographers working on the words at different times. Making consistent labelling choices across synonym sets through collaboration between LLMs and lexicographers requires clearer guidelines. However, even the present experiment demonstrates that this is a practical and worthwhile direction.

This returns us to the question posed above: should humans serve as the (gold) standard or a benchmark for LLMs, or could it be the other way around? For a fully prescriptive dictionary, this may be a less pressing question. For a mostly descriptive dictionary such as CombiDic,[13] when human annotators themselves disagree on the appropriate label, it represents a genuine challenge. Moreover, dictionaries carry significant social authority, and users often expect lexicographers to make the difficult calls, so they don't have to. But if human judgement – even that of language experts – is variable and inconsistent, are lexicographers' decisions more

[13] CombiDic contains approximately 2,500 usage notes, which are intended to support informed choices rather than function as strict guidelines or prohibitions; they reflect tendencies that emerge from patterns of language use.

reliable than those of LLMs? The question remains open: human inconsistency may itself be a virtue where LLM uniformity falls short, yet LLMs may equally help reduce human subjectivity.

References

Aasa, T., Pilvik, M-L. & Lindström, L. (2026). Intensiivsussõnade tugevus Eesti keeles: inimkõnelejate ja keelemudelite hinnangute võrdlus. Eesti Rakenduslingvstika Ühingu kevadkonverents, 24 April. Retrieved July 24, 2026, from https://www.youtube.com/watch?v=YrJb2ta8pfI&t=23404

Anthropic (2025). Claude Haiku 4.5. Retrieved July 24, 2026, from https://www.anthropic.com

Anthropic (2026). Claude Opus 4.6. Retrieved July 24, 2026, from https://www.anthropic.com/news/claude-opus-4-6

Arias-Arias, I., Domínguez Vázquez, M. J. & Valcárcel Riveiro, C. (2025). People, machines and dictionaries: Is artificial intelligence killing dictionaries? In I. Dias, R. H. Gouws, A. Lobenstein-Reichmann & S. J. Schierholz (Eds.), *Regarding the Significance of Lexicography/Zur Bedeutsamkeit der Lexikographie*. De Gruyter Brill. https://doi.org/10.1515/9783112216644-012

Biber, D. & Conrad, S. (2009). *Register, Genre, and Style* (Cambridge Textbooks in Linguistics). Cambridge: Cambridge University Press.

Biber, D. & Egbert, J. (2023). What is a register? Accounting for linguistic and situational variation within – and outside of – textual varieties. *Register Studies*, 5. Retrieved July 24, 2026, from https://www.jbe-platform.com/content/journals/10.1075/rs.00004.bib

Cruse, D. A. (1986). *Lexical Semantics*. Cambridge University Press.

Davies, M. (2025). Integrating AI/LLMs into English-Corpora.org. In *English-Corpora.org*. Retrieved July 24, 2026, from https://www.english-corpora.org/ai-llms

Eicher, J. E. & Irgolič, R. F. (2024). Reducing Selection Bias in Large Language Models. *arXiv.org*. https://doi.org/10.48550/arXiv.2402.01740

EKI Combined Dictionary/CombiDic 2026 = *EKI ühendsõnastik 2026*. Sõnaveeb. Estonian Language Institute. Retrieved July 24, 2026, from https://sonaveeb.ee/

Google, Gemini (2026). Gemini 3.1 Pro. Retrieved July 24, 2026, from https://deepmind.google/models/model-cards/gemini-3-1-pro/

Henriksson, E., Myntti, A., Hellström, S., Erten-Johansson, S., Eskelinen, A., Repo, L. & Laippala, V. (2024). From Discrete to Continuous Classes: A Situational Analysis of Multilingual Web Registers with LLM Annotations. In *Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities* (pp. 308–318).

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography* 1(1), pp. 7–36.

Koppel, K., Tavast, A., Langemets, M., & Kallas, J. (2019). Aggregating dictionaries into the language portal Sõnaveeb: issues with and without a solution. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek, & C. Tiberius (Eds.), *Proceedings of the eLex 2019 conference* (pp. 434–452). Lexical Computing CZ, s.r.o.

Koppel, K., Kallas, J., Jürviste, M. & Kaljumäe, H. (2023). *Estonian National Corpus 2023*. Lexical Computing Ltd./Estonian Language Institute. https://doi.org/10.15155/3-00-0000-0000-0000-08C04M

Kosem, I., Lew, R., Müller-Spitzer, C., Silveira, M.R., Wolfer, S., Dorn, A., Gurrutxaga, A., Ceberio, K., Etxeberria, E., Lefer, M.A., Geeraerts, D., Despot, K.Š., Stojanov, T., Ljubešić, N., Škrabal, M., Štěpánková, B., Vodrážková, V., Lorentzen, H., Trap-Jensen, L., Kallas, J . . . . Nesi, H. (2019). The Image of the Monolingual Dictionary Across Europe. Results of the European Survey of Dictionary use and Culture. *International Journal of Lexicography* 32(1), pp. 92–114. https://doi.org/10.1093/ijl/ecy022

Kostadinova, V. & Gardner, M. H. (2026). How prescriptive are Large Language Models? In *Book of abstracts of the 8th Prescriptivism Conference, 24-26 June, Brussel*, p. 18. Retrieved July 24, 2026,

from https://www.historicalsociolinguistics.be/wp-content/uploads/2026/06/Prescriptivism-Conference-book-of-abstracts-1.pdf

Langemets, M., Risberg, L. & Algvere, K. (2024). To Dream or Not to Dream About 'Correct' Meanings? Insights into the User Experience Survey. In *XXI EURALEX International Congress* (pp. 741–760). Cavtat, Croatia.

Leemets, T., Koppel, K. & Tuulik, M. (2024). Semantic Challenges of Bidirectional Synonymy in the EKI Combined Dictionary. In: Kristina Š. Despot, Ana Ostroški Anić, Ivana Brač (Eds.), *XXI EURALEX International Congress: Lexicography and Semantics. Book of Abstracts: Cavtat, Croatia, 8–12 October 2024* (pp. 152–154). Zagreb: Institut za hrvatski jezik.

Müller-Spitzer, C., & Koplenig, A. (2014). Online dictionaries: expectations and demands. In C. Müller-Spitzer (Ed.), *Using Online Dictionaries*. (Lexicographica. Series Maior 145.) (pp. 143–188). Walter de Gruyter.

OpenAI (2026). Introducing GPT-5.4. Retrieved July 24, 2026, from https://openai.com/index/introducing-gpt-5-4/

Paet, T. (2022). Solvavast keelekasutusest ['On the offensive language']. In *EKI teatmik* [EKI Handbook]. Retrieved July 24, 2026, from https://eki.ee/teatmik/solvavast-keelekasutusest/

Piper, A. (2026). A Theory-First Approach Towards Using Generative AI for Humanities Research. *Computatuional Humanities Research* (pp. 1–12). Retrieved July 24, 2026, from https://www.cambridge.org/core/journals/computational-humanities-research/article/theoryfirst-approach-towards-using-generative-ai-for-humanities-research/1D3B11E220735772BE68B5D6C83DB2B8

Proskurina, I., Carpentier, M., & Velcin, J. (2026). HatePrototypes: Interpretable and Transferable Representations for Implicit and Explicit Hate Speech Detection. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)* (pp. 4387–4399). European Language Resources Association (ELRA). https://doi.org/10.63317/3opu4zq9p6pc

Razghandi, A., Hosseini, S. M. H. & Baghshah, M. S. (2025). CER: Confidence Enhanced Reasoning in LLMs. *arXiv.org*. https://doi.org/10.48550/arXiv.2502.14634

Risberg, L., Tuulik, M., Langemets, M., Koppel, K., Vainik, E., Prangel, E. & Aedmaa, E. (2025a). Keelekorpus kui leksikograafi abiline kõnekeelsuse tuvastamisel. Keel ja Kirjandus, 7, pp. 605–624. https://doi.org/10.54013/kk811a3

Risberg, L., Aedmaa, E., Tuulik, M., Langemets, M., Vainik, E., Prangel, E., Koppel, K. & Pook, H. (2025b). The role of subjectivity in lexicography: Experiments towards data-driven labeling of informality. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda & A. Salgado (Ed.). *Electronic lexicography in the 21st century (eLex 2025): Intelligent lexicography. Proceedings of the eLex 2025 conference. Bled, 18–20 November 2025* (pp. 336–356). Bled: Lexical Computing CZ s.r.o.

Risberg, L., Aedmaa, E., Pook, H., Koppel, K., Tuulik, M., Prangel, E. & Langemets, M. (2026a). Korpuse kontekstiga või ilma? Suurte keelemudelite võimekus tuvastada ja märgendiga varustada eesti keele sõnatähendusi. *Eesti Rakenduslingvistika Ühingu aastaraamat = Estonian papers in applied linguistics*, 22, pp. 163–182. https://doi.org/10.5128/ERYa22.10

Risberg, L., Aedmaa, E., Pook, H., Koppel, K., Tuulik, M., Prangel, E., Langemets, M. & Vapper, S. (2026b). Suur keelemudel, sinu sädelev hoiakutevaba sõber? Objektiivse keeleanalüüsi võimalikkusest. At *Eesti humanitaarteaduste aastakonverents, 8.–10. aprill 2026*. Retrieved July 24, 2026, from https://confenta.com/ee/ehak/programm/

Rundell, M. (2002). Good old-fashioned lexicography: Human judgement and the limits of automation. In M. -H. Corréard (Ed.), *Lexicography and Natural Language Processing: A Festschrift in honour of B. T. S. Atkins*. (pp. 138–155). Stuttgart: EURALEX.

Smith, G., Fleisig, E., Rustagi, I. & Yin, X. (2026). Standard Language Ideology in AI-Generated Language. In *arXiv:2406.08726 [cs.CL]*. https://doi.org/10.48550/arXiv.2406.08726

Zhang, C., Zhu, C., Xiong, J., Xu, X., Li, L., Liu, Y. & Lu, Z. (2025). Guardians and Offenders: A Survey on Harmful Content Generation and Safety Mitigation of LLM. *arXiv.org*. https://doi.org/10.48550/arXiv.2508.05775

Zupic, I. (2026). A Human-Centered Workflow for Using Large Language Models in Content Analysis. *arXiv.org*. Retrieved July 24, 2026, from https://arxiv.org/pdf/2603.19271

Tavast, A., Koppel, K., Langemets, M. & Kallas, J. (2020). Towards the superdictionary: layers, tools and unidirectional meaning relations. In Z. Gavriilidou, M. Mitsiaki, & A. Fliatouras (Eds.), *Proceedings of XIX EURALEX Congress: Lexicography for Inclusion* (pp. 215–223). Democritus University of Thrace.

Tavast, A., Koppel, K., Langemets, M., Vapper, S. & Jürviste, M. (2023). Invisible meaning relations for representing near equivalents. In: M. Medveď, M. Měchura, C. Tiberius, K. Iztok, J. Kallas & M. Jakubíček (Ed.), *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference* (pp. 104–117). Brno, Czech Republic: Brno: Lexical Computing CZ, s.r.o.

Wang, L., Yang, N., Huang, X., Yang, L., Majumder, R. & Wei, F. (2024). Multilingual E5 Text Embeddings: A Technical Report. *arXiv.org*. https://doi.org/10.48550/arXiv.2402.05672

Yang, J., Lin, K. & Yu, X. (2025). Think When You Need: Self-Adaptive Chain-of-Thought Learning. *arXiv.org*. https://doi.org/10.48550/arXiv.2504.03234

Contact information

Lydia Risberg
Estonian Language Institute, University of Tartu
lydia.risberg@eki.ee

Kristina Koppel
Estonian Language Institute
kristina.koppel@eki.ee

Margit Langemets
Estonian Language Institute
margit.langemets@eki.ee

Hanna Maask
Estonian Language Institute
hanna.maask@eki.ee

Esta Prangel
Estonian Language Institute
esta.prangel@eki.ee

Maria Tuulik
Estonian Language Institute
maria.tuulik@eki.ee

Silver Vapper
Estonian Language Institute
silver.vapper@eki.ee

Acknowledgements

This paper was funded by R&D project “Applying large language models to lexicography: new opportunities and challenges” (project executor: Estonian Language Institute; funder: Estonian Ministry of Education and Research).

Ines Röhrer & Manuel Raaf

REFLECTIONS ON THE SEMANTIC UNDERSTANDING OF GENERATIVE AI

From LLM Experiments to Lexicographical Tools

Abstract This article presents an application evaluation of the use of large language models (LLMs) for the semantic classification of lexicographical content in three German dialect dictionary projects. Building on earlier experiments with LLM-based semantic classification that showed hit rates of over 80%, the study investigates whether generative AI can support the assignment of meanings to controlled semantic categories while preserving editorial oversight. Evaluation data from user ratings, surveys, and interviews indicate that the tool produces useful classifications in a majority of cases, reduces manual research effort, and contributes to greater consistency in semantic annotation. At the same time, the results reveal recurring limitations, especially in cases involving semantic granularity, context-dependent meanings, and definitions where modifiers are overemphasized. The paper argues that LLMs are most valuable not as replacements for expert lexicographical work, but as components of hybrid workflows that combine automated suggestion with human validation.

Keywords Lexicography; large language models; semantic classification; editorial workflows

1 Introduction

1.1 Motivation

Across the humanities, projects at all stages are currently exploring the most productive, cost-effective, ethically sound, and scientifically accepted ways of integrating generative language models (Large Language Models, LLMs) into their workflows and technical infrastructures. Lexicographic projects share many concerns and requirements with other disciplines engaged in this process. At the same time, their research practices are highly specific and often differ substantially between projects, so that they require their own custom starting points for LLM integration.

A common feature of most lexicographic projects is the use or provision of semantic classifications for their content. The present paper examines the use of LLMs for the automated assignment of meanings to subject groups, using a method tested in a productive lexicographical environment. Recent studies show that LLMs can compete with classical methods in zero- and few-shot scenarios for semantic classification tasks (cf.: Liu et al., 2024; Wang et al., 2024). In lexicography, they are increasingly being discussed as tools for annotation and editorial support (cf.: Lew, 2024; Stöckle et al., 2025).

1.2 Setting

The starting point is an evaluation study on the suitability of different models for automated subject group assignment (cf. Raaf & Röhrer, 2025a). Based on these experiments, a web service was developed and integrated into the workflows of three dialect dictionaries of the Bavarian Academy of Sciences: the Bavarian Dictionary (cf. Schnabel et al., 2021), the Dialectological Information System of Bavarian Swabia (cf.: Schwarz et al., 2021; Schwarz, 2022) and the Franconian Dictionary (cf. König et al., 2021). The service is named *PostKI* in reference to Rudolf

Post's work (1998). The aim is to investigate whether a hybrid workflow combining automatic suggestions and editorial control leads to more efficient and, at the same time, more consistent classification results. The classification is based on subject groups by Rudolf Post, a controlled semantic system based on the conceptual system of Hallig and von Wartburg (1963), utilised by several dialect dictionary projects, e.g. the Franconian Dictionary, the Dialectological Information System of Bavarian Swabia, the Dictionary of the Palatinate, and the Swiss Idioticon (Raaf, 2025, p. 189).[1] The manual assignment of meanings is time-consuming and becomes unfeasible for large quantities of data. It is also prone to inconsistent decisions, making automated support an obvious choice for this task.

While our findings are based on a very specific lexicographic tool implemented for three German dialect dictionaries, many of the conclusions drawn can be helpful for the exploratory assessment of the integration of large language models in a wide range of different projects. We examine how an LLM-based classification service can be integrated into dialect-lexicographical workflows, how editors evaluate its suggestions, and which limitations emerge in productive use.

2 From experiment to tool

2.1 Implementation

Since the evaluation study from 2025 produced promising results and a tool with this functionality appeared to be a helpful resource for the lexicographic workflow, an AI-based service for automated subject group assignment has been developed. The tool is implemented as a web application in which the user supplies at least the meaning, optionally supplemented by an example of usage and the lemma that could provide the LLM with more information. It can be used via graphical user interface or PostKI's own application programming interface (API).[2] The user can choose from a list of subscription-based models, e.g. *Claude Sonnet 4, Mistral Large 3, ChatGPT 4.1, or Llama 3.3 70B*, which are accessed via API of the particular provider. For doing so, an API key for the provider is needed, too. Due to the open character of PostKI, any provider and language model could be added in a configuration file. After submission of the mandatory API key and the meaning, two prompts are being sent to the AI-service in the back-end of PostKI.[3] First, the AI is requested to generate ten sentences based on the given meaning. These are then dispatched back to the AI in another, new request, along with the list of all subject categories and instructions to assign up to three appropriate categories. In rudimentary tests, we noticed better results in splitting the process into two single prompts rather than prompting all in one. The response, i.e. up to three entries of Post's list of subject groups ideally, is subsequently returned to the user. To facilitate the evaluation of each mapping, the user can rate the entirety of the generated answers on a scale from one to five stars, where one star means "very bad" and five stands for "very good". These ratings are stored and can later be used for the systematic evaluation of model performance (see Fig. 2 on page 5). They also provide a practical basis for identifying recurring error patterns and differences between models.

For the editorial process, it appeared more helpful not to limit the tool to a single perceived "correct" subject group, especially in edge cases where a meaning is not clear-cut and may fit multiple subject groups. The suggestions are subsequently reviewed by the editor, who can

[1] For the list of subject groups see https://postki.bdo.badw.de/sachgruppen/

[2] The service is available on https://postki.bdo.badw.de, its code is open source and publicly available on https://gitlab.lrz.de/badw-data/AI4SemClass/-/tree/main/webapp

[3] See *prompt_sentences* and *prompt_mapping* at https://gitlab.lrz.de/badw-data/AI4SemClass/-/blob/main/experiments/__vars__.py

accept, modify, or reject them where necessary. In this way, the editorial process becomes a hybrid approach that combines automated classification with expert validation and final editorial decision-making.

2.2 Integration into the editorial workflow

The integration into the editorial software tool of the particular dictionary project is achieved via the API of PostKI. For instance, a so-called user-defined action has been implemented into the XML-editor *Oxygen* that is used by the Bavarian Dictionary for writing entries, so that the lexicographers are able to use the service directly inside the program via clicking a specific button that then executes all necessary actions in the background, waits for the response, and finally presents them to the editor in a small view. Each subject group that has been returned by the LLM can be accepted and transferred into the XML-structure. Since there is no need for accepting any of the suggested groups, the response of the LLM can also be declined, thus leaving the full control to the lexicographer. The evaluation of the AI's response can be conducted through a rating, with a maximum of five stars indicating optimal performance. This process is integrated in the same way into the respective working environments of the two other projects.[4]

In everyday editorial work, the system is primarily used as a suggestion tool; the final decision remains with the editorial team. For large existing data sets, such as the Bavarian Dictionary with over 40,000 meanings that have not been linked to subject groups yet, the assignment is automated to enable basic semantic structuring, making previously unclassified data accessible for semantic research.

2.3 Evaluation

Up to this point, the tool has gone through different stages of evaluation. The first experiments, which form the basis for PostKI, are summarised in the following sub-section. This is followed by evaluations of the current tool in the sections on ratings of the responses of the large language models and the feedback from the lexicographers on utilising PostKI in everyday work.

2.3.1 Initial study

The first evaluation study covered 1,000 meanings from Post's assignments (cf. Raaf & Röhrer, 2025a): A random sample of meanings with already manually assigned subject groups was selected, and new suggestions were generated by the model. These suggestions were then evaluated both against the existing subject groups as a gold standard and by experienced editors. An assignment was considered correct if it was assessed as equivalent or better than the existing or newly made editorial decision. The results show a hit rate of over 80% (see Fig. 1). In practical use, this leads to a significant reduction in manual search and decision-making times, as editors can usually choose from a pre-structured selection of suitable subject groups. The values vary depending on the granularity of the subject groups and the available contextual information, which is consistent with general findings on LLM-based classification (cf. Wang et al., 2024).

[4] The Franconian Dictionary uses Excel sheets that are edited via user-form written in Visual Basic for Applications (VBA); the Dialectological Information System of Bavarian Swabia uses a web-based tool.

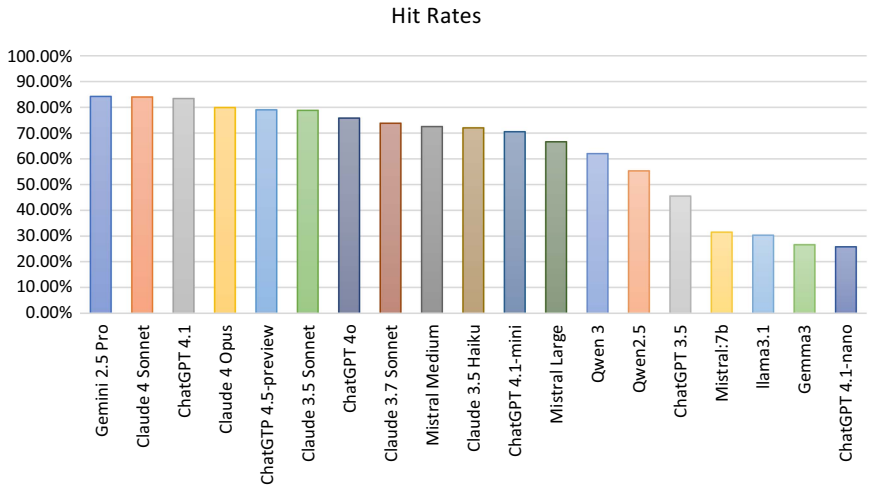


FIGURE 1 Results of the experiments (Raaf & Röhrer, 2025a, p. 4)

2.3.2 Surveys on the use of PostKI

1. Ratings of model suggestions
   PostKI was evaluated in May 2026 both by enabling the lexicographers to rate the AI's response on a scale from 1 ("very bad") to 5 stars ("very good") and by editor feedback. For the internal tests, the latest model from *Anthropic* in May 2026, *Claude Sonnet 4.6*, was used.[5] Some editors took notes, which are discussed in the interview section. The ratings indicate that the mappings of meanings to subject groups were rated as good or very good in about 67% of the cases, with the majority being the latter (see Fig. 2).[6] Hence, the integration of PostKI into the workflow of the three dictionary projects creates added value. In the interview (see results of the interviews below), the editors explained that a good rating was given if at least one subject group was identified correctly by the LLM.Of the ten editors who work in the three dictionary projects, six participated in the evaluations. All ratings are based on productive testing within the actual editorial workflow. In total, 564 ratings were collected in one week of work.

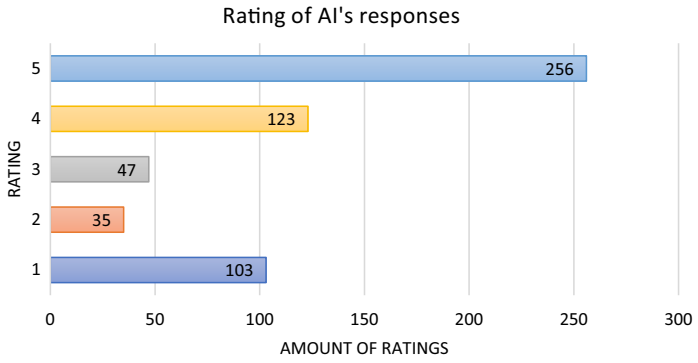


FIGURE 2 Ratings of AI's mapping of meaning to subject groups

[5] We chose *Sonnet* over Anthropic's flagship model *Opus* because – at least for this task – it is better (see Fig. 1) and significantly cheaper.

[6] It is noteworthy that this is congruent with the rate that was achieved in our first experiments in 2025 regarding the direct hits of the LLM (Raaf & Röhrer, 2025b, p. 18) as well as the hit rate of German embedding models. The results of the latter study have not yet been published.

2. Questionnaire on editorial usability
   In addition to the evaluation via rating, the overall user experience of the app should also be assessed. To this end, a set of thirteen statements was devised, which respondents were invited to rate using a Likert scale, indicating their degree of disagreement (-3) or agreement (3). The list of statements (Table 1) also contains pairs that are semantically identical but which are expressed in opposite terms (positive and negative). However, the negation was occasionally overlooked, resulting in both answers receiving the same rating. This led to a distortion of the standard deviation and the average rating. However, given the limited number of personnel involved in the projects and thus the evaluation, it has been decided that these errors will neither be corrected nor deleted.[7]

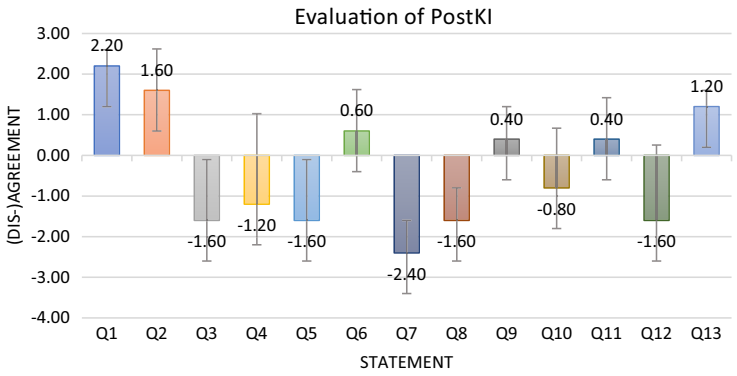


FIGURE 3 Evaluation of the use of *PostKI* as part of the editorial workflow

TABLE 1 List of statements on the use of PostKI

ID	Statement
Q1	Integrating PostKI into the workflow is straightforward.
Q2	The classifications generated by PostKI are useful.
Q3	The classifications generated by PostKI are mostly poor.
Q4	PostKI is not a useful tool.
Q5	It is cumbersome to use PostKI in day-to-day work.
Q6	Using PostKI saves time.
Q7	I can accept the AI's classifications without hesitation and without double-checking them.
Q8	AI-assisted classification takes more time than manual classification.
Q9	PostKI sometimes provides new insights into the semantic diversity of a meaning.
Q10	PostKI is too slow for use in daily work.
Q11	PostKI suggested a subject group that I consider better than the one I had (initially) in mind.
Q12	I am concerned that I will rely too heavily on the AI's response (in the future).
Q13	The AI's classifications are a useful supplement to human expertise.

[7] In large-scale questionnaire surveys, errors of this kind are usually eliminated either individually or in their entirety—that is, the entire data set—in order to analyse data sets that are as clean as possible and therefore as valuable as possible.

2.3.3 Interviews

In addition to the evaluation method described above, the authors wished to adopt a direct, non-statistical approach in order to gain a clearer understanding of how the system is used, what its benefits are, where any problems lie, and what improvements might be suggested by the lexicographers. To this purpose, interviews were conducted with the editors, in which they were primarily asked to speak openly about their impressions on usability, speed, reliability or quality of responses, and overall impression. In response to these statements, we asked appropriate follow-up questions, e.g., to clarify in which cases the results appeared to be less favourable.

Although we felt that the response time was slow, it is by no means a problem for the editors. There is a consensus that a time frame of five to ten seconds constitutes a marked improvement over the manual research method, which involves looking up meanings and their subject groups in various dialect dictionaries—a process that may well yield no results—or consulting the full list of subject categories. Overall, the editors rated the quality of the AI responses as good and identified the integration of PostKI into the editorial workflow as a beneficial development. The classification of abstract and concrete terms was found to be equally effective. However, weaknesses were particularly noticeable in definitions consisting exclusively of verbs, or of adverbs, modal particles, or adjectives as parts of the meaning: One such example is the meaning “freches Kind” (cheeky child), where the AI tended to assign semantic categories based mainly on the adjective “cheeky” rather than the head “child”. This behaviour is indeed challenging to understand, as both the prompt and the fundamental semantic understanding of modern language models recognise the attribute as such or, at least, do not prioritise it over the base. In certain cases, the classification proved to be less specific than desired. For instance, the term “Oskar” was classified as name rather than as first name. In other cases, AI has returned a subject group that is correct, but not the most appropriate one: First, “Ehemann (verächtlich)” has not been linked to the assumed and correct subject group, since the LLM seemed to concentrate more on “verächtlich” (derogatory) than on “Ehemann” (husband). The editor repeated the action several times because she was surprised that the category for “husband” was missing in AI’s suggestions and thus noticed that the responses of the model vary more than those of other meanings.[8] The editors’ assessment of incorrect versus correct classifications consistently falls within the range of 1.5 to 2 on a scale from -3 (very bad) to 3 (very good), thus aligning with the consensus that the classifications are good.

3 Results and discussion

3.1 Responsible use of AI

Following Virginia Dignum’s insistence that the “question zero” should always be “why AI?” (cf. Dignum, 2025, 00:20:00ff.), the project did not begin from the availability of generative AI, but from a concrete linguistic and editorial problem.[9] Throughout the project’s development, it has therefore been important to employ artificial intelligence in a conscientious manner, reserving its use for situations where it provides a clear methodological benefit for semantic classification.[10] Consequently, the prevailing AI-first approach was not adopted, nor was the assumption made that the humanities must adapt to IT solutions. Instead, the linguistic problem itself was identified as the starting point, with the objective of developing the most

[8] Other models than *Sonnet* also showed this issue in further tests afterwards.

[9] For a broader discussion of the responsible development and use of artificial intelligence, see Dignum (2019).

[10] Also see chapter “Conclusion and outlook”.

efficient solution possible. While there are many potential benefits of AI, it is argued that its implementation is occurring too rapidly in a variety of fields, including the Humanities. This accelerated deployment may be partly attributed to unrealistic expectations regarding its performance, particularly its capacity for understanding. Thus, the objective was to obtain results that could be directly utilised and contextualised within the problem—even if they turned out to be poor—without adapting the humanities to the needs of AI simply to make the use of cutting-edge technology appear sensible.

3.2 Reliability and error profiles

The evaluation shows that the models perform well in a controlled classification scenario. Errors in the sense of misclassifications occur primarily in cases of granularity conflicts between semantically similar subject groups, in highly context-dependent meanings, and in model-typical preferences for prototypical interpretations. As an example for context-depending meanings, "ruhen lassen" (let smth. rest) was assigned to the group 9800 "Bewegung, Wechsel, Veränderung, Stillstand" (movement, change, transformation, stagnation) by Rudolf Post and to several incorrect groups by AI. The dialect example that the assignment is based on, however, shows that the topic is of legal matter so that 8300 "Recht, Justiz" (law, justice) would be correct. These findings correspond to general observations on semantic annotation (cf. Pedersen et al., 2022). In cases where the mapping did not hit the gold standard nor the assumption of the lexicographer, AI's suggestions rarely were entirely wrong. The LLM still hit the semantic scope of the particular meaning in a way that editors would accept it.

The provision of objectivity or neutrality by the AI was cited as a further advantage. As statements of this kind must be viewed critically, particularly given the inherently probabilistic nature of generative AI, further clarification was sought regarding the intended meaning. It was elucidated that the AI's classification of subject groups is impartial, exhibiting a consistent quality that remains unaffected by external factors such as the editor's mood on a given day. Also, the service produced classifications that furnished novel insights into the semantic domain of meaning, which the human editor had not directly ascertained. For instance, the meaning "hastig, übereilt" (hasty, rushed) was assigned to the group 9600 "Zeit" (time) by Rudolf Post. AI, however, assigned it to 5570 "sich eilen, tummeln, trödeln" (to hurry, to run about, to dawdle), which has been considered a very fitting additional group by our evaluating lexicographer. This phenomenon can be considered an 'enrichment despite experience', a concept that is applicable to the utilisation of PostKI in general, as asserted by one editor.

3.3 Impact on editorial practice

The use of generative models influences not only the speed but also the systematics of classification. While manual assignments often vary, the model suggestions serve as a consistent reference point. In individual cases, model decisions were rated as more accurate than previous manual assignments according to the lexicographers (cf. statements Q9, Q11, and Q13). Algorithmically mediated homogenisation can contribute to quality assurance but also raises questions about normative effects and the possible standardisation of editorial decisions. The responses to Q12 also demonstrate that the utilisation of artificial intelligence does not substitute for subject-specific expertise but rather enhances it. Furthermore, the lexicographers did not perceive a tendency to become excessively dependent on AI or to relinquish personal responsibility.

Hence, the feedback also shows that human oversight continues to be regarded as an essential final step in the classification process. The purpose of the system is not to replace editorial work through automation, but to support and accelerate existing workflows through targeted suggestions. Since linguistic work requires a particularly high degree of precision and consistency, the integration of an automated system that is itself not free of errors would not be considered appropriate without continued expert validation. At the same time, acceptance among editors plays an important role in the productive use of such tools. Positive experiences with the quality and usefulness of the generated suggestions appear to increase trust in the system and contribute to a more efficient integration into everyday editorial practice.

Some editors reported that their initial scepticism gave way to surprise at the overall performance of the system, stating that they see clear advantages to integrating the tool into their established workflow.

3.4 Limitations

The evaluation of PostKI is subject to several constraints. First, the number of participating lexicographers is small, as the tool was tested within three specialised dictionary projects with limited editorial teams. The survey results thus provide valuable insight into expert use; however, they should not be interpreted as statistically representative user data.

Second, the ratings assess PostKI's usefulness as a suggestion system rather than the accuracy of a fully automated classifier. Because the system can return up to three subject groups, editors may rate a response positively when at least one suggestion is appropriate. The results therefore primarily indicate whether the tool provides useful editorial support, rather than whether it consistently identifies a single correct category without human validation.

Third, the findings are tied to the models and prompting strategy used in the study. LLMs are characterised by their rapid development, and it is important to note that model performance, costs, latency, and output behaviour may vary over time and across providers, making it impractical to evaluate every newly released system comprehensively. Experiments with more recent models produced broadly comparable outcomes, however, it is essential to recognise that the present study is an evaluation of a specific workflow, rather than a definitive evaluation of individual models or LLM-based semantic classification in general. The observed error patterns demonstrate that LLM-generated classifications still require expert validation. The identification of challenging cases is of particular importance, and these include meanings with insufficient context, conflicts between semantically related subject groups, overly general classifications, and cases in which the model assigns greater weight to modifiers than to the semantic core of the definition. These limitations provide a foundation for the central assumption of the paper: PostKI is most useful as part of a hybrid workflow in which automated suggestions accelerate and structure editorial work, while final responsibility remains with the lexicographers.

4 Conclusion and outlook

This study demonstrates that generative language models can provide practical support for the semantic classification of lexicographical content. The hybrid approach achieves high acceptance among participating editors and suggests more consistent results and reduced editorial effort. Its principal value lies not in full automation, but in targeted decision support. The editorial teams of the three dictionary projects have welcomed the tool as a

valuable addition. This may help increase acceptance of the use of modern digital methods in lexicography particularly, and the humanities in general.

At the same time, such systems influence editorial decision-making processes and require reflection on transparency, error control, and normative effects. Future work will focus on larger user studies, detailed error analyses, and comparisons of model types, classical machine learning methods (such as support vector machines or embedding models), and training/prompting strategies. PostKI will also be developed into a multi-layered tool: In the first stage, the system will check whether the exact definition has already been assigned to a subject category in one of the participating dictionaries. If this is not the case, the second layer will check whether traditional machine learning methods can assign it with a high degree of probability. Only if this also fails will the LLM be consulted. This would create a responsible, efficient, and reliable problem-first workflow that adds value to the editorial everyday work.

References

Dignum, V. (2019). *Responsible Artificial Intelligence: How to Develop and Use AI in a Responsible Way*. Springer International Publishing. https://doi.org/10.1007/978-3-030-30371-6

Dignum, V. (2025). *Responsible AI: Governance, Ethics, and Sustainable Innovation* [Munich AI lectures]. https://youtu.be/WQZQV1zlX_c

Hallig, R., & von Wartburg, W. (1963). *Begriffssystem als Grundlage für die Lexikographie. Versuch eines Ordnungsschemas* (2. neu bearb. und erw. Aufl.). Deutsche Akademie der Wissenschaften zu Berlin.

König, A., Raaf, M. & Klepsch, A. (2021). Fränkisches Wörterbuch (WBF). In A. N. Lenz and P. Stöckle (Eds.), *Germanistische Dialektlexikographie zu Beginn des 21. Jahrhunderts* (pp. 77–104). Franz Steiner Verlag.

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, 11(1), 426. https://doi.org/10.1057/s41599-024-02889-7

Liu, C., Zhang, H., Zhao, K., Ju, X. & Yang, L. (2024). LLMEmbed: Rethinking Lightweight LLM's Genuine Function in Text Classification. In L.-W. Ku, A. Martins & V. Srikumar (Eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 7994–8004. https://doi.org/10.18653/v1/2024.acl-long.433

Pedersen, B., Sørensen, N. C. H., Nimb, S., Flørke, I., Olsen, S. & Troelsgård, T. (2022). Compiling a Suitable Level of Sense Granularity in a Lexicon for AI Purposes: The Open Source COR Lexicon. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk, & S. Piperidis (Eds.), *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 51–60). European Language Resources Association. https://aclanthology.org/2022.lrec-1.6/

Post, R. (1998). Möglichkeiten der elektronischen Strukturierung, Vernetzung und Verfügbarmachung von lexikographischen Daten bei der Arbeit am Pfälzischen Wörterbuch. In R. Große (Ed.), *Bedeutungserfassung und Bedeutungsbeschreibung in historischen und dialektologischen Wörterbüchern* (pp. 211–220). Hirzel.

Raaf, M. (2025). *Bayerns Dialekte Online—Ein dynamisches Sprachinformationssystem zu den Dialektwörterbüchern Bayerisches Wörterbuch, Dialektologisches Informationssystem von Bayerisch-Schwaben und Fränkisches Wörterbuch*. [Dissertation, Friedrich-Alexander-Universität Erlangen-Nürnberg]. https://open.fau.de/handle/openfau/34909

Raaf, M., & Röhrer, I. (2025a). Semantische Klassifikation lexikographischer Inhalte mithilfe künstlicher Intelligenz neu denken? Ergebnisse einer Studie zur Erweiterung des Bayerischen Wörterbuchs um Ontologie mithilfe von LLMs. In U. Henny-Krahmer, A. Theise, & P. Helling (Eds.), *FORGE 2025—Forschungsdaten in den Geisteswissenschaften: Daten neu denken. Konferenzabstracts*. https://doi.org/10.5281/zenodo.17178256

Raaf, M., & Röhrer, I. (2025b). *Semantische Klassifikation lexikographischer Inhalte mithilfe künstlicher Intelligenz neu denken? Ergebnisse einer Studie zur Erweiterung des Bayerischen Wörterbuchs um Ontologie mithilfe von LLMs*. FORGE 2025 - Daten neu denken. [Presentation]. Rostock. https://doi.org/10.5281/ZENODO.17227619

Schnabel, M., Raaf, M., & Schwarz, D. (2021). Bayerisches Wörterbuch (BWB). In A. N. Lenz & P. Stöckle (Eds.), *Germanistische Dialektlexikographie zu Beginn des 21. Jahrhunderts* (pp. 47–76). Franz Steiner Verlag.

Schwarz, B. (2022). *Das dialektologische Informationssystem von Bayerisch-Schwaben: Dokumentation und mögliche Präsentation von Sprachdaten mit Multimedia im Internet*. Franz Steiner Verlag.

Schwarz, B., Funk, E., Raaf, M., & Welsch, U. (2021). Dialektologisches Informationssystem von Bayerisch-Schwaben (DIBS). In A. N. Lenz & P. Stöckle (Eds.), *Germanistische Dialektlexikographie zu Beginn des 21. Jahrhunderts* (pp. 105–142). Franz Steiner Verlag.

Stöckle, P., Elsner, D., Koppensteiner, W., & Korecky-Kröll, K. (2025). LLM-Assisted Dialect Lexicography: Challenges and Opportunities in Processing Historical Bavarian Dialects. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda & A. Salgado (Eds.), *Electronic lexicography in the 21st century (eLex 2025): Intelligent lexicography. Proceedings of the eLex 2025 conference*. Lexical Computing CZ s.r.o.

Wang, Z., Pang, Y., Lin, Y., & Zhu, X. (2024). Adaptable and Reliable Text Classification using Large Language Models. In Y. He, W. Hamidouche, I. Razzak, H. Hacid & M. Panov (Eds.), *2024 IEEE International Conference on Data Mining Workshops (ICDMW)*, (pp. 67–74). https://doi.org/10.1109/ICDMW65004.2024.00015

Contact information

Ines Röhrer
Bavarian Academy of Sciences and Humanities
ines.roehrer@badw.de

Manuel Raaf
Bavarian Academy of Sciences and Humanities
manuel.raaf@badw.de

Acknowledgements

We gratefully acknowledge the computational resources as well as the support provided by the Leibniz Supercomputing Centre (https://www.lrz.de/en/), an institute of the Bavarian Academy of Sciences and Humanities. Also, we would like to thank Rudolf Post once again for granting us permission to use, modify, and publish his subject groups under the open licence CC-BY.

Antonio San Martín & Catherine Trekker

INTRODUCING HUMAN-CENTEREDNESS IN AI-ASSISTED LEXICOGRAPHY

Abstract This paper proposes a human-centered artificial intelligence (HCAI) framework for AI-assisted lexicography. While generative AI offers significant opportunities to enhance lexicographic work, it also raises concerns regarding the future role of lexicographers and the preservation of linguistic and cultural diversity. Drawing on HCAI principles and previous applications in other language professions, the paper identifies four interrelated dimensions through which AI integration in lexicography can be understood and critically examined: the augmented lexicographer, the sociotechnical context of AI integration, bias, and the design of AI-powered lexicographic tools. The framework argues that AI should augment rather than replace lexicographers, combining automation with meaningful human control. It further emphasizes the importance of preserving professional agency, mitigating AI-generated biases, and designing tools around the needs of lexicographers. By doing so, the paper provides a foundation for future research and the beneficial integration of AI into lexicographic workflows.

Keywords Human-centered AI; generative AI; AI-assisted lexicography

1 Introduction

The widespread availability of generative artificial intelligence (GenAI), particularly large language models (LLMs) such as those powering ChatGPT and Claude, has the potential to transform lexicographic practice (Fuertes-Olivera & Tarp, 2025; Lew, 2024). On the one hand, the growing ability of LLMs to provide lexical information has raised concerns about the future role of lexicographers and the continued relevance of dictionaries and other lexicographic products (de Schryver, 2023). On the other hand, artificial intelligence (AI) offers powerful new opportunities to support lexicographic work. These developments place lexicography at a critical juncture and raise a fundamental question: how can AI be integrated so that it enhances rather than undermines the work of lexicographers?

In this context, the paradigm of human-centered AI (HCAI) (Capel & Brereton, 2023; Schmager et al., 2023; Shneiderman, 2020) provides a useful framework for guiding the integration of AI into lexicography. HCAI places human needs and well-being at the center of AI development and challenges approaches that prioritize machine performance and cost reduction at the expense of human values. Applied to lexicography, this perspective recognizes that lexicographic resources provide important social benefits by supporting reliable access to lexical knowledge and preserving linguistic and cultural diversity. To protect these benefits, lexicographers must retain meaningful control over AI-assisted workflows, so that automation serves to augment professional expertise rather than displace or marginalize it.

Against this background, this paper proposes a conceptual framework for AI-assisted lexicography grounded in HCAI.[1] It identifies a set of interrelated dimensions through which the relationship between AI and lexicographic work can be understood and critically examined.

[1] This paper focuses on the application of HCAI to lexicographic work from the perspective of the lexicographer. Although the same approach can also be applied to the impact of AI on end users of lexicographic products, the present discussion concentrates on the former while acknowledging its direct implications for the latter.

By doing so, the paper seeks to provide a conceptual starting point for future research and support the beneficial integration of AI into lexicographic practice.

The remainder of the paper is organized as follows. Section 2 introduces the principles of human-centered AI and their application to language professions. Section 3 develops a human-centered framework for AI-assisted lexicography organized around four interrelated dimensions: the augmented lexicographer, the sociotechnical context of AI integration, bias, and the design of AI-powered lexicographic tools. Section 4 concludes by reflecting on the implications of the framework and identifying priorities for future research.

2 Human-Centered Artificial Intelligence

Much contemporary AI development prioritizes speed and efficiency over human values and societal impacts (Schmager et al., 2023, p. 1). In response, HCAI defines human flourishing as the central aim of AI (Vallor, 2024a, p. 13) and calls for equitable outcomes and broadly shared benefits for humanity (Stanford Institute for Human-Centered AI, 2025). It reconceptualizes AI as a tool designed to serve human needs and challenges the assumption that progress necessarily demands minimizing human involvement through automation. Instead, HCAI frames AI as amplifying human capabilities while preserving human control.

This reorientation also requires acknowledging that AI functions within sociotechnical ecosystems of people, institutions, and norms (Dignum & Dignum, 2020, p. 2). In profit-driven environments, organizations may impose AI to cut costs rather than empower users. HCAI also addresses ethical challenges in AI deployment (including bias, environmental impact, or intellectual property concerns) as well as the design of AI tools that prioritizes users' needs and goals.

Translation was the first language profession to which HCAI was applied (Jiménez-Crespo, 2025a; O'Brien, 2024). Although lexicography faces distinct challenges, translation offers a useful cautionary example. Neural machine translation has disrupted the field for over a decade, reshaping workflows and labor conditions. Organizations have often imposed AI tools to maximize speed and reduce costs at the expense of translators' needs. Moorkens (2020) describes this trend as "*digital Taylorism*," in which automation reduces translators to "*cogs in a large machine*," diminishing their agency. An HCAI approach to translation seeks to retain the benefits of automation while reducing its harms.

Among language professions, terminology work arguably faces AI-related challenges that most closely resemble those of lexicography, as both involve the systematic extraction and representation of lexical knowledge. This paper adapts the HCAI framework proposed for terminology (San Martín, 2025) to the specific context of lexicographic work.

3 Human-Centered AI-Assisted Lexicography

GenAI is arguably the most significant development in lexicography since the introduction of corpus analysis tools. For users, it enables new ways of accessing lexicographic information; for lexicographers, it offers the potential for greater efficiency. At the same time, however, it raises concerns about the future of lexicography, lexicographers, and lexicographic products (de Schryver, 2023).

GenAI chatbots are increasingly capable of fulfilling users' lexicographic needs and, in some respects, offer advantages over traditional dictionaries. They provide access to lexical information

in a flexible and user-friendly manner, allowing users to formulate queries in context, receive customized answers, and ask follow-up questions.

However, GenAI-based chatbots are also prone to errors and hallucinations, may reproduce biases, and, due to their non-deterministic nature, generate inconsistent answers. These limitations raise important concerns about their reliability. In contrast, dictionaries produced through established lexicographic processes provide information that has undergone expert validation. Moreover, dictionaries are not merely tools for accessing lexical data. As Arias-Arias et al. (2025, p. 252) argue, they also hold cultural, social, and emotional value, functioning as "*repositories of cultural heritage and linguistic diversity, reflecting the evolution of language and society. They serve as educational artifacts, historical records, and symbols of linguistic identity*." In the case of prescriptive dictionaries, they additionally provide guidance on recommended usage.

From a human-centered perspective, the wholesale replacement of lexicography with GenAI would conflict with broader social values and benefits. Yet cost pressures and growing overtrust in AI risk marginalizing lexicographic expertise. AI will likely transform lexicography, which had already extended beyond dictionary compilation before GenAI emerged (Tarp, 2019, p. 231). Although some traditional dictionaries, particularly online ones, will probably survive, lexicographic authority may increasingly become embedded within broader AI-mediated language infrastructures. In response, Sennrich & Ahmadi (2025) propose the term "*conversational lexicography*" to describe an approach in which AI chatbots serve as interfaces for lexicographic content curated by lexicographers, combining the reliability of expert-curated content with conversational flexibility and user-friendliness.

HCAI envisions AI as augmenting rather than replacing lexicographers. In this view, GenAI serves as a powerful tool whose assistance can be broadly grouped into two overlapping modes: post-editing lexicography and intelligent support. Although post-editing lexicography (a term coined by Jakubíček et al., 2018) predates GenAI, recent advances enable unprecedented levels of automation, such as the automatic generation of definitions or lexicographic illustrations. Intelligent support, by contrast, includes tasks such as analyzing lexical data, assisting corpus analysis, and conducting quality assurance. Emerging agentic systems, including OpenAI's Codex and Anthropic's Claude Code and Cowork, further enable lexicographers to analyze and manipulate lexical data without requiring programming expertise.

Despite these powerful capabilities, GenAI has clear limitations as a lexicographic tool. Existing evaluations of LLMs in diverse lexicographic tasks[2] suggest that, while they can achieve strong results in certain contexts, particularly for high-resource languages, their performance remains imperfect and inconsistent. Furthermore, persistent biases continue to pose significant challenges.

As AI both threatens and transforms lexicography, a human-centered approach is essential to ensure that AI supports rather than marginalizes lexicographers. The following sections examine four interrelated dimensions of human-centered AI-assisted lexicography: the augmented lexicographer, the sociotechnical context of AI integration, bias, and the design of AI-powered tools.

[2] Due to space limitations, it is not possible to cite or review all relevant studies. Representative evaluations of LLMs include those by Chen et al. (2024), Kosem et al. (2024), and Rees & Lew (2024).

3.1 The Augmented Lexicographer

Technological tools have enhanced lexicographers' capabilities for decades. This computational augmentation enables lexicographers to work more efficiently and potentially produce higher-quality output than would be possible without technological support. Since lexicography is now inseparable from technology, the contemporary lexicographer can be understood as an augmented lexicographer[3]. This transformation was already accelerating before the emergence of GenAI. Krek (2019), for instance, argued that lexicography had undergone a computational turning point driven by rapid advances in natural language processing and the growing availability of language data.

From a HCAI perspective, the augmentation of lexicographers through AI should be guided by the two principles formulated by Shneiderman (2020), which are discussed in the following two subsections. The final subsection examines how GenAI is transforming the role of lexicographers and the skills required to effectively leverage AI-based augmentation.

3.1.1 High automation and strong human control

The first HCAI principle holds that high levels of automation and human control are not mutually exclusive but can coexist productively. Traditional AI approaches often frame automation and human oversight as a zero-sum relationship, a view HCAI explicitly rejects (Shneiderman, 2020, p. 115). Accordingly, under the right circumstances, AI-assisted lexicographers are expected to outperform both fully automated systems and entirely manual approaches, since human expertise and computational efficiency operate synergistically (Capel & Brereton, 2023, p. 6). Such augmentation, however, requires a careful balance between machine and human intervention. Excessive automation may compromise the quality of lexicographic output, whereas excessive human involvement can unnecessarily limit the efficiency gains that AI systems may provide (Shneiderman, 2020, p. 116).

This balance between automation and human control must be determined at the level of individual subtasks rather than for an entire task. In definition generation, for example, the lexicographer may retain control by designing the prompt that guides the process, specifying elements such as the target audience, intended perspective, and stylistic constraints, and providing a relevant corpus. The LLM may then automatically generate a draft definition. Human control may resume during the review and validation stage, where the lexicographer evaluates and revises the output as needed. At this point, automation may once again support the process by assisting with corpus analysis, while the lexicographer retains final authority over the definition ultimately adopted.

The balance cannot be defined uniformly across all lexicographic contexts. It depends on project goals and LLM performance in different tasks, domains, and languages. Consequently, the extent and form of automation may differ according to factors such as the work setting (commercial, institutional, academic, etc.), the type of lexicography involved (general, bilingual, learner, etc.), the languages concerned, and even the individual preferences of the lexicographer.

Achieving this balance requires systematic research on the performance of AI systems for lexicography. This includes benchmarking LLMs across lexicographic tasks and subtasks to identify strengths and limitations. It also requires examining AI use in situated practice: how lexicographers interact with tools throughout their workflow and whether these interactions

[3] Adaptation of the notion of "augmented translator" (DePalma & Lommel, 2017).

actually improve their capabilities, in a manner comparable to translation post-editing studies (Vieira, 2019). Bedard et al. (2026) show that AI use can generate cognitive fatigue in some work contexts, underscoring the need to identify forms of automation that provide clear benefits. Following O'Brien's (2024, p. 400) proposal for translation, research should also elicit lexicographers' views on the types of AI augmentation they consider most useful and on how they would like these systems to support their work. For example, since corpus evidence constitutes a central component of lexicographic practice (Hanks, 2012), lexicographers will likely value its integration into AI-assisted workflows.

3.1.2 Amplification rather than emulation

Shneiderman's (2020) second HCAI principle holds that AI should not aim to replicate human intelligence. Artificial and human intelligence are fundamentally different in nature, and treating AI as if it should mimic human cognition misconstrues both forms of intelligence. Accordingly, HCAI conceives AI not as a substitute for human intelligence, but as a tool that extends it.

This distinction is crucial in lexicography. AI systems excel at processing large datasets and detecting patterns at great speed (Ozmen Garibay et al., 2023, p. 422). LLMs, however, manipulate language without direct access to the realities to which it refers (Poibeau, 2025, p. 48). As Vallor (2024b, p. 24) argues, no dataset can capture the open-ended complexity of lived human experience. Lexicographic work therefore cannot be reduced to statistical pattern recognition alone. Beyond corpus evidence, lexicographers can draw on world knowledge, embodied experience, introspective linguistic competence, and direct engagement with language communities.

Within this framework, AI contributes to lexicography because it offers strengths humans lack. For instance, it can quickly identify candidate senses, synonyms, antonyms, or interlingual equivalents that are time-consuming to extract from corpora. However, producing and validating reliable lexical data ultimately requires expert lexicographic judgement and human understanding of language and the world.

3.1.3 The new role and skills of lexicographers

As Tiberius et al. (2024) note, lexicographers do more than edit dictionary entries. Their roles increasingly encompass project management, data management, communication with computational specialists and end users, fundraising, and public relations. The integration of AI is expected to further transform the profession by reshaping both the tasks that lexicographers perform and the competencies that define their professional value.

Drawing on AI literacy frameworks proposed by Annapureddy et al. (2025) and Zhang & Magerko (2025), lexicographers will need to acquire skills such as understanding how different AI models can support lexicographic tasks, selecting and using appropriate AI tools, crafting effective prompts, and recognizing the ethical and legal implications of AI use. They must also critically evaluate AI-generated outputs, identifying and correcting errors and potential biases.

Lexicographers must also develop sufficient technical expertise to retain control over the integration of AI. Tarp (2015, p. 219) already argued that lexicographers require adequate computational knowledge to guide programmers, rather than allowing them to shape lexicographic decisions. Similarly, lexicographers today need a solid understanding of AI to supervise

its integration and ensure that implementation decisions align with lexicographic principles and objectives.

Finally, developing AI-related competencies must not undermine the fundamental expertise that defines the lexicographic profession, a concern similarly raised by Jiménez-Crespo (2025b, p. 279) in relation to translators. If the role of lexicographers is reduced to merely validating AI-generated output, the profession risks progressive de-skilling.

3.2 The Sociotechnical Context of AI Integration in Lexicography

AI integration in lexicography is more than a technical matter, since it always takes place within a broader social context including a variety of *"stakeholders, institutions, cultures, norms and spaces"* (Dignum & Dignum, 2020, p. 2). This sociotechnical environment involves multiple stakeholders whose roles vary across commercial, institutional, and academic settings, including companies, universities, publishers, public institutions, lexicographers, programmers, end users, and the wider language community. These actors collectively shape both the ways AI systems are integrated into lexicographic workflows and the consequences of that integration.

While the traditional AI goals of machine autonomy and performance are not inherently problematic, augmenting lexicographers through AI cannot be considered human-centered if it dehumanizes them by prioritizing profit extraction or cost reduction. As Kalluri (2020, p. 169) argues, human-centeredness ultimately depends on where power resides. In lexicography, AI integration should primarily support lexicographers and the broader social benefits of their work, rather than allowing institutional and economic interests to undermine them.

AI-related tensions intensify pre-existing structural problems in lexicography, since many dictionary projects were already operating under conditions of financial fragility (Tiberius et al., 2024, p. 19). In this context, from the perspective of sustainable lexicography (Colman, 2016; Fuertes-Olivera, 2024), AI augmentation may help preserve some lexicographic projects by reducing costs. However, the same economic logic may also encourage funders to maximize automation in ways that marginalize lexicographers. The difference lies in whether cost savings are achieved by supporting expert work or by removing expert authority from decisions that determine the reliability and societal contribution of the lexicographic resource.

The sociotechnical conditions under which AI is implemented also directly affect lexicographers' well-being and their ability to carry out their professional responsibilities. For AI integration to remain human-centered, lexicographers must retain meaningful agency over how AI is incorporated into their workflows. When AI systems are imposed without such control, lexicographers risk being reduced to mere validators of machine-generated output, with negative consequences for professional autonomy, motivation, job satisfaction, and stress levels (Oviatt, 2021, p. 278). More importantly, such conditions may limit lexicographers' capacity to produce reliable and socially valuable resources that reflect linguistic and cultural diversity.

In this context, the benchmarking and evaluation of LLMs acquire an additional human-centered function. Beyond measuring technical performance, such research can help identify which aspects of lexicographic expertise current AI systems cannot reliably reproduce. It can therefore provide further evidence that expert human judgment remains necessary for ensuring the quality and social value of lexicographic resources. In doing so, it also provides lexicographers with empirical grounds to defend their professional role and to claim meaningful authority over AI-assisted workflows.

3.3 Bias

Major sources of LLM bias relevant to lexicography include training data composition and algorithmic design (Ferrara, 2023). While the exact composition of the training data used by major proprietary LLMs remains undisclosed (Ravichander et al., 2025, p. 1962), the Common Crawl corpus is widely considered their largest component (Baack, 2024, p. 2199). This corpus consists largely of uncurated web data, over 40% of which is in English. Biases present in the training data are subsequently amplified by algorithmic processes. Because LLMs reproduce statistical patterns in their training data and are optimized to generate high-probability outputs, they reinforce and overrepresent dominant patterns (Sourati et al., 2026, p. 5). These factors produce distinct but often overlapping forms of bias, discussed separately below.

Linguistic bias. Linguistic bias denotes the tendency of LLMs to favor certain languages or linguistic features in their outputs. Since English is overrepresented in the training data of many leading LLMs, AI-generated lexicographic content may be more accurate in this language. Cross-linguistic interference may also occur, whereby one language, typically English, influences outputs in other languages, as shown by Rigouts Terryn and de Lhoneux (2024). Linguistic bias may also operate within a single language, when dominant regional varieties are projected onto less-resourced ones.

Temporal bias. Temporal bias refers to the risk that LLMs generate lexicographic information that is misaligned with the relevant time frame. Models may present diachronic data as synchronic, or vice versa, and may hallucinate when evidence is sparse. Although temporal bias in AI-assisted lexicography remains understudied, studies report LLM weaknesses in date-sensitive factual accuracy and chronology (Wallat et al., 2024). A key cause is that training corpora often mix historical and contemporary texts without clear temporal boundaries (B. Zhao et al., 2024, p. 15015). In the case of neology, temporal bias may further stem from the training data cutoff.

Cultural bias. Cultural bias denotes the tendency of LLMs to reproduce the cultural assumptions, values, norms, and perspectives that dominate their training data while underrepresenting or distorting others. Consequently, LLM outputs may reflect Anglophone and Western cultural interpretations as universal, a tendency reported by Tao et al. (2024).

Epistemic bias. Epistemic bias refers to the tendency of LLMs to overrepresent dominant forms of knowledge by amplifying what is most prevalent in their training data while marginalizing less represented perspectives. In practice, this often entails privileging mainstream conceptualizations prevalent in the Global North. This is particularly relevant for specialized language, since terms may be conceptualized differently across domains, theoretical traditions, and ideological currents (San Martín, 2022). Consequently, AI-generated lexicographic output may obscure conceptual variation, disciplinary disagreement, or emerging forms of knowledge.

Social bias. Social bias refers to the reproduction of prejudicial views, stereotypical representations, and discriminatory attitudes directed toward particular social groups on the basis of factors such as gender, race, ethnicity, age, sexual orientation, and disability (Gallegos et al., 2024, p. 1098). Although post-training techniques are intended to mitigate such biases, social bias remains an unresolved problem in LLMs (McIntosh et al., 2024).

The forms of AI bias discussed above have different consequences. Linguistic and temporal bias primarily produce descriptive distortions, since they compromise the faithful representation of language use. The role of the lexicographer is therefore to ensure descriptive adequacy by critically evaluating AI-generated output against corpus evidence and other reliable sources.

In prescriptive contexts, lexicographers must also prevent such distortions from being transformed into normative recommendations. This task specifically requires lexicographic expertise, since identifying subtle linguistic interferences and distortions cannot be fully automated. If left uncorrected, these forms of bias may contribute to the erosion of linguistic diversity.

Cultural, social, and epistemic bias raise a different type of problem because they concern visibility and legitimacy rather than descriptive accuracy alone. Although such biases in lexicography long predate AI (Kaplan, 2020), LLMs introduce new mechanisms for reproducing and amplifying them at scale. AI-generated lexicographic content may reproduce existing asymmetries while presenting them under the appearance of computational objectivity.

Critical lexicography has already emphasized that dictionaries are not neutral representations of language (W. Chen, 2019, p. 362), but rather products shaped by social, cultural, and ideological values. Given that, from the point of view of cognitive linguistics, meaning cannot be separated from encyclopedic knowledge (Evans, 2019, p. 386), representing language use necessarily involves selecting and organizing culturally and socially situated knowledge about the world. Consequently, even a strictly descriptivist approach cannot be fully neutral, as lexicographers must decide, in accordance with the aims and characteristics of a given project, which usages and perspectives to include and which to background or omit. Descriptivism, therefore, cannot be regarded as entirely neutral, because any description of language inevitably reflects particular social and epistemic standpoints.

From a human-centered approach, the role of lexicographers is therefore not to pursue an unattainable neutrality, but to critically manage AI-generated lexical representations through fairness and contextually adequate pluralism. This responsibility requires expert lexicographic judgment that cannot be delegated entirely to AI systems. Without the intervention of a lexicographer, AI outputs may privilege dominant cultural perspectives while distorting and marginalizing others, suppress epistemic diversity, and reinforce discriminatory representations affecting already marginalized social groups.

Managing AI-generated bias in lexicography requires intervention at multiple stages of the workflow. Upstream, lexicographers can reduce bias through careful prompt design and by constraining retrieval-augmented generation (RAG)[4] to appropriate sources. Downstream, bias mitigation requires the critical evaluation of AI-generated output through expert judgment, corpus analysis, and, in some cases, even the strategic use of AI systems themselves as bias-detection aids.

If lexicographers do not retain meaningful control over the integration of AI into lexicographic workflows, inaccurate and biased representations of the lexicon may circulate and acquire unwarranted authority. Such lexicographic information may subsequently influence real language use. The problem becomes even more significant when the resulting texts enter corpora used in lexicographic work or are incorporated into the training data of LLMs. This may contribute to self-reinforcing feedback loops that progressively amplify linguistic distortions as well as cultural, epistemic, and social inequities.

Research can contribute to bias mitigation by systematically evaluating LLMs across different types of bias, lexicographic tasks, and linguistic contexts. Such research would enable lexicographers to anticipate systematic distortions, make more informed methodological decisions, and maintain effective critical oversight over AI-assisted lexicographic workflows.

[4] RAG is a mechanism that allows LLMs to base their responses on external information.

3.4 The Design of AI-Powered Lexicographic Tools

Some publicly available tools used by lexicographers, such as TLex and AntConc, have begun integrating GenAI. However, the extent to which AI has been incorporated into in-house solutions, which are widely used in many lexicographic projects (Tiberius et al., 2024), is understudied. It is nevertheless reasonable to assume that much of the interaction between lexicographers and GenAI still occurs through standalone chatbots. From a human-centered perspective, this mode of interaction is often suboptimal. The need to move continuously between tools and chatbots introduces unnecessary friction into the workflow and limits the extent to which AI can meaningfully augment professional practice. Consequently, AI functionality should be integrated directly into lexicographic environments and workflows, provided that this does not undermine the lexicographer's autonomy. Alternatively, lexicographic tools can be integrated into AI agentic systems (e.g., Codex and Claude Code)[5]. Such integrations are not merely a technical issue but a design challenge that should be guided by established principles of human-computer interaction, particularly the well-developed field of human-centered design (HCD)[6], which addresses the physical, cognitive, social, and emotional factors of the human-machine relationship (Boy, 2011, p. 1).

Applying HCD principles (Benyon, 2019, p. 13) to AI-assisted lexicographic tools involves two fundamental requirements. First, systems and interfaces must be designed around the needs of lexicographers. Second, lexicographers themselves must be actively involved in the conception and refinement of AI-powered tools. This participatory approach helps avoid reverse adaptation (Winner, 2020, p. 174), whereby lexicographers and lexicographic projects are forced to adapt to technological constraints instead of having technology designed around lexicographic requirements. By placing lexicographers at the center of the design process, HCD also supports the practical implementation of broader HCAI principles, particularly the preservation of human control over AI systems and the genuine augmentation of lexicographers' capabilities.

With regard to preserving human control over AI, useful guidance can be drawn from recommendations developed in the context of HCAI-oriented translation research (Jiménez-Crespo, 2025a; O'Brien, 2024). These recommendations emphasize that AI-powered tools should combine strong usability with features that allow users to selectively enable or disable particular system functions and calibrate automation levels depending on the task or individual preferences. Role-aware permissions, version control, and the capacity of lexicographers to override AI suggestions could also ensure that they retain ultimate authority over lexicographic outputs.

Beyond preserving control, the interface of AI-assisted lexicographic tools should augment the lexicographer's capabilities by supporting decision-making. One important avenue for such augmentation is explainability, that is, the use of techniques that help users understand and evaluate AI outputs (Capel & Brereton, 2023, p. 6). Even though the internal processes of LLMs remain opaque (H. Zhao et al., 2024, p. 2), approaches such as RAG, chain-of-thought prompting, and reasoning models can provide, to a certain extent, evaluative cues about outputs.

Among these approaches, RAG is arguably one of the most valuable for lexicographic work because it enables AI outputs to be linked to identifiable and verifiable sources. The most desirable implementation involves retrieval from corpora selected by lexicographers themselves, since unrestricted web-based RAG may rely on unreliable, contextually inadequate, or even

[5] Agentic systems combine the knowledge and reasoning capabilities of LLMs with the ability to write and execute code to operate and orchestrate external tools, such as corpus analysis tools and databases.

[6] HCD has already been applied to lexicography from the perspective of the end users (Tarp & Gouws, 2020).

AI-generated material. This approach also has the advantage of pointing to specific linguistic evidence, a key issue in lexicography. Nonetheless, despite the value of RAG, it should not be the sole basis for LLM-generated suggestions, as doing so would limit the model's ability to draw on the broader knowledge acquired during training. LLMs can provide useful contributions that are not directly traceable to specific sources, but in such cases the relevant linguistic evidence must subsequently be identified and validated.

The augmentation of lexicographers also requires careful attention to the cognitive effects of automation. Two particularly important risks are reduced critical engagement and anchoring. Reduced critical engagement refers to a decline in the rigor with which users evaluate and validate AI-generated outputs (Heer, 2019, p. 1844). This tendency may be driven by fatigue, time constraints, or an automation-induced overtrust in the reliability of AI outputs that discourages verification (Klingbeil et al., 2024, p. 7). Tool design may mitigate this risk through empirically calibrated intentional friction, such as confirmation dialogs that encourage lexicographers to reassess AI suggestions before validation.

The anchoring effect may also emerge in AI-assisted lexicographic work when initial AI suggestions constrain subsequent human judgment and limit consideration of alternatives (Furnham & Boo, 2011). Early exposure to AI-generated proposals may thus narrow expert reasoning. One potential mitigation strategy is therefore to delay the presentation of AI suggestions until after lexicographers have completed an initial independent analysis. Such an approach allows AI to function as a source of support and comparison rather than as a starting point that shapes the entire decision-making process.

Taken together, these design recommendations should be regarded as provisional hypotheses rather than established best practices. Their effectiveness will ultimately depend on systematic empirical validation across a range of lexicographic contexts and workflows. Advancing AI-assisted lexicography therefore requires sustained collaboration between researchers, developers, and lexicographers, as well as the rigorous application of established HCI methodologies, including interviews, focus groups, usability studies, participatory design approaches, and observational research (Capel & Brereton, 2023, p. 4).

4 Conclusions

AI is likely to transform lexicography profoundly. The key question is not whether AI will be integrated into lexicographic work, but how. This paper has argued that HCAI provides a principled framework for ensuring that AI supports and strengthens lexicographic practice rather than undermines it. To this end, the paper has proposed a human-centered framework for AI-assisted lexicography organized around four interrelated dimensions: the augmented lexicographer, the sociotechnical context of AI integration, bias, and the design of AI-powered lexicographic tools.

A central premise of this orientation is that lexicographers should remain at the core of AI-assisted lexicography. From a human-centered perspective, AI should augment rather than replace the lexicographer's capabilities. High levels of automation and meaningful human control are not mutually exclusive; instead, they can coexist synergistically when AI systems are designed around the needs of lexicographers, preserving their decision-making authority and professional agency.

The implications of this framework extend well beyond lexicography as a profession, since lexicographic resources shape and support human communication. Preserving lexicographic expertise therefore serves broader societal goals. Decisions made now about AI deployment

will influence not only professional practice but also the long-term preservation of linguistic and cultural diversity.

Advancing a human-centered approach to AI-assisted lexicography requires a sustained empirical research agenda. Future studies should investigate the performance of LLMs across lexicographic tasks. They should also examine how lexicographers interact with AI in situated practice and determine which forms of automation genuinely augment their capabilities. Further research should also assess the impact of AI on lexicographers' working conditions and autonomy, evaluate the forms of bias that emerge in AI-assisted workflows in order to identify strategies to manage them, and compare alternative design approaches to integrating AI into the tools used by lexicographers. The framework proposed in this paper serves as a starting point for research aimed at ensuring that AI empowers lexicographers to continue contributing to human flourishing in an increasingly AI-mediated society.

References

Annapureddy, R., Fornaroli, A. & Gatica-Perez, D. (2025). Generative AI literacy: Twelve defining competencies. *Digital Government: Research and Practice*, 6(1), 1–21. https://doi.org/10.1145/3685680

Arias-Arias, I., Domínguez Vázquez, M. J. & Riveiro, C. V. (2025). People, machines and dictionaries: Is artificial intelligence killing dictionaries? In I. Dias, R. H. Gouws, A. Lobenstein-Reichmann & S. J. Schierholz (Eds.), *Regarding the Significance of Lexicography/Zur Bedeutsamkeit der Lexikographie* (pp. 235–258). De Gruyter. https://doi.org/10.1515/9783112216644-012

Baack, S. (2024). A critical analysis of the largest source for generative AI training data: Common Crawl. *The 2024 ACM Conference on Fairness, Accountability, and Transparency* (pp. 2199–2208). https://doi.org/10.1145/3630106.3659033

Bedard, J., Kropp, M., Hsu, M., Karaman, O. T., Hawes, J. & Rosen Kellerman, G. (2026). When using AI leads to "brain fry." *Harvard Business Review*. https://hbr.org/2026/03/when-using-ai-leads-to-brain-fry

Benyon, D. (2019). *Designing user experience: A guide to HCI, UX and interaction design* (4th ed.). Pearson.

Boy, G. A. (2011). A human-centered design approach. In G. A. Boy (Ed.), *The Handbook of Human-Machine Interaction* (pp. 1–20). CRC Press. https://doi.org/10.1201/9781315557380-1

Capel, T. & Brereton, M. (2023). What is human-centered about human-centered AI? A map of the research landscape. In A. Schmidt, K. Väänänen, T. Goyal, P. Ola Kristensson, A. Peters, S. Mueller, J. R. Williamson & M. L. Wilson (Eds.), *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (pp. 1–23). https://doi.org/10.1145/3544548.3580959

Chen, L., Dao, H. L. & Do-Hurinville, D.-T. (2024). AI empowerment: Where are we in the automation of lexicography? A metaphraseographic study. In A. Inoue, N. Kawamoto, & M. Sumiyoshi (Eds.), *AsiaLex 2024 Proceedings* (pp. 90–98). The Asian Association for Lexicography. HAL (hal-04664242).

Chen, W. (2019). Towards a discourse approach to critical lexicography. *International Journal of Lexicography*, 32(3), 362–388. https://doi.org/10.1093/ijl/ecz003

Colman, L. (2016). Sustainable lexicography: Where to go from here with the ANW (Algemeen Nederlands Woordenboek, An Online General Language Dictionary of Contemporary Dutch)? *International Journal of Lexicography*, 29(2), 139–155. https://doi.org/10.1093/ijl/ecw008

de Schryver, G.-M. (2023). Generative AI and lexicography: The current state of the art using ChatGPT. *International Journal of Lexicography*, 36(4), 355–387. https://doi.org/10.1093/ijl/ecad021

DePalma, D. A. & Lommel, A. (2017). *Augmented Translation Powers up Language Services*. Common Sense Advisory. https://csa-research.com/Blogs-Events/Blog/Augmented-Translation-Powers-up-Language-Services

Dignum, F. & Dignum, V. (2020). How to center AI on humans. In A. Saffiotti, L. Serafini & P. Lukowicz (Eds.), *Proceedings of the First International Workshop on New Foundations for Human-Centered AI (NeHuAI), Co-Located with the 24th European Conference on Artificial Intelligence (ECAI 2020), CEUR Workshop Proceedings*, *2659* (pp. 59–62). https://ceur-ws.org/Vol-2659/dignum.pdf

Evans, V. (2019). *Cognitive linguistics: A complete guide*. Edinburgh University Press.

Ferrara, E. (2023). Should ChatGPT be biased? Challenges and risks of bias in large language models. *First Monday*. https://doi.org/10.5210/fm.v28i11.13346

Fuertes-Olivera, P. A. (2024). Making lexicography sustainable: Using ChatGPT and reusing data for lexicographic purposes. *Lexikos*, 34(1), 123–140. https://doi.org/10.5788/34-1-1883

Fuertes-Olivera, P. A. & Tarp, S. (2025). Innovative lexicography: New products and methods in the generative AI age. *Revista Internacional de Lenguas Extranjeras/International Journal of Foreign Languages*, (24), 71–95. https://doi.org/10.17345/rile24.4164

Furnham, A. & Boo, H. C. (2011). A literature review of the anchoring effect. *The Journal of Socio-Economics*, *40*(1), 35–42. https://doi.org/10.1016/j.socec.2010.10.008

Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Dernoncourt, F., Yu, T., Zhang, R. & Ahmed, N. K. (2024). Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3), 1097–1179. https://doi.org/10.1162/coli_a_00524

Hanks, P. (2012). Corpus evidence and electronic lexicography. In S. Granger & M. Paquot (Eds.), *Electronic Lexicography* (pp. 57–82). Oxford University Press. https://doi.org/10.1093/acprof:oso/9780199654864.003.0004

Heer, J. (2019). Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences*, 116(6), 1844–1850. https://doi.org/10.1073/pnas.1807184115

Jakubíček, M., Měchura, M., Kovář, V. & Rychlý, P. (2018). Practical post-editing lexicography with Lexonomy and Sketch Engine. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (Eds.), *The XVIII EURALEX International Congress: Lexicography in Global Contexts, 17–21 July 2018, Ljubljana, Book of Abstracts* (pp. 65–67). Ljubljana University Press, Faculty of Arts. https://doi.org/10.4312/9789610600954

Jiménez-Crespo, M. A. (2025a). Human-centered AI and the future of translation technologies: What professionals think about control and autonomy in the AI era. *Information*, *16*(5), 387. https://doi.org/10.3390/info16050387

Jiménez-Crespo, M. A. (2025b). "If students translate like a robot . . ." or how research on human-centered AI and intelligence augmentation can help realign translation education. *The Interpreter and Translator Trainer*, 19(3–4), 277–295. https://doi.org/10.1080/1750399X.2025.2542022

Kalluri, P. (2020). Don't ask if artificial intelligence is good or fair, ask how it shifts power. *Nature*, *583*(7815), 169–169. https://doi.org/10.1038/d41586-020-02003-2

Kaplan, S. M. (2020). *A theoretical model for the preparation of an inclusive and bias-free expression dictionary* [PhD dissertation, Stellenbosch University]. https://doi.org/10.13140/RG.2.2.28882.09928

Klingbeil, A., Grützner, C. & Schreck, P. (2024). Trust and reliance on AI — An experimental study on the extent and costs of overreliance on AI. *Computers in Human Behavior*, *160*, 108352. https://doi.org/10.1016/j.chb.2024.108352

Kosem, I., Gantar, P., Arhar Holdt, Š., Gapsa, M., Zgaga, K. & Krek, S. (2024). AI in lexicography at the University of Ljubljana: Case studies. In S. Krek (Ed.), *Book of Abstracts of the Workshop Large Language Models and Lexicography* (pp. 29–32). https://www.cjvt.si/wp-content/uploads/2024/10/LLM-Lex_2024_Book-of-Abstracts.pdf

Krek, S. (2019). Natural language processing and automatic knowledge extraction for lexicography. *International Journal of Lexicography*, 32(2), 115–118. https://doi.org/10.1093/ijl/ecz013

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, 11(1), 426. https://doi.org/10.1057/s41599-024-02889-7

McIntosh, T. R., Susnjak, T., Liu, T., Watters, P. & Halgamuge, M. N. (2024). The inadequacy of reinforcement learning from human feedback—Radicalizing large language models via semantic vulnerabilities. *IEEE Transactions on Cognitive and Developmental Systems*, 16(4), 1561–1574. https://doi.org/10.1109/TCDS.2024.3377445

Moorkens, J. (2020). "A tiny cog in a large machine": Digital Taylorism in the translation industry. *Translation Spaces*, 9(1), 12–34. https://doi.org/10.1075/ts.00019.moo

O'Brien, S. (2024). Human-centered augmented translation: Against antagonistic dualisms. *Perspectives*, 32(3), 391–406. https://doi.org/10.1080/0907676X.2023.2247423

Oviatt, S. (2021). Technology as infrastructure for dehumanization: Three hundred million people with the same face. *Proceedings of the 2021 International Conference on Multimodal Interaction* (pp. 278–287). https://doi.org/10.1145/3462244.3482855

Ozmen Garibay, O., Winslow, B., Andolina, S., Antona, M., Bodenschatz, A., Coursaris, C., Falco, G., Fiore, S. M., Garibay, I., Grieman, K., Havens, J. C., Jirotka, M., Kacorri, H., Karwowski, W., Kider, J., Konstan, J., Koon, S., Lopez-Gonzalez, M., Maifeld-Carucci, I., … & Xu, W. (2023). Six human-centered artificial intelligence grand challenges. *International Journal of Human–Computer Interaction*, 39(3), 391–437. https://doi.org/10.1080/10447318.2022.2153320

Poibeau, T. (2025). *Understanding conversational AI: Philosophy, ethics, and social impact of large language models*. Ubiquity Press. https://doi.org/10.5334/bde

Ravichander, A., Fisher, J., Sorensen, T., Lu, X., Lin, Y., Antoniak, M., Mireshghallah, N., Bhagavatula, C. & Choi, Y. (2025). Information-guided identification of training data imprint in (proprietary) large language models. In L. Chiruzzo, A. Ritter & L. Wang (Eds.), *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 1962–1978). https://doi.org/10.18653/v1/2025.naacl-long.99

Rees, G. P., & Lew, R. (2024). The effectiveness of OpenAI GPT-generated definitions versus definitions from an English learners' dictionary in a lexically orientated reading task. *International Journal of Lexicography*, 37(1), 50–74. https://doi.org/10.1093/ijl/ecad030

Rigouts Terryn, A. & de Lhoneux, M. (2024). Exploratory study on the impact of English bias of generative large language models in Dutch and French. In S. Balloccu, A. Belz, R. Huidrom, E. Reiter, J. Sedoc & C. Thompson (Eds.), *Proceedings of the Fourth Workshop on Human Evaluation of NLP Systems (HumEval) @LREC-COLING 2024* (pp. 12–27).

San Martín, A. (2022). Contextual Constraints in Terminological Definitions. *Frontiers in Communication*, *7*. https://doi.org/10.3389/fcomm.2022.885283

San Martín, A. (2025). *Toward human-centered AI-assisted terminology work*. arXiv. https://doi.org/10.48550/arxiv.2512.18859

Schmager, S., Pappas, I. & Vassilakopoulou, P. (2023). Defining human-centered AI: A comprehensive review of HCAI literature. In M. Holkkola, J. Nyrhinen, M. Makkonen, L. Frank, H. Karjaluoto & T.-A. Wilska (Eds.), *MCIS 2023: Proceedings of the 15th Mediterranean Conference on Information Systems* (p. 13). https://aisel.aisnet.org/mcis2023/13

Sennrich, K. & Ahmadi, S. (2025). Conversational lexicography: Querying lexicographic data on knowledge graphs with SPARQL through natural language. In M. Alam, A. Tchechmedjiev, J. Gracia, D. Gromann, M. Pia di Buono, J. Monti & M. Ionov (Eds.), *Proceedings of the 5th Conference on Language, Data and Knowledge* (pp. 289–300). https://aclanthology.org/2025.ldk-1.29/

Shneiderman, B. (2020). Human-centered artificial intelligence: Three fresh ideas. *AIS Transactions on Human-Computer Interaction*, 109–124. https://doi.org/10.17705/1thci.00131

Sourati, Z., Ziabari, A. S. & Dehghani, M. (2026). The homogenizing effect of large language models on human expression and thought. *Trends in Cognitive Sciences*. https://doi.org/10.1016/j.tics.2026.01.003

Stanford Institute for Human-Centered AI. (2025). *About* [Stanford HAI]. https://hai.stanford.edu/about

Tao, Y., Viberg, O., Baker, R. S. & Kizilcec, R. F. (2024). Cultural bias and cultural alignment of large language models. *PNAS Nexus*, 3(9), 1–9. https://doi.org/10.1093/pnasnexus/pgae346

Tarp, S. (2015). Structures in the communication between lexicographer and programmer: Database and interface. *Lexicographica*, 31, 217–245. https://doi.org/10.1515/lexi-2015-0011

Tarp, S. (2019). Connecting the dots: Tradition and disruption in lexicography. *Lexikos*, 29(1), 224–249. https://doi.org/10.5788/29-1-1519

Tarp, S. & Gouws, R. (2020). Reference skills or human-centered design: Towards a new lexicographical culture. *Lexikos*, 30(1), 470–498. https://doi.org/10.5788/30-1-1600

Tiberius, C., Kallas, J., Koeva, S., Langemets, M. & Kosem, I. (2024). A lexicographic practice map of Europe. *International Journal of Lexicography*, 37(1), 1–28. https://doi.org/10.1093/ijl/ecad023

Vallor, S. (2024a). Defining human-centered AI. In C. Régis, J.-L. Denis, M. L. Axente & A. Kishimoto (Eds.), *Human-Centered AI* (pp. 13–20). Chapman and Hall/CRC. https://doi.org/10.1201/9781003320791-3

Vallor, S. (2024b). *The AI mirror: How to reclaim our humanity in an age of machine thinking*. Oxford University Press. https://doi.org/10.1093/oso/9780197759066.001.0001

Vieira, L. N. (2019). Post-editing of machine translation. In M. O'Hagan (Ed.), *The Routledge Handbook of Translation and Technology* (pp. 319–336). Routledge. https://doi.org/10.4324/9781315311258-19

Wallat, J., Jatowt, A. & Anand, A. (2024). Temporal blind spots in large language models. *Proceedings of the 17th ACM International Conference on Web Search and Data Mining* (pp. 683–692). https://doi.org/10.1145/3616855.3635818

Winner, L. (2020). *The whale and the reactor: A search for limits in an age of high technology* (2nd ed.). The University of Chicago Press.

Zhang, C. & Magerko, B. (2025). *Generative AI literacy: A comprehensive framework for literacy and responsible use*. arXiv. https://doi.org/10.48550/arxiv.2504.19038

Zhao, B., Brumbaugh, Z., Wang, Y., Hajishirzi, H. & Smith, N. (2024). Set the clock: Temporal alignment of pretrained language models. *Findings of the Association for Computational Linguistics ACL 2024*, 15015–15040. https://doi.org/10.18653/v1/2024.findings-acl.892

Zhao, H., Chen, H., Yang, F., Liu, N., Deng, H., Cai, H., Wang, S., Yin, D. & Du, M. (2024). Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2), 1–38. https://doi.org/10.1145/3639372

Contact information

Antonio San Martín
University of Quebec in Trois-Rivières
antonio.san.martin.pizarro@uqtr.ca

Catherine Trekker
University of Quebec in Trois-Rivières
catherine.trekker-seguin@uqtr.ca

Acknowledgements

This research was funded by Canada's Social Sciences and Humanities Research Council (grant number 430-2023-0248), Quebec Research Fund (grant 366841), and Spain's Ministry of Science and Innovation (grant number PID2020-118369GB-I00).

Jesús Torres-del-Rey & M.ª Teresa Fuentes-Morán

A HYBRID APPROACH TO WEB ACCESSIBILITY EVALUATION IN DIGITAL LEXICOGRAPHY

Comparing and Integrating Rule-Based Validators and Generative AI

Abstract Electronic dictionaries pose particular accessibility challenges for blind screen reader users, yet the field lacks evaluation frameworks that account for the complexity of their information structures. This paper reports on a three-phase study comparing Automated Web Accessibility Evaluation Tools and Large Language Models in evaluating three online dictionary entries from different lexicographical traditions. Phase 1 establishes a baseline through automated WAET scanning and zero-shot LLM auditing. Phase 2 compares their outputs and derives a four-category framework, reframing compliance as a design question. Phase 3 tests two enhanced prompting strategies — literature-augmented and design-oriented. The former uses specific dictionary accessibility literature while the latter combines a structured checklist derived from Phase 1 with persona-based use-type simulation, surfacing a macro-to-micro hierarchy of dictionary-specific barriers. The study argues for an iterative, multi-condition hybrid workflow integrating WAETs, LLMs, and expert human review for e-dictionaries.

Keywords Web accessibility evaluation; electronic dictionaries; large language models; artificial intelligence; WCAG; screen readers; blind users; dictionary microstructure; automatic validators

1 Introduction

Online (web-based) dictionaries have made lexicographical knowledge available to broader audiences. However, web design also creates new barriers for users with disabilities (Botelho, 2021), and digital dictionaries are no exception. In fact, the complexity of their information structures makes them particularly difficult to accommodate within general web accessibility guidelines (Arias-Badia & Torner, 2024). Yet the information dictionaries carry is essential to communication across all its linguistic dimensions, from the formal and semantic to the pragmatic. It is therefore imperative that dictionaries are presented in as accessible a form as possible — a goal that calls for research combining technological and textual design, lexicographical knowledge, and user-centred studies (Torres-del-Rey & Fuentes Morán, 2024).

Web accessibility is currently governed by the Web Content Accessibility Guidelines (WCAG) 2.2. Automated Web Accessibility Evaluation Tools (WAETs) such as WAVE, Axe DevTools, and AChecker reliably detect only approximately 30% of WCAG 2.2 AA success criteria (Abascal et al., 2019; Bassi et al., 2025). Beyond coverage rates, actual accessibility depends on semantic, pragmatic, and contextual information for which WAETs and strict adherence to standards show important limitations (Abou-Zahra & Brewer, 2019; Power et al., 2012), as they lack the reasoning required for semantic and structural interpretation (Huda et al., 2025).[1]

[1] Most barriers therefore require manual evaluation by experts and users with disabilities, a process that is often financially and logistically prohibitive for many lexicographical projects (Abascal et al., 2019; Mateus et al., 2021).

Recent research has explored Large Language Models (LLMs) in order to bridge the gap between technical compliance and actual accessibility (Aljedaani & Mollik, 2026). Fathallah et al. (2025) proposed a tripartite taxonomy classifying violations into Syntactic (detectable via parsing), Semantic (requiring contextual interpretation), and Layout (necessitating visual or spatial reasoning). While WAETs dominate the syntactic realm, they can also address basic layout issues when integrated with tools like Playwright (ibid.).[2] Conversely, multimodal LLMs demonstrate superior performance in evaluating semantic criteria traditionally requiring manual intervention — such as the appropriateness of non-text content, link purpose, or the language of parts (Fathallah et al., 2025; Gu et al., 2026; López-Gil & Pereira, 2025; Morado Vázquez & Rodríguez Vázquez, 2026) — as well as interpreting complex visual sensory characteristics, logical reading order, hierarchical headings, and page landmarks or regions (He et al., 2025; Kumar et al., 2025; Paternò et al., 2025).

The efficacy of LLMs is, however, critically dependent on prompt engineering (Abu Doush et al., 2023; Huang et al., 2024; Morado Vázquez & Rodríguez Vázquez, 2026; Paternò et al., 2025). The dependency is compounded by a structural bias in training data: with 95.9% of the top one million homepages failing WCAG standards (WebAIM, 2026), models learn predominantly from inaccessible patterns and may therefore fail to identify common violations or hallucinate compliance (Aljedaani et al., 2024). Key strategies include role-play prompting, task decomposition, dynamic contextualisation with relevant WCAG guidelines (Fathallah et al., 2025; Gu et al., 2026; He et al., 2025), and iterative or corrective prompting loops (Abu Doush et al., 2023). This points to the need for a multi-tiered evaluation framework (Kumar et al., 2025) integrating, in hybrid pipelines, automated tools, LLM-driven semantic analysis, and expert human review — a need corroborated by a recent systematic review of 33 studies showing that existing LLM-based approaches concentrate overwhelmingly on generic web content (Aljedaani & Mollik, 2026).

This human-AI partnership, positioning LLMs as copilots that assist human auditors through cross-modal reasoning and high-effort remediation tasks, is particularly relevant for the *information programme* of e-dictionaries (Arias-Badia & Torner, 2024; Torres-del-Rey & Fuentes Morán, 2024). Previous research on accessible lexicography (Rees, 2023, 2025; Torres del Rey & García Garmendia, 2026) has pointed out the limitations of rule-based Web Accessibility Evaluation in lexicography. Whilst WAETs provide a necessary technical baseline, the literature suggests that LLMs offer the interpretive depth required to evaluate the functional affordances of complex digital structures — a shift supporting a more inclusive auditing paradigm in specialised domains such as digital lexicography.

2 Methodology

This research adopts an incremental, multi-layered methodology to evaluate the accessibility of digital lexicographical resources through a combination of traditional automated testing, human-AI hybrid auditing, and the emerging capabilities of LLMs. It follows a functionalist and design-based perspective, prioritising both the user's ultimate goal — efficient access to lexicographical information — and lexicographical affordances — offering rich, interconnected information structures to guide and expand the user's original query and scope — over mere technical compliance.

Analysis was limited to the dictionary entries for *say* (*Merriam-Webster*, MW), *decir* (*Diccionario de la lengua española*, DLE), and *sagen* (*Digitales Wörterbuch der deutschen Sprache*, DWDS),

[2] A browser automation library capturing the live rendered state of the page, and also allowing for dynamic integration with LLMs (Fathallah et al., 2025).

focusing on blind screen reader users and on two dimensions of the entry — interactive elements and information hierarchy. Given the small number of entries analysed and the l ack, at this stage, of validation with blind users, the results reported below should be read as exploratory, and as a springboard for such user testing. All data, full prompts, inputs, outputs and full analyses are openly available at Zenodo.[3] The methodology is structured into three phases:

1. **Phase 1: Initial Multi-Tool Assessment and Baseline Testing.** On the one hand, we deploy industry-standard rule-based WAETs: WAVE (WebAIM) and Axe DevTools (Deque).[4] In addition, we incorporate Intelligent Guided Tests (IGTs) from the Pro version of Axe DevTools, hereafter Axe-G. On the other hand, following a zero-shot, blind prompting strategy, LLM models are provided with web content without explicit information about its lexicographical nature, assessing their raw capability to detect accessibility barriers from their training corpus and multimodal inference alone. Three leading LLMs are benchmarked: Gemini 3.1 Reasoning (Google), Claude Sonnet 4.6 (Anthropic) and ChatGPT 5.3 Instant, Free Tier (OpenAI). Each model receives three complementary inputs per entry, generated by Playwright: the live DOM (a hierarchical representation of the rendered HTML code), the accessibility tree (a11y tree, see Fig. 2), and a high-resolution screenshot. Our aim is not to catalogue all possible error types but to assess complementarity and overlap between WAETs and LLMs, and to identify accessibility patterns relevant to lexicographical design.

2. **Phase 2: Comparative Evaluation.** The second phase compares selected WAET and LLM outputs systematically, and discusses key implications. Issues are classified by WCAG 2.2 success criterion; and a complementary, more lexicographically-oriented design taxonomy is also derived in accordance with our goals and to enhance the next steps in the assessment.

3. **Phase 3: Hybrid Strategies and Iterative Dialogue.** The third phase proposes procedures for dialogical refinement of accessibility evaluation in e-dictionaries. Building on the lexicographical findings of Phase 2, we present and evaluate two enhanced prompting strategies — knowledge-augmented and design-oriented prompting — and propose a hybrid evaluation workflow grounded in dictionary design.

3 Comparison of WAET and zero-shot LLM evaluation results

3.1 Overview: Coverage by WCAG principle and tool category

Table 4 (Section 5) provides a cross-condition overview of detection rates across all WCAG 2.2 success criteria and all four evaluation conditions from Phase 1 through Phase 3 of this study — over a full dataset of 78 valid issues. The results discussed in this section correspond to the WAETs and LLMs (0-shot) columns alone.[5]

[3] https://doi.org/10.5281/zenodo.20422503.

[4] "Errors" and "Alerts" were examined in WAVE; "Problems", "Advanced Problems," and "Best Practices," in Axe.

[5] Table 4 covers all four evaluation conditions (phases 1 to 3) for comparability across sections; specific Phase 1-only figures (50 unique issues) and percentages, calculated over that subset, are available on Zenodo.

LLMs outperform WAETs in eleven of thirteen criteria. Nine criteria favour LLMs strongly (≥30 pp) and two moderately. Two criteria strongly favour WAETs: 2.1.1 Keyboard and 2.4.4/2.4.9 Link Purpose (both +40 pp). The disparity is starkest at the principle level: in Understandable (Principle 3), WAETs detect no issues whereas LLMs detect 4 issues; in Perceivable (Principle 1), WAETs identify 11 issues versus 22 for LLMs.

3.2 Where WAETs excel: Syntactic reliability and interaction testing

WAETs demonstrate strong and consistent reliability for issues that map onto deterministic, rule-checkable HTML markup and ARIA attributes.[6] The clearest domain of WAET strength is 2.1.1 where the full WAET stack detects all five Phase-1 issues whilst LLMs capture only three. WAETs can interact with JavaScript — the code that drives a page's dynamic, on-screen behaviour — to detect mouse-only events and unfocusable elements:[7] a capacity that LLMs, if provided only with DOM snapshots and screenshots, lack (Gur et al., 2023).

A relevant exception within this criterion is the treatment of abbreviation tooltips (see Fig. 3a):[8]P WAETs do not flag them as violations, since assistive technologies can in principle read the title attribute that carries the expansion. However, accessing the tooltip requires either mouse hover (a mouse-only event) or, sometimes, an explicit screen reader command — a practical barrier for keyboard-only and novice users that WAETs overlook but LLMs *consistently* identify as an accessibility concern, even when the underlying code is technically conformant.

Four of these five detections are due to Axe-G, which contributes detections that WAVE and standard Axe miss: DLE's unfocusable clickable elements and mouse-only abbreviation tooltips, DWDS's mouse-only abbreviation and inflection tooltips (Fig. 1), and MW's mouse-only relevance tooltip (Fig. 2b). This confirms that guided tools serve as a non-redundant bridge between static automated scanning and full manual auditing, specifically for interaction-layer failures. However, the guided workflow requires substantially greater time investment than automated scanning, and its heavy reliance on human interpretation limits its scalability.

Beyond keyboard operability, WAETs demonstrate reliable syntactic coverage across all three dictionaries: WAVE and Axe jointly detect formally styled but non-programmatic headings (1.3.1) in MW (Fig. 2) and DWDS; WAVE alone detects a broken link in MW; Axe detects same-link-different-purpose cases in both MW and DWDS (2.4.9), and the absent table caption in DLE's conjugation tables (2.4.6); and WAVE and Axe both detect the CSS-produced underlining in DLE (DLE-1) that has no programmatic correspondence for assistive technologies (1.3.3). In all these cases, the issue has a direct code-level correlate that fires a deterministic rule.

[6] HTML markup refers to the tags that structure a page's code — for instance, a heading or a table caption, both discussed below. A rule-checkable HTML attribute is *alt* for images: WAETs can verify whether it is present, though not whether its wording is adequate. Accessible Rich Internet Applications (ARIA) attributes are additional markers describing an element's role for assistive technology, such as identifying a navigation landmark or marking an element as an interactive button — the kind of attribute that helps WAETs detect the keyboard issues discussed next.

[7] A mouse-only event is one that responds only to a pointing device, not to the keyboard; an element is focusable if a keyboard user can reach it by pressing Tab, and unfocusable if they cannot.

[8] Small pop-up boxes displaying additional information, typically when a user hovers over an element.

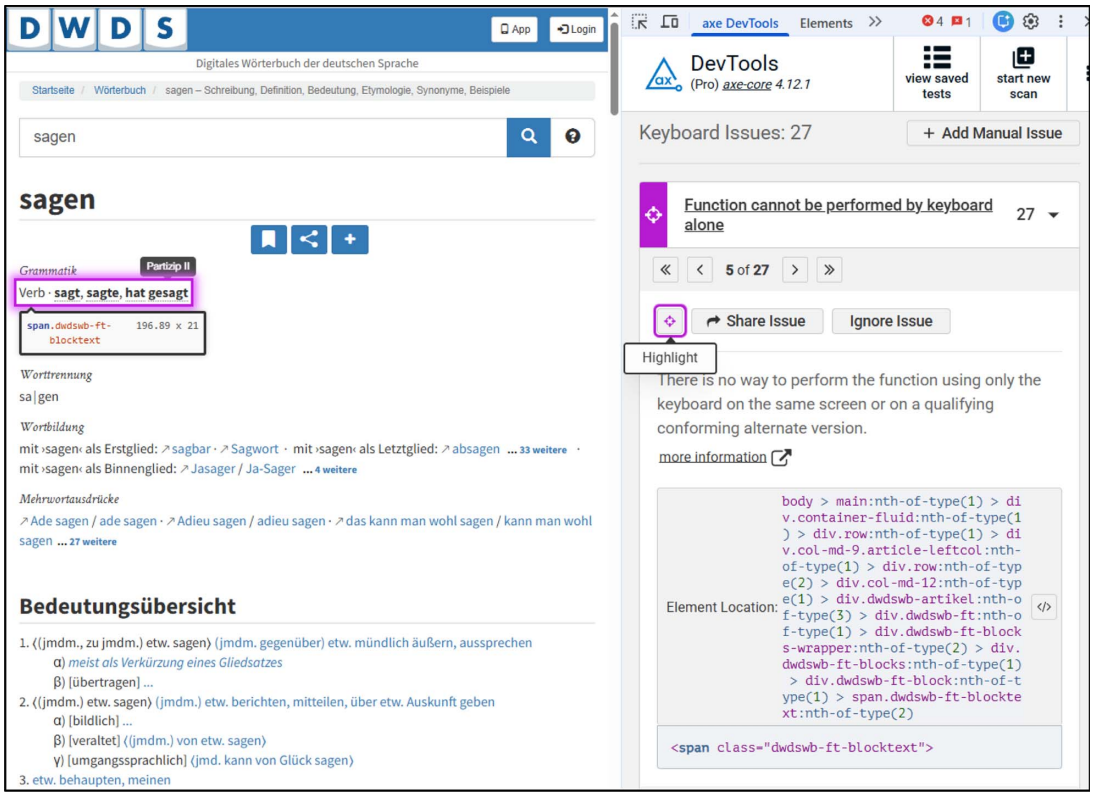


FIGURE 1 Axe-G flags the grammar tooltip for *sagen*'s conjugated forms in DWDS as a keyboard barrier: triggered by mouse hover, with no keyboard-operable equivalent

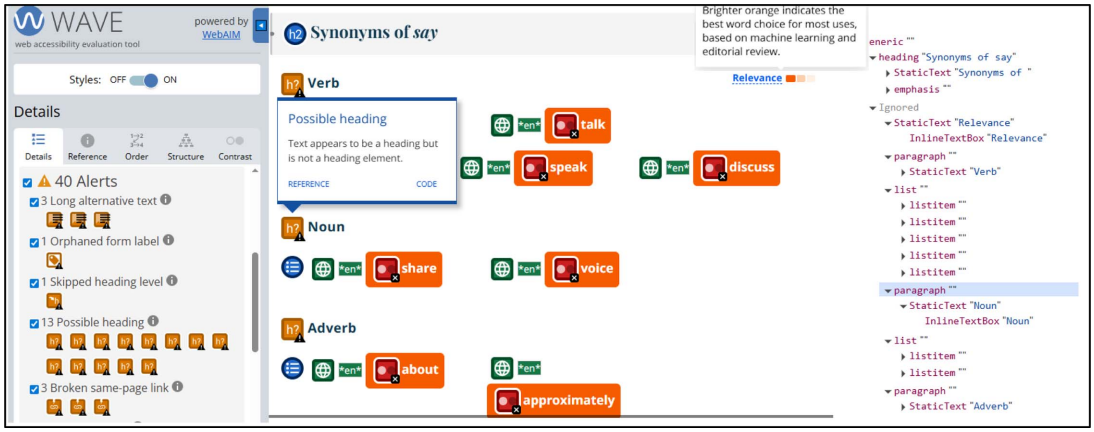


FIGURE 2 WAVE (left/centre) and the Chrome accessibility tree (right), applied to MW's entry for *say*: (a) "Verb", "Noun" and "Adverb" are flagged as possible headings yet carry no heading role; (b) the coloured bars encoding each synonym's relevance score carry no text or alt attribute

3.3 Where LLMs excel: Semantic reasoning and reading comprehension

LLMs achieve substantially higher coverage across criteria that require contextual interpretation. In the lexicographical domain, this distinction is particularly important, as dictionary entries involve specialised text conventions — grammatical abbreviations, etymological notation, sense hierarchies, homograph numbering — that WAETs treat as undifferentiated web content. The most structurally important LLM-dominant criterion is 1.3.1 Info and Relationships (+23 pp), which has the highest number of issues in the Phase-1 round (13). WAETs detect

the structurally verifiable failures (e.g. MW-4's absent main landmark), whilst LLMs detect ten of the thirteen issues, six of them exclusively, all requiring reading "comprehension" — such as DLE-13's unmarked synonym/antonym list (see Fig. 3b). In these six cases the HTML is formally valid; the barrier lies in the communicative logic of the content, not its code.

The criterion 1.3.3 Sensory Characteristics illustrates a characteristic split within a single criterion. WAETs reliably detect the one case where sensory-only communication has a direct code correlate (the DLE-1 underlining case above) but fail entirely on the three cases where identifying the barrier requires pragmatic inference: colour as the sole carrier of meaning distinguishing DLE's different data categories, homograph disambiguation through superscript alone in DLE (Fig. 3), and a word cloud in DWDS conveying meaning through visual layout alone (Fig. 4). All three are detected exclusively by LLMs.

Two criteria show total WAET failure against 100% LLM coverage. For 3.1.2 Language of Parts, both instances of Latin etymological content without a *lang* attribute are detected exclusively by Gemini — this requires recognising that a language shift has occurred, a judgment that presupposes reading comprehension rather than attribute inspection. For 3.1.4 Abbreviations, WAETs detect nothing; the DLE case (DLE-8, excessive grammatical abbreviations, Fig. 3a) is identified by all three LLMs, whilst the DWDS case (DWDS-18, the same abbreviation density, Fig. 1) is detected by ChatGPT alone. This illustrates an accessibility challenge specific to print-derived dictionaries: condensed typographic conventions create cognitive barriers invisible to rule-based tools.

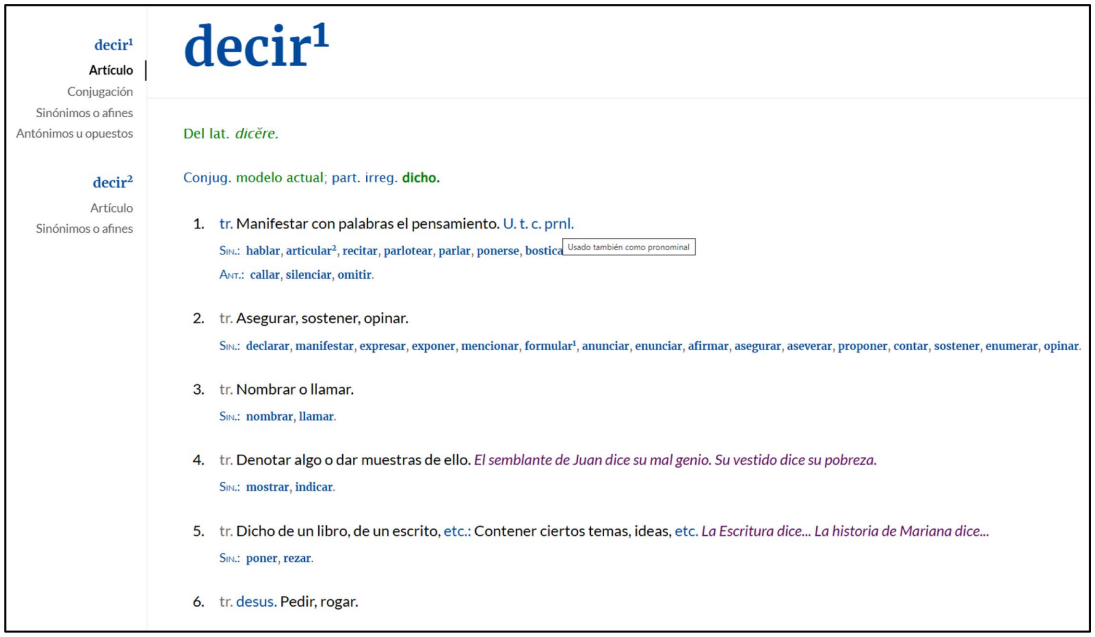


FIGURE 3 DLE's entry for *decir*: (a) grammatical abbreviations ("tr.", "U. t. c. prnl.") stack before each sense, expandable only via a mouse-hover tooltip; (b) the "Sin.:"/"Ant.:" lists present synonyms and antonyms as comma-separated clickable elements, with no structural marker distinguishing the two; (c) the homographs *decir*[1]/*decir*[2] are distinguished only by superscript, read aloud by screen readers with no indication these are separate entries

LLMs also spontaneously identify barriers rooted in visual density: absent or non-descriptive alternative text on complex visual elements (MW's relevance meter — colour-coded score, Fig. 2b — and DWDS's frequency line, geographic map and timeline) is flagged as a failure



FIGURE 4 In DWDS's collocation cloud for *sagen*, font size shows how strongly each word associates with the headword — but only visually: a screen reader user hears the words as plain links, with no sense of which ones matter most

under 1.1.1 or 4.1.2. WAETs detect only cases where the attribute is altogether missing, with no mechanism for judging the adequacy of provided descriptions.

3.4 Comparative profiles: Consistency, domain sensitivity, and qualitative dimension

Whilst LLMs demonstrate broader semantic coverage, their output is characterised by meaningful inconsistency (Andruccioli et al., 2025) — a structural contrast with WAETs, for which a given rule either fires or it does not. Three examples may suffice: for 2.4.1 Bypass Blocks (providing a mechanism to bypass blocks of content that are repeated on multiple web pages), the three LLMs collectively produce ten detections but only five are distinct, meaning half are redundant across models; for 1.1.1 Non-text Content, detections are near-perfectly distributed across models with minimal duplication — the criterion where multiple LLMs add the most genuine incremental value; for 3.1.4, all three LLMs flag DLE's abbreviation overload, but only ChatGPT flags the equivalent in DWDS.

Another significant finding is the degree to which LLMs spontaneously adopt a lexicographical frame of reference without explicit prompting. In the zero-shot session, models were not informed that the pages were dictionary entries, yet consistently identified issues specific to dictionary microstructure: (1) Homograph disambiguation: both Gemini and Claude note, independently and unprompted, that superscript numerals distinguishing homographs in DLE (*decir*[1], *decir*[2]) are announced by screen readers without clarification that these represent distinct lexical entries (Fig. 3c). (2) Cross-reference legibility: LLMs identify in MW that links labelled "see more" or "read more" fail the link-purpose criterion specifically because, within the dictionary microstructure, users need to understand whether a reference leads to a related definition, a synonym, or an extended sense — information the generic label does not convey.

As Table 1 illustrates, WAETs and LLMs are not substitutes but structurally complementary tools whose blind spots are determined by their fundamental operating logics (Fathallah et al., 2025; Paternò et al., 2025): WAETs for deterministic code-level correlates, LLMs for qualitative, semantic interpretation and contextual reading.

TABLE 1 Comparative profile of WAETs and LLMs across key evaluation dimensions

Feature	WAETs	LLMs
Primary focus	Syntactic/programmatic compliance	Semantic/cognitive user experience
Landmarks	Detect duplicate or missing elements	Evaluate the "mental map"; flag disorienting logic
Headings	Identify skipped levels or missing markup	Critique logical consistency of heading hierarchy
Lexicography	Treat content as generic web data	Identify entry-specific issues: superscripts, etymology, abbreviation overload
Visual density	Ignored if contrast/size passes rules	Flagged as cognitive accessibility barrier
Language of parts	Not detected (0%)	Detected by Gemini exclusively (100%)
Hallucination risk	False positives possible on syntactic edge cases	Occasional hallucinations (e.g., non-existent landmark)

3.5 A design-pattern view of dictionary accessibility

The criterion-level analysis in Sections 3.1–3.4 addresses which compliance rules are violated. A complementary view, more directly useful to lexicographers, organises the same issues around the aspects of dictionary use they affect. We propose four categories, as shown in Table 5 (Section 5).[9] This reframes accessibility evaluation from a compliance question into a design question, in line with the *information programme* approach proposed above. A definition of each category is provided now, with figures only considering Phase 1 (50 unique issues identified):

- **Information Grouping, Segmentation and Relations (IGS&R)** is the largest category (16 issues) and the one with most overlap (WAETs 62%, LLMs 75%). The internal split between Landmarks — WAETs leading by 25 pp — and Lists/Tables — clearly LLM-leaning — is itself a microcosm of the syntactic/semantic divide: structural failures with a code-level correlate are caught by WAETs; those requiring interpretation of adequately rendered hierarchy are caught by LLMs.

- **Element Function (EF)** is the dimension most invisible to automated workflows (WAETs 33%, LLMs 100%). It groups features whose accessibility depends not only on the *presence* but primarily on the *adequacy* of identification — alt text on a relevance meter, a language attribute, colour or typographic convention as the sole differentiator between definition and example. This is the clearest evidence that dictionary accessibility cannot be evaluated by rule-based scanning alone.

- **Navigation and Interaction (N&I)** is the only category where both families perform identically (67% each), and it is genuinely complementary: WAETs detect all four link-level failures — empty, broken or same-text/different-destination links (4 of 4) — but miss bypass repeated blocks entirely (0 of 3); LLMs show the inverse pattern. The guided WAET workflow (Axe-G) is uniquely valuable here for detections requiring simulated user interaction.

[9] Subcategories are available on our Zenodo spreadsheet; nonetheless, they are progressively distilled in the four category descriptions provided in this section.

- **Complexity and Noise (C&N)** is small in count (four issues) but qualitatively distinctive (WAETs 0%, LLMs 100%). It captures failures emerging from the density of dictionary content — abbreviation overload, complex conjugation tables, proliferation of cross-references. WAETs are constitutively unable to operate at this level: there is no rule for *too many*. This dimension corresponds most directly to the cognitive-architectural concerns of the *information programme* framework.

Several issues cut across categories: mouse-only abbreviation expansion is both an interaction (N&I) and an identification (EF) failure; complex conjugation tables are both structural (IGS&R) and a cognitive-load problem (C&N). This reinforces the argument that dictionary accessibility failures do not partition along any single technical or design dimension — precisely what a strict WCAG-based audit cannot capture and what evaluation at the level of the *information programme* can.

4 Beyond zero-shot: Enhanced LLM evaluation strategies

4.1 Prompting strategies and tool selection

Phase 2 established that WAETs and LLMs are structurally complementary — WAETs for syntactic, code-level violations; LLMs for semantic barriers requiring contextual interpretation. Phase 3 investigates whether enhanced prompting strategies improve LLM coverage, depth, and reliability, and how iterative dialogue with LLMs can serve dictionary designers and lexicographers.

Two strategies are examined: knowledge-augmented prompting (Section 4.2), providing LLMs with specialised literature via a Gemini Gem and a Claude Project, approximating RAG (Retrieval-Augmented Generation) pipelines (Fathallah et al., 2025; He et al., 2025); and a hybrid structured prompt (Section 4.3) combining a structured decomposition checklist of the four-category framework from Phase 2 with a persona-based scenario simulation across three dictionary use types — comprehension, production, and translation. Both rounds used Gemini and Claude only; ChatGPT was excluded due to subscription constraints.

4.2 Literature-augmented prompting: effects and limits

For the literature-augmented (lit.) round, both LLMs received the same knowledge base: Arias-Badia & Torner (2024), Rees (2023), Torres del Rey & Fuentes Morán (2024), Torres del Rey & García Garmendia (2026), and Vasik & Fuentes Morán (2025). For DLE, two sessions were run: a "Full" session (including Torres del Rey & García Garmendia, 2026 — a DLE-specific user study with blind screen reader users) and a "Without" session (excluding that study), to isolate the effect of dictionary-specific knowledge.[10] Table 2 summarises recovery and discovery rates.

Both DLE sessions recovered the same 11/17 zero-shot issues (65%). The Full session produced three new issues citing the DLE-specific study directly; the Without session produced two different ones not documented in that study — among them DLE-21, section links labelled

[10] Rees (2023) and Arias-Badia & Torner (2024) also evaluated MW and DLE, among other dictionaries, but on earlier site versions. Unlike Torres del Rey & García Garmendia (2026), which is devoted specifically to DLE, their coverage of MW and DLE is version-mismatched with our corpus, so we judged their likely effect on the LLMs' output too small to warrant a parallel Full/Without comparison for MW. A later piece by Rees (2025) synthesises the first two sources above to inform a prototype dictionary design, so it was excluded to avoid overweighting issues and dictionaries already covered, whereas the three sources co-authored by Torres del Rey and/or Fuentes Morán address different dictionaries and aspects.

TABLE 2 Zero-shot issue recovery and new issue discovery in the literature-augmented round

Dictionary	0-shot issues	Recovered (lit.)	Recovery rate	New issues (lit.)
MW	14	6	43%	10
DLE — Full	17	11	65%	3
DLE — Without	17	11	65%	2
DWDS	19	11	58%	6

"Artículo" ("Article") without specifying *decir*[1] or *decir*[2] (Fig. 3c, left column), which the Full session missed. What might be called a *selective amplifier effect* — "Full" increasing depth on documented issues whilst narrowing sensitivity to undocumented ones — means the two conditions are complementary. It follows that knowledge-augmented evaluation should probably be one component of an iterative multi-condition workflow, not a self-sufficient replacement for other approaches (Andruccioli et al., 2025; Fathallah et al., 2025).

Knowledge augmentation enhanced output quality alongside coverage. The two LLMs produced complementary output profiles: Claude's precision, anchored in specific DOM elements and attributes, suited developer-facing remediation; Gemini's more abstract, typological framing — patterns transferable across dictionary types — suited lexicographer-facing design guidance. A follow-up prompt inviting re-categorisation by lexicographical design aspect or information category elicited the latter from both models — grouping findings by phonology, grammar, semantics, and access structure — with Claude consistently more granular (5–10 categories per session) than Gemini's more compact four- or five-category summaries.

A significant finding was a non-standard tooltip-attribute pattern originating in the Bootstrap framework,[11] which recurs across DWDS and is catalogued as a single issue (DWDS-22): that single decision causes multiple inaccessible abbreviation tooltips, inflection labels, and citation descriptions simultaneously (Fig. 1). This systemic pattern was invisible to both WAETs and zero-shot evaluation; detecting it required combining Claude's DOM-reading with the literature's conceptual framework for abbreviation handling. It is the clearest example of knowledge augmentation enabling a qualitatively different class of finding: a root-cause analysis allowing a developer to fix multiple failures with a single intervention, rather than a list of individual violations.

Two limitations merit note. First, WAET-exclusive syntactic violations (broken links, ARIA widget errors, empty anchors) do not reappear in any literature-augmented session. Second, where the knowledge base describes print lexicographical conventions, *knowledge base drag* is a risk: Gemini Full described punctuation symbols used in print DLE editions as present in DLE v23.8 when some are not.

4.3 Hybrid structured prompting: Design categories and persona-based scenarios

The design-oriented (DO) round applied a two-prong prompt to the same three entries. For both prongs, the two LLMs were asked to draw on core concepts from e-dictionary

[11] Bootstrap is a widely used framework for building website interfaces. In DWDS, the tooltip text is stored in a way that only sighted users hovering with a mouse can see it, which was detected by Axe-G and ChatGPT as a mouse-only event (DWDS-15); what is a new discovery (DWDS-22) is that screen readers cannot access the tooltip under any interaction method, making it inaccessible for blind users.

design theory — information programmes, access structures, lexicographical affordances, user typology by consultation purpose, or condensation conventions — so that they would adopt a shared analytical context oriented towards lexicographers and designers rather than generic web accessibility auditors. The first prong — the design-category checklist — used the four lexicographical design categories from Phase 2 as a structured checklist, directing the model through each category and its sub-items systematically. The second prong — the user-scenario simulation — asked the model to simulate three dictionary use types — comprehension, production, and translation — adopting the perspective of a blind screen reader user, activating knowledge about real usage needs. This persona approach is grounded in the design thinking tradition (Horton & Quesenbery, 2013).

The round produced 36 unique issue detections — fewer than zero-shot LLMs (41) — but a different *type* of evidence: issues framed as barriers to specific lexicographical tasks rather than WCAG violations. The two prongs also produced partially non-overlapping findings: the checklist prong recovered most issues already identified in previous rounds, whilst the scenario prong surfaced issues specific to particular use types. The degree of overlap between the two prongs varied by dictionary, reflecting genuine differences in how each dictionary's design challenges map onto user needs.

The two-prong structure of the DO round, read against the findings of the previous conditions, allowed us to distil a consistent macro-to-micro hierarchy of accessibility failures across all three dictionaries — a synthesis no single round had provided, and one directly useful for prioritising remediation. Table 3 presents this hierarchy with barriers described in terms of what a blind user cannot do — the language that resonates with design thinking and with designers who need to understand user impact.

TABLE 3 Macro-to-micro hierarchy of accessibility failures across MW, DLE, and DWDS

Level	Barrier for blind/screen reader users	Detected by	Design priority and lexicographical impact
1. Macro-architecture	No programmatic boundaries between major entry blocks: user cannot tell by heading navigation where numbered senses end and phrasemes, paradigm tables, or synonym sections begin	LLMs (chiefly lit. and DO) — no WAET detections: boundaries are rendered visually, with no code correlate	Highest priority — affects all use types equally. All three dictionaries affected
2. Sense-internal structure	Definition, example, usage note, and sub-sense structurally merged: screen reader reads them as undifferentiated text	LLMs	High — affects comprehension (finding the definition) and production (finding the right example for a sense)
3. Cross-sense navigation	No named anchors on individual senses: translator needing to jump non-linearly between senses must read the entire entry sequentially	LLMs (DO) only — surfaces only via translation scenario	Rated Critical for translation by both LLMs across all three dictionaries

TABLE 3 (*CONTINUED*)

Level	Barrier for blind/screen reader users	Detected by	Design priority and lexicographical impact
4. Micro-element	Abbreviated grammatical labels, phonetic transcriptions, typographic symbols (‖, :), usage badges, cross-reference link types carry meaning without programmatic equivalents	Syntactic instances by WAETs; semantic adequacy by LLMs (all conditions)	Variable — some WAET-detectable (CSS-only underlining, empty alt); most require LLM semantic evaluation

Level 3 — absent sense-level navigation anchors across all three dictionaries (MW-29, and equivalent issues in DLE and DWDS) — carries the most immediate practical implications. This issue was invisible in all previous rounds; it required explicitly simulating the translation scenario to surface, because translation use is constitutively non-linear: a translator needs to jump between specific senses without reading the entire entry. The fix is technically minimal — exposing existing DOM element IDs as navigation targets[12] — but its discovery required the persona-based simulation, although it should be confirmed with real users.

Claude and Gemini diverged in the DO round along the same lines observed in the literature round — Claude producing longer, code-referenced analyses suited to developer-facing remediation, Gemini producing shorter, design-pattern analyses suited to lexicographer-facing guidance — a complementarity that suggests using both in tandem rather than selecting one.

5 Conclusions: A multi-condition, design-oriented evaluation framework

Tables 4 and 5 synthesise detection coverage across the full dataset of 78 valid issues under four evaluation conditions — WAETs, zero-shot LLM prompting (Phase 1), literature-augmented LLM prompting, and design-oriented LLM prompting (Phase 3) — from two complementary perspectives: by WCAG 2.2 success criterion (Table 4) and by lexicographical design category (Table 5). This dual perspective is a key output for dictionary designers and developers: Table 4 shows which condition detects which barrier from a standards-compliance angle, while Table 5 reframes the same coverage around lexicographical design categories, supporting teams who approach accessibility from a more holistic, design-led perspective.

Three patterns emerge from the data, summarised in Tables 4 and 5. First, no single condition achieves consistent coverage across all categories. Second, the literature-augmented round produces the highest total unique detections (49 vs. 41 for 0-shot LLMs). Third, the design-oriented round detects fewer unique issues than other LLM conditions (36) but more than WAETs (25), contributes 7 issues detected by no other condition, and is the only condition surfacing Level 3 cross-sense navigation failures. These observations jointly make the case for iterative multi-condition evaluation.

Despite WAETs' lower overall coverage, their visual interface (see Fig. 1 and Fig. 2) — highlighting exact issue locations and linking directly to WCAG documentation — remains

[12] Each sense already carries a unique identifier in the page's underlying code; the fix consists of turning these into clickable anchors, without any redesign of the entry itself.

TABLE 4 Detection rates by WCAG 2.2 success criterion and evaluation condition (n = issues per criterion; percentages within criterion; issues may be assigned to multiple criteria)[13]

WCAG 2.2 criterion	n	WAETs	LLMs 0-shot	LLMs lit.	LLMs DO
1. Perceivable					
1.1.1 Non-text Content	8	25%	75%	75%	38%
1.3.1 Info & Relationships	26	27%	38%	62%	58%
1.3.2 Meaningful Sequence ⚠	2	0%	50%	50%	0%
1.3.3 Sensory Characteristics	8	12%	50%	100%	38%
1.3.6 Identify Purpose	3	33%	67%	67%	33%
2. Operable					
2.1.1 Keyboard	6	83%	50%	67%	33%
2.4.1 Bypass Blocks	9	22%	56%	56%	33%
2.4.3 Focus Order	5	0%	40%	60%	20%
2.4.4/2.4.9 Link Purpose	9	44%	22%	44%	22%
2.4.6 Headings & Labels	5	40%	60%	80%	60%
2.4.10 Section Headings ⚠	1	0%	0%	0%	100%
3. Understandable					
3.1.2 Language of Parts ⚠	2	0%	100%	100%	100%
3.1.3 Unusual Words ⚠	1	0%	0%	100%	0%
3.1.4 Abbreviations	3	0%	67%	100%	100%
4. Robust					
4.1.2 Name, Role, Value	12	42%	67%	58%	42%
Total (unique issues)	**78**	**32%**	**53%**	**63%**	**46%**

TABLE 5 Cross-condition coverage matrix: percentage of unique issues in each design category detected by each evaluation condition; each issue is assigned to a single primary category; the overlaps discussed in Section 3.5 are conceptual

Design category (n issues)	WAETs	LLMs 0-shot	LLMs lit.	LLMs DO
Information Grouping, Segmentation & Relations (n=26)	38%	46%	54%	50%
Element Function (n=25)	20%	60%	84%	48%
Navigation and Interaction (n=21)	48%	48%	48%	33%
Complexity and Noise (n=6)	0%	67%	67%	67%
TOTAL (n=78)	**32%**	**53%**	**63%**	**46%**

indispensable for developer-facing reports. For syntactic and interaction-layer barriers in particular, their rule-based checks offer a reliability that LLM judgement cannot match: a pass or fail is verifiable against the code itself. LLMs complement this with semantic depth: explaining *why* an issue matters lexicographically, *what* information a blind user cannot access,

[13] Shading: dark grey ≥ 60%, mid grey 30–59%, no fill < 30% (within-criterion percentages). ⚠ = fewer than 3 issues.

and — going beyond the generic WCAG-based remediation guidance WAETs already provide — *how* it can be fixed and *who* is specifically affected. A practical hybrid workflow could proceed in three loops: automated WAET scanning establishes a syntactic baseline; structured LLM evaluation adds semantic coverage and designer-oriented explanations; and expert review filters false positives at every stage, retaining ultimate accountability for accessibility decisions.

The four-category framework, complemented by the macro-to-micro hierarchy, provides a shared vocabulary for developers and lexicographers. The identification of the three sub-dimensions for the IGS&R category (previously, "Chunking and Relations") was due to Phase 3 dialogic refinement: *grouping* (whether content blocks are marked as discrete units), *segmentation* (whether boundaries between adjacent blocks are explicitly delimited), and *relations* (whether hierarchical dependencies are expressed programmatically). These sub-dimensions allow designers to prioritise by type of structural failure.

Several caveats apply. This is a study of three entries across three dictionaries, and no direct testing with blind screen reader users was conducted; the findings should therefore be read as exploratory rather than confirmatory, pending replication with broader coverage of dictionary types, entries, and entry structures. Some of the issues identified are best understood as best-practice recommendations rather than confirmed accessibility barriers, and require validation through real-user studies with blind screen reader users before being adopted as design requirements. LLM outputs are also inherently variable across sessions; performance may be influenced by differences in subscription tiers, prior interaction history within a session, and frequent model updates. We have attempted to mitigate these risks through consistent prompt design, multiple evaluation rounds, and disabling session memory features, but session-level variability remains an inherent limitation of the LLM conditions, underscoring why expert review remains responsible for final accessibility judgements. Full data, prompts, and recorded sessions are available on Zenodo for replication and extension.

References

Abascal, J., Arrue, M., & Valencia, X. (2019). Tools for web accessibility evaluation. In Y. Yesilada & S. Harper (Eds.), *Web accessibility: A foundation for research* (pp. 479–503). Springer.

Abou-Zahra, S., & Brewer, J. (2019). Standards, guidelines, and trends. In Y. Yesilada & S. Harper (Eds.), *Web accessibility: A foundation for research* (pp. 225–246). Springer.

Abu Doush, I., Sultan, K., Al-Betar, M. A., Almeraj, Z., Alyasseri, Z. A. A., & Awadallah, M. A. (2023). Web accessibility automatic evaluation tools: To what extent can they be automated? *CCF Transactions on Pervasive Computing and Interaction*, 5(3), 288–320.

Aljedaani, W., Habib, A., Aljohani, A., Eler, M., & Feng, Y. (2024). Does ChatGPT generate accessible code? Investigating accessibility challenges in LLM-generated source code. *Proceedings of the 21st International Web for All Conference, W4A '24*, 165–176.

Aljedaani, W., & Mollik, R. H. (2026). Large language models for web accessibility: A systematic literature review. *Proceedings of the Web4All 2026 23rd International Web for All Conference (W4A '26)*.

Andruccioli, M., Bassi, B., Delnevo, G., & Salomoni, P. (2025). Leveraging large language models for sustainable and inclusive web accessibility. *Big Data and Cognitive Computing*, 9(10).

Arias-Badia, B., & Torner, S. (2024). Bridging the gap between website accessibility and lexicography: Information access in online dictionaries. *Universal Access in the Information Society*, 23(2), 545–560.

Bassi, B., Delnevo, G., Franco, M., Gaggi, O., Gatto, S., Mirri, S., & Olaiya, K. (2025). An assessment of LLM-based auditing and validation for web accessibility. *Proceedings of the 2025 International Conference on Information Technology for Social Good, GoodIT '25*, 297–305.

Botelho, F. H. F. (2021). Accessibility to digital technology: Virtual barriers, real opportunities. *Assistive Technology*, 33(sup1), 27–34.

Fathallah, N., Hernández, D., & Staab, S. (2025). AccessGuru: Leveraging LLMs to detect and correct web accessibility violations in HTML code. *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS '25*, 1–22.

Gu, M., Wang, Z., Lai, S., Gao, Z., Zhou, S., & Bu, J. (2026). Towards scalable web accessibility audit with MLLMs as copilots. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(45), 38515–38523.

Gur, I., Nachum, O., Miao, Y., Safdari, M., Huang, A., Chowdhery, A., Narang, S., Fiedel, N., & Faust, A. (2023). Understanding HTML with large language models. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 2803–2821.

He, Z., Huq, S. F., & Malek, S. (2025). Enhancing web accessibility: Automated detection of issues with generative AI. *Proceedings of the ACM on Software Engineering*, 2(FSE), 2264–2287.

Horton, S., & Quesenbery, W. (2013). *A web for everyone: Designing accessible user experiences*. Rosenfeld Media.

Huang, C., Ma, A., Vyasamudri, S., Puype, E., Kamal, S., Garcia, J. B., Cheema, S., & Lutz, M. (2024). *ACCESS: Prompt engineering for automated web accessibility violation corrections*. ICLR 2024.

Huda, M., Setyanto, A., Kusrini, K., & Wahyuni, S. N. (2025). Web accessibility evaluation in the AI era: A systematic review on conventional tools and the emerging use of large language models (LLMs). *2025 10th International Conference on Information Technology and Digital Application (ICITDA)*, 1–7.

Kumar, A., Padath, T., & Wang, L. L. (2025). Benchmarking PDF accessibility evaluation: A dataset and framework for assessing automated and LLM-based approaches for accessibility testing. *Proceedings of the 27th International ACM SIGACCESS Conference on Computers and Accessibility, ASSETS '25*, 1–24.

López-Gil, J.-M., & Pereira, J. (2025). Turning manual web accessibility success criteria into automatic: An LLM-based approach. *Universal Access in the Information Society*, *24*(1), 837–852.

Mateus, D. A., Silva, C. A., Oliveira, A. F. B. A. de, Costa, H., & Freire, A. P. (2021). A systematic mapping of accessibility problems encountered on websites and mobile apps: A comparison between automated tests, manual inspections and user evaluations. *Journal on Interactive Systems*, 12(1), 145–171.

Morado Vázquez, L., & Rodríguez Vázquez, S. (2026). Cuando las palabras necesitan etiqueta: Accesibilidad y uso del atributo lang en los diccionarios en línea. In F. Domínguez Santana & C. Travieso Rodríguez (Eds.), *Uso de los diccionarios: Avances metodológicos y enfoques inclusivos en los estudios de usuarios*. Tirant lo Blanch.

Paternò, F., Vinci, M., Manca, M., & Iannuzzi, N. (2025). How an LLM can improve automatic web accessibility validation? *Proceedings of the 16th Biannual Conference of the Italian SIGCHI Chapter, CHItaly '25*, 1–8.

Power, C., Freire, A., Petrie, H., & Swallow, D. (2012). Guidelines are only half of the story: Accessibility problems encountered by blind users on the web. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems, CHI '12*, 433–442.

Rees, G. P. (2023). Online dictionaries and accessibility for people with visual impairments. *International Journal of Lexicography*, 36(2), 107–132.

Rees, G. P. (2025). Making dictionary content accessible for people with visual impairments. *Lexikos*, 35(2), 165–184.

Torres del Rey, J., & García Garmendia, M. (2026). Más allá de las pautas de accesibilidad: Estudio de interacción de usuarios ciegos mediante lector de pantalla con el Diccionario de la lengua española, versión 23.8 en línea. In F. Domínguez Santana & C. Travieso Rodríguez (Eds.), *Uso de los diccionarios: Avances metodológicos y enfoques inclusivos en los estudios de usuarios*. Tirant lo Blanch.

Torres-del-Rey, J., & Fuentes Morán, M. [a] T. (2024). PRELemma: Por unos recursos léxicos multilingües accesibles. In I. Lozano & A. B. Sánchez-García (Eds.), *Una educación inclusiva y de calidad: Ideas y estrategias para seguir avanzando* (pp. 397–416). Universidad de Salamanca.

Vasik, Y., & Fuentes Morán, M. ª T. (2025). Monolingual lexicography and visual impairment: Parameters for more accessible dictionaries. *1st International Conference on Lexicology and Lexicography*. LexiKonf 2025, Budapest.

WebAIM. (2026). *The WebAIM million. The 2026 report on the accessibility of the top 1,000,000 home pages*. https://webaim.org/projects/million/

Contact information

Jesús Torres-del-Rey
Department of Translation and Interpreting, Faculty of Translation and Documentation & INICO, University of Salamanca
jtorres@usal.es

M.ª Teresa Fuentes-Morán
Department of Translation and Interpreting, Faculty of Translation and Documentation, University of Salamanca
tfuentes@usal.es

Acknowledgements

This article was developed within the framework of the PReLemma project (Parameters for more accessible multilingual lexical resources), supported by the Spanish Ministry of Science and Innovation (Project ID: PID2022-137210OB-I00).

Thordis Ulfarsdottir & Ellert Thor Johannsson

THE REVERSAL OF A BILINGUAL DICTIONARY USING AI

Abstract This paper presents a pilot study exploring the use of artificial intelligence in reversing a bilingual dictionary. The Modern Icelandic-English Dictionary is a freely available online resource. The aim was to transform this Icelandic-English dictionary into a functional English-Icelandic dictionary while minimising manual labour. Using the DANTE English lexical database as a reference, the study focused on English words beginning with the letter m. The process was divided into five phases. First, SQL queries were used to generate an initial reverse listing. Second, the Gemini large language model was employed to process the output and combine it with data from the DANTE list. The resulting dataset contained 4,750 potential entries, of which 1,750 lacked an Icelandic equivalent. In the third phase, Gemini supplied missing translations, while in the fourth phase it performed sense disambiguation for polysemous entries. The fifth phase involved post-processing, including the elimination of unnecessary entries and extensive proofreading. The findings suggest that AI can successfully handle three of the five phases. However, substantial post-processing still requires human expertise, although AI might also be used to facilitate parts of that work.

Keywords Icelandic; bilingual lexicography; reversal; LLM

1 Introduction

The topic of this paper is the use of artificial intelligence (AI) in lexicographic work, more specifically the use of AI in reversing an existing bilingual dictionary. Such reversals have been attempted and described in the literature before the advent of generative AI (Corda et al., 1998; Deksne & Veisbergs, 2018; Veldi, 2010; Tarp, 2022). Earlier approaches have used computational methods to automate parts of the reversal process, including data extraction, matching, filtering, and restructuring. For example, Deksne & Veisbergs (2018) developed a semi-automatic workflow in which potential translation equivalents were extracted from parallel corpora using statistical methods and subsequently evaluated by a lexicographer. Data from a reversed dictionary were also filtered and merged automatically, while the resulting entries required manual lexicographic post-editing. However, the emergence of large language models introduces a new possibility: extending automation beyond the processing of existing lexical data to tasks, such as generating missing translation equivalents and structuring polysemous entries.

In recent years, lexicographers have begun to test the performance of AI models for various lexicographic tasks (e.g. Jakubíček & Rundell, 2023; Trap-Jensen, 2025). The results of such experiments have been mixed, although it is clear that AI can assist with aspects of dictionary work that require linguistic and lexicographic judgement. This paper presents a pilot study in which a large language model (LLM) is employed to assist with the reversal of a recent Icelandic–English online dictionary, officially launched in 2025. Given the significant improvements in AI performance for Icelandic in recent years, from non-existence to its current state of near native fluency, it is pertinent to investigate whether such models are now capable of performing more complicated lexicographic tasks. The study therefore explores whether generative AI can extend semi-automatic dictionary reversal into areas traditionally dependent on human expertise, while also assessing whether it can meaningfully accelerate the reversal process.

2 Background

Icelandic is a "small language" with around 400.000 speakers. Despite its size the country boasts of a long lexicographic tradition, with the first dictionary (Icelandic-Latin) published already in 1650. Modern bilingual lexicography took a great leap forward in the 19th century with the publication of Konráð Gíslason's Danish-Icelandic dictionary (1851), and in the early 20th century other bilingual dictionaries appeared, of those especially noticeable is Icelandic-Danish dictionary of Sigfús Blöndal (1920-1924), which is still the largest lexical description of Icelandic to date with 108.000 headwords.

As in many countries, commercial lexicographic publications have been greatly reduced in recent years and as a result academic institutions have taken on increased responsibility for furthering and updating bilingual dictionaries. The Árni Magnússon Institute for Icelandic studies (AMI) has been instrumental in providing the Icelandic language community with various modern online resources, such as a database of morphological inflections, a large corpus of Modern Icelandic, historical text archives, and various lexicographic works (see e.g. *málið.is*). These include several dictionaries such as ISLEX (officially launched in 2011), a multilingual Icelandic-Scandinavian dictionary, with Icelandic as a source language and the target languages Danish, Swedish, Norwegian (Bokmål and Nynorsk), Faroese and Finnish. In recent years there have been further efforts in publishing high quality bilingual lexical resources. AMI has also developed and made available online Icelandic-French, Icelandic-German and Icelandic-Polish dictionaries, and the latest work is an Icelandic-English dictionary, which is the basis of the pilot study we describe in this article.

Work on a new Icelandic-English online dictionary commenced in 2021. It was based on the Icelandic lemma list of previous bilingual projects which contained around 55.000 words, covering a broad base of the contemporary vocabulary. When this dictionary was compiled, the workflow was somewhat novel, as the method of using pivot-languages of the previous lexicographic works was employed to identify English equivalents to the Icelandic headwords (Úlfarsdóttir & Steingrímsson, 2022). This sometimes resulted in a long list of potential equivalents, e.g., *endless, unending, infinite, interminable, incessant, continuous*, etc., for a single Icelandic headword, *óendanlegur*. The editorial work largely involved reviewing the results of the automatic process and filtering suggested equivalents down to a more manageable level. After this, the editors filled in the blanks where English equivalents were missing as well as translating usage examples and phrases within the entries. A typical entry is shown in Figure 1.

This online dictionary contains a wide range of contemporary Icelandic vocabulary. The user can choose a metalanguage as either Icelandic or English, giving it a broader user base (Figure 1 illustrates the English interface). This new dictionary has filled a growing need in Iceland for a modern and up-to-date Icelandic-English lexical resource, and it has been very well received (January – June 2026 had 130,000 users).

With a successful portfolio of multiple bilingual dictionaries focusing on Icelandic as source language, the AMI institute has contemplated the feasibility of producing parallel dictionaries where Icelandic is the target language. There is some demand for such resources, especially an English-Icelandic dictionary. The last printed English-Icelandic dictionary was published in 2012 (Sanders et al., 2012) and has not been updated. Producing such work of sufficient quality would be a huge undertaking, at least if conventional lexicographic methods were exclusively employed. Therefore, researchers at the institute have been weighing the benefits and disadvantages of using automatic processes to facilitate a reversal of the existing lexical resources where the Icelandic source language would become the target language of a new dictionary.

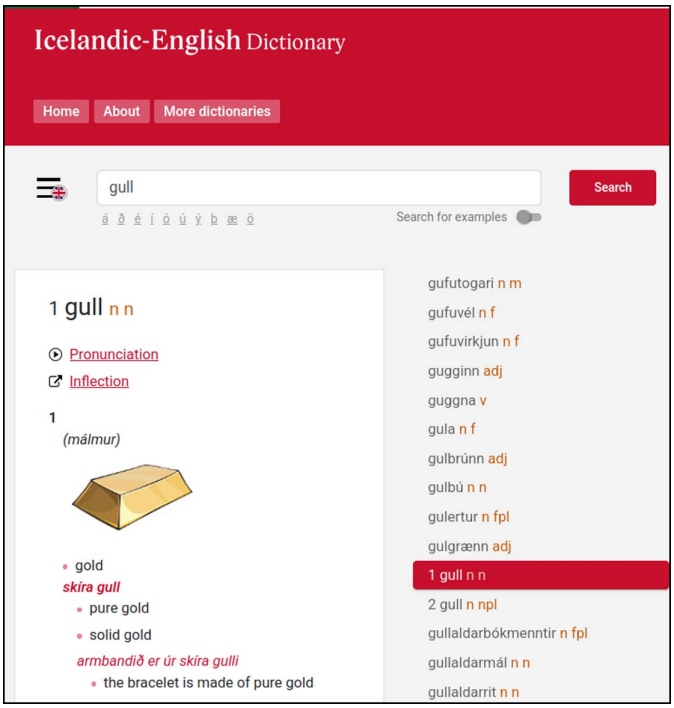


FIGURE 1 Screenshot from the Icelandic-English online dictionary, the noun *gull* 'gold'

The current pilot study is the first step in developing a potential workflow for such an endeavour. We were interested in testing AI models to see how they would perform when they were given certain tasks relating to the reversal of the Icelandic-English dictionary. If successful, the methods could become a standard in future lexicographic projects. At least it would become clearer to what degree LLM's could be of use and how much human intervention was still needed.

3 The study

For the study we defined different tasks in the reversal process and used AI where it could be of the greatest assistance. The reversal process can be divided into five different phases which are not all suitable for AI involvement. The five phases we defined were:

- Generating the English lemma list from IE (Human)
- Supplementing the list with a list of common English headwords creating a combined list with data from both IE and a secondary source (AI)
- Filling in Icelandic equivalents where they are missing (AI)
- Structuring multisense entries (AI)
- Post-processing, editing and proofing (Human (+ partial AI))

3.1 Generating an English lemma list from IE

The first step was to create a simple lemma list based on the English equivalents to the Icelandic headwords. This part did not require the use of AI, just some SQL-queries to the dictionary's database. We wanted to limit the scope of the study and therefore chose to focus

on a single letter of the alphabet, i.e., all English words starting with the letter *m*. The query generated a list of English words from starting with *m*, a part of which is shown in Figure 2.

manufacture	framleiða	v
manufactured	verksmiðjuframleiddur	adj
manufactured goods	iðnaðarvara	n f
manufacturer	iðnrekandi	n m
manufacturer	framleiðandi	n m
manufacturing	framleiðsla	n f
manufacturing	iðnaður	n m
manufacturing defect	smíðagalli	n m
manufacturing fault	framleiðslugalli	n m
manufacturing industry	framleiðsluatvinnuvegur	n m
manufacturing industry	framleiðslugrein	n f
manufacturing process	framleiðsluferli	n n
manure	áburður	n m

FIGURE 2 A snippet of English m-words plus their Icelandic headwords, exported from the Icelandic-English dictionary database

When reviewing the resulting list, the contrast of different lexicographic traditions became apparent. In English dictionaries it is common to find different parts of speech under the same headword, i.e., the "same" word can be used as a noun, verb and even an adjective. Thus, a noun and its corresponding verb (e.g. *run* or *scream*) are contained in the same main entry. In Icelandic lexicography this is never the case. Such cognate word pairs of different parts of speech are always separate entries (e.g. *orka* n.f. 'energy' and *orka* v. 'have strength').

Another difference between English and Icelandic lexicography concerns the treatment of compounds. Since many English compounds are written as open compounds rather than single words, they are often treated lexicographically as separate lexical items. In many cases, a single Icelandic word corresponds to an English expression consisting of a noun modified by an adjective or another noun. Some of the English equivalents in column 1 of Figure 2 illustrate this, for example *manufactured goods*.

Thus, the expression *married couple*, which also appears in the initial list of equivalents, would not normally occur as an independent headword in an English dictionary. Instead, it would typically be treated under *married, marry*, or possibly *couple*. The corresponding Icelandic headword, however, is the single lexical item *hjón*. As a result, the list of potential English lemmas contained many open compounds and multiword expressions that would not ordinarily qualify as independent headwords in English lexicography. We nevertheless chose to retain these expressions, as they correspond directly to single Icelandic words and are therefore of interest to users.

3.2 Supplementing the list and combining data

When looking at the English lemma list it became obvious that it contained significant gaps. In order to make sure that no important English words were missing, it was essential to compare the reversal dataset to a recognised list of lemmas reflecting the current English language. We used a list that was easily obtainable, the DANTE list (Database of Analysed Texts of English). DANTE is a lexical database containing over 42,000 headwords, 23,000 multiword expressions, and 27,000 idioms/phrases. This database was originally developed between 2008 and 2010 by Sue Atkins, Adam Kilgarriff and Michael Rundell (Atkins et. al., 2010) and released under an open-source license (CC-BY 4.0) in 2024. DANTE is available on the Lexonomy

dictionary writing system platform and can be freely used in lexicographic work. We decided that this resource would complement nicely our existing data.

The next step in the process was to make use of the *m*-section of the DANTE database and compare it with the list of lemmas generated from the IE data. The idea was that the DANTE list could fill the gaps and make the English wordlist more comprehensive and useful in further processing. The size of the DANTE database also fits well with the scope of our original data. The list from IE (beginning with *m*) contained 3000 English equivalents. The DANTE list of *m*-words contained 2850 entries. Of those there were 1100 overlapping items, i.e., they were present on both lists indicating that 1750 words in the DANTE list were not recorded in our IE data. As a result, the combined data set was 4750 lemmas.

To consolidate the DANTE and IE datasets into a single, non-redundant list based on part-of-speech (POS) tags, we evaluated several Large Language Models (LLMs). Preliminary pilot tests e.g. with OpenAI's [e.g., GPT-4o] (accessed May 2026) yielded significant rates of omission errors, lexical hallucinations, and duplicate entries. In contrast, Google Gemini 3.5 Flash (accessed May 2026) successfully adhered to the POS constraints. Consequently, Gemini was selected to compile the final dataset. Because a formal benchmark of LLM performance is beyond the scope of this study, a detailed comparative analysis of these trials is omitted. The Gemini model settings were left neutral, but "thinking mode" was selected.

Here is the original prompt to Gemini:

> I am building an English-Icelandic dictionary. I am doing a pilot study where I take an existing Icelandic-English dictionary and "reverse" it and run the results together with an open access Modern English word list. I have two lists that are in txt format. One is all English equivalents starting in *m-* and the Icelandic headwords they belong to . . . . often these are more than one. And the other list is an open access English word list. I want to run those together to combine them.

Gemini responded by suggesting writing a Python script to perform this task (cf. Figure 3). It asked us for a short sample of both lists which we provided. It then proceeded to write the script and after a few modifications it produced a working script that was able to combine the data from both lists, taking into consideration POS tags.



FIGURE 3 A screenshot of part of the Python script written by Gemini

It was not necessary for us to copy the script and run it with a separate program or application. Gemini offered to run the script itself locally, and subsequently the AI produced a combined list of lemmas where the added items from the DANTE list were left empty as they did not have an Icelandic equivalent. The list was returned as a CSV-file, a small part of which is shown in Figure 4.

malt	n; n n; v	malt
malt drink	n n	maltöl
malt loaf	n n	maltbrauð
malt sugar	n m	maltsykur
malted	adj; n	
maltose	n m	maltsykur
maltreat	ptp_adj; v	kúska
maltreated	adj	kúgaður
maltreatment	n; n f	misþyrming
maltster	n	
malware	n f	tölvuveira

FIGURE 4 An example from the combined lemma list of IE and DANTE. Two words on the DANTE list (on the left) have no Icelandic equivalent (on the right)

When reviewing the list we found 15 repeated entries, mostly due to minor spelling differences, i.e., *micro-organism* (IE) vs. *microorganism* (DANTE). Once all such duplicates had been eliminated the list contained 4702 potential entries.

3.3 Filling in Icelandic equivalents

After this rather successful production of a workable lemma list, the next step was to process the 1750 English headwords that did not have an Icelandic equivalent. The AI was asked to perform this task. The prompt was very simple:

> Can you suggest equivalents for the entries missing Icelandic words?

The AI responded by explaining that processing all the words in one go could result in some errors. It suggested that we would divide the list into eight batches of 200 words to be processed one at a time followed by a reduced batch of the remaining words. It then proceeded to process the data and returned an CSV-file for each batch. When all the batches were done, we combined the results into a single Excel-table and used the data to fill in the gaps in the combined list it had produced earlier.

Figure 5 illustrates nine English headwords from the combined list. The Icelandic equivalents in red indicate that the word was originally supplied from DANTE and subsequently processed by Gemini to fill in the missing information.

3.3.1 Some observations and embellishments

It is interesting to take a closer look at how the AI worked to fill in the missing information in the combined list. In many cases it went beyond the prompt, which was to simply provide an equivalent to the English word. It would sometimes generate multiple equivalents and even usage notes and other information that we had not specified (cf. Table 1).

mutant	adj	stökkbreyttur
mutant	n	stökkbrigði (n.)
mutate	v	stökkbreytast (so.)
mutated	adj	stökkbreyttur
mutation	n	stökkbreyting (f.)
mutational	adj	stökkbreytingar- (lo.)
mutatis mutandis	adv	að breyttu breytanda
mute	adj	mállaus, málstola
mute	n	málleysingi (m.), hljóðdeyfir (m.)

FIGURE 5 A sample of the combined list after the AI had supplied the missing Icelandic equivalents. The AI additions are shown in red

TABLE 1 Equivalents provided by AI

English lemmas from DANTE	POS	Icelandic suggested by Gemini	Translations of the Icelandic (for the purpose of this article)
motherfucking	adj	helvítis (lo. - blótsyrði)	damned (adj. – swearword)
motion	v	benda (so.), gefa merki (so.)	point (v.), give signal (v.)
Motown	n	Motown (tónlistarstíll)	Motown (music style)
MIT	n	tækniháskóli Massachusetts (skst.)	The technical University of Massachusetts (abbreviation)
mitch	v	skrópa (so. - írska)	play hooky (v. Irish)
Mitchelstown	n	Mitchelstown (bær á Írlandi)	Mitchelstown (a town in Ireland)

Table 1 illustrates some of the embellishments the AI performed on the data. For the headword *motherfucking* it added a usage note in parenthesis that it is "blótsyrði" 'swearword'. The verb *motion* received two equivalents that each convey a slightly different semantic nuance, "benda" 'point', "gefa merki" 'give signal'. *Motown* is supplied with the refinement "tónlistarstíll" 'music style'. The abbreviation *MIT* is dissolved as "tækniháskóli Massachusetts" 'The technical university of Massachusetts' and in parentheses we find the word *skammstöfun* 'abbreviation' abbreviated (*skst.*). The headword *mitch*, 'play hooky', is labelled Irish and *Mitchelstown* is further explained as a town in Ireland.

These examples demonstrate that the AI supplied the data with some additional meta-information, in order to make the entries more understandable for Icelandic speakers.

3.3.2 Error rates and error types

To check the quality of the AI generated equivalents, we manually checked the first 250 words and rated them as "good", "needing minor adjustment" or "needing major adjustment/ replacement". The result of the sample dataset indicated that 81% of the suggested equivalents were good, 15% needed minor adjustment and 4% were either severely erroneous or wrong. The 19% of entries needing minor or major adjustment were of various types, e.g. deficient equivalents, misspelled forms, Icelandified English words, incorrect translations and lexical fabrications. Most of these errors are easily corrected by simple proofreading, but they underline the fact that AI is not completely reliable and some post-processing is required.

3.3.3 Closer look at the errors

If we take a closer look at only the entries that needed minor or major adjustment in the sample evaluation, 46% were deficient in some way and needed additional editing. For example the word *mage* (n.) was correctly translated as *vitringur* but also as *töframaður* 'magician', which would have to be removed in the post-processing stage. In other cases, a more precise equivalent was missing, i.e. *magna cum laude* was translated as *með miklu lofi* 'with great praise' which is a possible Icelandic translation of this Latin phrase but most commonly it is translated as *með miklu láði*, which would have to be added by hand.

There were examples where Icelandic equivalents had spelling errors or typos. For example, the suggested equivalent "rykagögn" for the English word *mote*. This appears to be some corrupted variation of *rykögn* 'dust particle', which would be a good equivalent. In our sample data we found 8% of entries to be deficient in this way. Sometimes the Icelandic equivalents were "Icelandified" English words (12%), e.g. *malayskur* for expected *malasískur* as translation of *Malay* (adj.). There were also some examples where the translation was wrong (25%), e.g. *magenta* (adj.) was translated *fúksíurauður* 'foccasia-red' instead of the ususal *purpurarauður* 'purple red'. All such cases would need to be corrected in post-processing.

Most noteworthy were instances where the AI would fabricate a completely new Icelandic word. The clearest example is the suggested Icelandic equivalent to the word *munchies*. Here the AI generated the word *nartsultur* followed by a usage note *(oft vegna kannabisneyslu)* '*snack-hunger (often a result of cannabis consumption)*'. We have not been able to find this word in any Icelandic lexicographic database nor corpus data. It appears to be a compound word of *nart* 'snack' and *sultur* 'hunger, starvation' so it is comprehensible to Icelanders and seems like a soundly formed Icelandic neologism. It is interesting that in such cases where there really is no corresponding word in Icelandic the AI invents a plausible word based on the assumed meaning instead of paraphrasing. In our small sample, 8% of errors were of this kind.

3.3.4 Inconsistencies

During the processing of the nine batches, we suspected an error had occurred in one of the batches with words starting with the prefix *micro-*. As a result, we decided to ask the model to process this batch of words again. Even though the suspected error turned out to be non-existant we noticed that the processing was not completely consistent with the previous batch. There was no modification of the prompt nor the LLM settings, but nonetheless the AI did not return the completely same set of equivalents. In some cases, the second run produced additional equivalents, and in other cases important information that had been produced in the first run was left out in the second run. This can be seen in Table 2.

TABLE 2 A table showing some examples of different results from the same data being processed by the AI in two separate runs

Words	POS	1st run	2nd run
microprocessing	n	örgjörvavinnsla (f.)	örvinnsla (f.)
microprocessor	n	örgjörvi (m.)	örgjörvi (m.)
microprogram	n	örforrit (n.)	örforrit (n.)
microscopical	adj	smásjárlegur (lo.), örsmár (lo.)	smásjár- (lo.)
microscopically	adv	í smásjá (ao.), á örsmáan hátt (ao.)	í smásjá (ao.)
microscopy	n	smásjárskoðun (f.)	smásjárfræði (f.)

While most of the Icelandic equivalents remain the same, we find different equivalents in several instances, e.g. *örgjörvavinnsla* vs. *örvinnsla* for *microprocessing*. The first equivalent is more precise, but the second one can also refer to the same concept although it is more commonly associated with micromachinery. Similarly, we find *smásjárskoðun* 'examination with a microscope' vs. *smásjárfræði* 'microscopy studies' for *microscopy*. There are also some examples where multiple equivalents in the first run were reduced to only one in the second run. This begs the question of whether the whole dataset should be processed more than once. Perhaps several runs of the same dataset would result in more nuanced equivalents and would increase the accuracy of the dictionary. For this pilot study we decided to limit the processing to a single run, but it is important to be aware of the inconsistency in how the AI treats the same dataset. This highlights the probabilistic nature of LLMs and how reproducing results is impossible even with the same model on the same day using the same settings.

3.4 Sense division by AI

The fourth phase in processing the results was the treatment of words with multiple senses. After the initial SQL-query and its output from the IE database, the sense structure was not clearly indicated. In 36 cases the reversal process produced excessive number of words (more than six per sense), such as the English adjective *mad* that has 26 Icelandic equivalents. This was not an AI problem, it just reflects the rich synonymity of Icelandic adjectives meaning 'mad'. The AI could in fact be employed to consolidate those and reduce the number of synonyms during a post-processing phase.

Generally the number of equivalents was more manageable. Additional instructions to the AI were given to correctly order the different senses and to add numbers (1, 2, 3 . . .) to separate them. The prompt was:

> This is an entry in a bilingual dictionary. Divide it into senses and provide a dictionary layout.

Here are three example words after the processing:

> **magazine** n. 1. (publication) tímarit n., blað n. 2. (storage, military) birgðageymsla f.
>
> **motif** n. 1. (recurring theme in art, literature, or music) minni n., stef n., mótíf n. 2. (visual design or pattern) munstur n. 3. (specific literary or folkloric elements) leiðarminni n., sagnaminni n., farandminni n.
>
> **mould** n. 1. (fungus, mildew) mygla f., myglusveppur m., fúkki m. 2. (hollow container for shaping, cast) mót n., mát n., form n. 3. (loose rich earth, soil) gróðurmold f.

The input for these words was: *magazine* n birgðageymsla (f.), tímarit (n.), blað (n.); *motif* n farandminni (n.), munstur (n.), stef (n.), leiðarminni (n.), sagnaminni (n.), mótíf (n.), minni (n.); *mould* n mygla (f.), gróðurmold (f.), myglusveppur (m.), fúkki (m.), mót (n.), mát (n.), form (n.).

The generated sense divisions were evaluated manually by the authors. A division was considered successful if the Icelandic equivalents corresponding to the same English sense were grouped together, distinct senses were kept separate, and the resulting structure was suitable for dictionary presentation with no misplaced equivalents. Our evaluation indicates that in all cases the AI was able to figure out the different semantic groupings and correctly identify two or three different senses. It provided meta-information in the form of a semantic label

or specification in parenthesis to provide a simple overview of the different senses. This was quite an acceptable outcome even if in some cases the English meta-information was too long (e.g. '*recurring theme in art, literature, or music*').

3.5 Human intervention and postprocessing

Once AI had performed the three defined tasks we moved on to post-processing. This involved humans reading through the suggested entries with a critical eye to correct errors and seeing if there were still some improvements that needed to be made. Because the DANTE lexical database was originally developed for monolingual English lexicography it contains a relatively high number of proper nouns and placenames — such as *Madison Avenue* or *Macgillycuddy's Reeks* — that are unsuitable for a bilingual English-Icelandic dictionary. In addition, some of the words from DANTE seemed very specialised and could be eliminated, e.g. *magneto* and *magnetron*. At the same time, the review also showed that many of the English multiword expressions from IE that corresponded to a single Icelandic word were superfluous or could be consolidated. In total, the critical analysis of the lemma list indicated that 15% of the processed entries were not needed for the type of general bilingual dictionary we had in mind.

The post-processing also revealed other aspects of the lemma list that would need further processing. Certain entries, especially high-frequency verbs, needed a more hands-on approach. Such words have complex semantic structures and should be looked at separately. In the sample data we found that multiword expressions featuring the verb *make* were quite numerous and would have to be structured in a single semantically nuanced entry. Other verbs that should be treated separately are modal verbs such as *may* and *must*. Since such entries are not very many (only 3 in our sample of *m*-words), we decided not to attempt to process these further as a part of the current pilot study.

Our conclusion is that the output from the AI needs to be checked and corrected, so postprocessing by humans is absolutely essential in order to obtain the desired final result. Since the study focuses on the AI component, all the details of the workflow for post-processing have yet to be formalised.

4 Conclusion

The purpose of this pilot study was to investigate the feasibility of employing AI in lexicographic work, specifically in the process of reversing a bilingual dictionary. The study demonstrates that dividing the reversal process into defined tasks is useful and makes the process easier to manage. Some tasks, such as the initial reversal of the dictionary data, do not require AI assistance. Many subsequent tasks can be largely performed by AI but post-processing should be done by humans.

The study included three specific tasks where the AI was used: a) creating a combined lemma list from equivalents from the original dictionary and supplementary data from an external wordlist; b) filling out information where equivalents from the new target language were missing; c) creating a sense structure for specific entries with multiple equivalents which belonged to different semantic domains. For all three tasks the AI performed well. The combined list covered most aspects of the expected vocabulary even though slight spelling differences could result in repetitive entries. These were, however, few and the final result was a functioning starting point for further lexicographic processing. The supplying of missing Icelandic equivalents was also satisfactorily executed, although there were some minor errors and occasional lapses. It is possible that the results could be improved by adding multiple processing of the

same dataset, perhaps by different LLMs and then comparing the results. The structuring of entries with multiple senses was also in general executed well although some of the more complicated verbs should be treated separately.

The results of the study indicate that using AI to assist in the reversal process of a bilingual dictionary can be of significant help. In some respects, the AI outperformed our expectations, e.g., by adding supplementary meta-information. At the same time, it is clear that AI cannot resolve all problems and that human lexicographic judgement is often required in the post-processing phase.

The reversal process also exposes many structural asymmetries between the two languages. Some Icelandic lemmas appear repeatedly as equivalents across many English *m*-lemmas, while others occur only in highly specific use. The DANTE database also contains several types of entries that are superfluous in a general bilingual context. The post-processing phase might involve additional use of AI in ranking the DANTE entries with the aim of eliminating entries of limited relevance. It could also involve using AI to identify unnecessary multiword lemmas. Finally, it could be worth exploring to ask AI to suggest additional entries that might still be missing from the list, i.e. words that are neither present in IE nor DANTE. However, such post-processing tasks were not part of the current study. It is clear that further refinement of the processed list needs to be revisited at a later stage if the workflow described in the pilot study becomes the basis for a reversal of an entire dictionary. The conclusion is that the final lemma list needs to be critically reviewed and refined, possibly by combining AI and human efforts.

The observations discussed in this article underscore that at its current stage AI can support but not replace human lexicographic analysis. The primary value of this study is that it offers a tested framework for potentially extending the method to a full dictionary between English and Icelandic, using AI under a human supervision. Furthermore, the results indicate that similar methodology can be employed when reversing other bilingual dictionaries, not only Icelandic ones, but also dictionaries in other languages.

References

Atkins, S., Kilgarriff, A. & Rundell, M. (2010). Database of Analysed Texts of English (DANTE): The NEID database project. In A. Dykstra & T. Schoonheim (Eds.), *Proceedings of the 14th EURALEX International Congress* (pp. 549–556).

Blöndal, S. (1920–1924). *Íslensk-dönsk orðabók*. Prentsmiðjan Gutenberg.

Corda, A., Lo Cascio, V. & Pipolo, M. (1998). Automatic reversal of a bilingual dictionary: Implications for lexicographic work. In T. Fontenelle, P. Hiligsmann, A. Michiels, A. Moulin & S. Thei (Eds.), *Proceedings of the 8th EURALEX International Congress* (pp. 433–443). Euralex.

DANTE lexical database. Retrieved January 31, 2025, from https://dantedictionary.com/#/dante

Deksne, D. & Veisbergs, A. (2018). A workflow for supplementing a Latvian-English dictionary with data from parallel corpora and a reversed English-Latvian dictionary. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (Eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts* (pp. 127–135). Ljubljana University Press, Faculty of Arts. https://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202018/118-4-2926-1-10-20180820.pdf

Gíslason, K. (1851). *Dönsk orðabók með íslenzkum þýðingum*. Copenhagen.

IE = *Icelandic-English Dictionary*. Ulfarsdottir, T. (Ed.). The Árni Magnússon Institute for Icelandic Studies. Retrieved January 22, 2026, from https://enska.arnastofnun.is/en/

ISLEX = ISLEX: Íslensk-norræn veforðabók [Icelandic-Nordic Dictionary]. Ulfarsdottir, T. (Ed). The Árni Magnússon Institute for Icelandic Studies. Retrieved May 10, 2026, from https://islex.arnastofnun.is/

Jakubíček, M. & Rundell, M. (2023). The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography? In M. Medveď, M. Měchura, I. Kosem, J. Kallas, C. Tiberius, M. Jakubíček & S. Krek (Eds.), *Electronic Lexicography in the 21st Century* (pp. 518–533). Lexical Computing CZ s.r.o. https://elex.link/ojs/index.php/elex/article/view/46

málið.is. Retrieved May 10, 2026, from áðhttps://málið.is/

Sanders, C., Hólmarsson, S., Tucker, J., Skaptason, J., Hafsteinsdóttir, D., Þórarinsdóttir, H., Markan, I., Harðardóttir, S. & Hafstein, Þ. (Eds.). (2012). *Íslensk-ensk, ensk-íslensk vasaorðabók*. Forlagið.

Tarp, S. (2022). Turning bilingual lexicography upside down: Improving quality and productivity with new methods and technology. *Lexikos, 32*(1), 66–87. https://doi.org/10.5788/32-1-1686

Trap-Jensen, L. (2025). Fremtidens ordbøger: Udfordringer og tilpasninger. *LexicoNordica, 32*, 177–186. https://doi.org/10.7146/ln.v1i32.160908

Úlfarsdóttir, T. & Steingrímsson, S. (2022). Dannelsen af en tosproglig ordbog med hjælp af sprogteknologiske metoder. *LexicoNordica, 29*, 153–173. https://doi.org/10.7146/ln.vi29.134780

Veldi, E. (2010). Reversing a bilingual dictionary: A mixed blessing? In A. Dykstra & T. Schoonheim (Eds.), *Proceedings of the 14th EURALEX International Congress* (pp. 861–865).

Contact information

Thordis Ulfarsdottir
The Árni Magnússon Institute for Icelandic Studies
thordis.ulfarsdottir@arnastofnun.is

Ellert Thor Johannsson
The Árni Magnússon Institute for Icelandic Studies
ellert.thor.johannsson@arnastofnun.is

Carlos Valcárcel Riveiro, María Teresa Sanmarco Bande,
Iván Arias-Arias & María José Domínguez Vázquez

CAN LEXICOGRAPHIC ENTRIES EMBED DIDACTIC MODULES?

A Case Study on LLM-Generated Outputs for a Dictionary Interface

Abstract Digital dictionaries can become more than consultation tools when their internal data are sufficiently structured to be reused and transformed into embedded learning components. This paper evaluates whether large language models (LLMs) can convert PORTLEX entries, represented as reviewed JSON files, into interactive HTML didactic modules integrable into their corresponding dictionary entries. PORTLEX is a multilingual valency-oriented dictionary of the nominal phrase, and its microstructure offers a controlled source for activities on formal links, ordering, complement function and semantic compatibility. The experiment compares a GPT-based custom system and a Gemini-based Gem across three languages, four semantically aligned nouns per language and two model outputs per entry, producing 24 modules and 48 expert evaluations. Results show that both systems generated broadly usable modules and generally preserved the linguistic-valency information encoded in the JSON input. The main limitations did not concern basic HTML functionality, but pedagogical regulation: feedback, correction, scaffolding, distractor quality and ambiguity management. Overall scores did not differ significantly; only feedback and correction favoured GPT after statistical adjustment. Language-based patterns were diagnostic rather than conclusive. The results support treating LLM-generated modules as controlled prototypes within a documented, reusable workflow, not as ready-made pedagogical products.

Keywords LLMs; PORTLEX; didactic lexicography; valency; prompt engineering

1 Introduction

Generative AI changes the position of dictionaries in the lexicographic workflow. Not because large language models can replace editorial judgement, but because they lower the cost of transformation. A structured lexical resource can now be reformulated as an explanation, an interface, an exercise or an evaluative tool with limited technical effort. In lexicography, however, speed is not value. The central question is whether the generated object remains accountable to curated data and to the use for which it is produced.

Current debate has often focused on the generation of lexicographic content, such as definitions, examples or draft entries (De Schryver, 2023; Fuertes-Olivera, 2024; Lew, 2024). This paper examines a different use of LLMs. The model is not asked to supply linguistic authority. It receives lexicographic data already selected and organised by experts and transforms them into another user-facing format. The authority of the dictionary entry remains in place: the LLM does not determine the linguistic content, the relations it establishes or the constraints it defines.

The experiment therefore does not ask whether LLMs can write dictionary entries. It asks whether they can transform structured PORTLEX entries (PORTLEX, 2018)[1] into didactic HTML

[1] PORTLEX (Domínguez Vázquez & Valcárcel Riveiro, 2020) is a multilingual valency-oriented dictionary of the nominal phrase whose entries describe the combinatory behaviour of nouns. For a description of its microstructure and didactic potential, see Section 2.2.

modules that are linguistically reliable, pedagogically usable and technically integrable into a dictionary interface. This format matters because PORTLEX entries are generated through an interface that outputs editable HTML pages, where additional code can be embedded. The modules are bounded learning components derived from the microstructure of the entry, not independent lessons. Here, the purpose is to make nominal valency visible through four evaluative activities: formal links, ordering of the nominal construction, complement function and semantic compatibility.

Each entry was processed using two LLMs—a custom GPT (GPT-5.5 Thinking) and a Gemini Gem (Gemini 3.1 Pro)—with identical prompting architecture and structured input, generating a dataset of 24 HTML modules across German, French and Spanish (four nouns per language), each independently evaluated by two authors with relevant language expertise, yielding 48 evaluations. To what extent can LLMs transform structured PORTLEX entries into linguistically reliable, pedagogically usable and technically functional HTML didactic modules? The analysis addresses model performance, evaluation dimensions, recurrent problems and language as a diagnostic variable. The contribution is limited but practical. The paper evaluates whether curated valency data can be converted into usable didactic prototypes, where the transformation succeeds, and which parts of the workflow still require expert control.

2 Background

2.1. LLMs, lexicographic workflows and pedagogical transformation

Current work has shown that LLMs can generate lexicographic content at scale (De Schryver, 2023; Fuertes-Olivera, 2024; Lew, 2024). Scale, however, is not a lexicographic criterion. A fluent output is useful only if it remains accountable to evidence, editorial decisions and a defined use scenario. This study takes a different position: rather than using LLMs as sources of linguistic authority, it employs them as instruments for transforming curated lexicographic data into a format accessible to learners. This productivity changes the problem rather than solving it. LLM-generated material may contain unsupported claims, invented examples or inaccurate generalisations. Experiments on ChatGPT and lexicographic tasks have also shown variation across sessions, limited metalexicographic knowledge and strong dependence on user expertise (Arias-Arias et al., 2025). These risks connect directly with wider concerns about opacity, linguistic representativity and reproducibility in AI-assisted lexicography (Gouws & Tarp, 2024; Rundell, 2023).

A controlled workflow gives the model a different function. The entry remains the authority, while the LLM renders it in another form. This position is compatible with e-lexicographic models in which centralised lexical databases feed several products, interfaces and applications, and in which data reuse becomes both a methodological and sustainability issue (Fuertes-Olivera, 2024; Klosa-Kückelhaus & Tiberius, 2025). The task is therefore not to generate new lexicographic knowledge, but to reorganise already validated information.

This distinction makes the question of supported and unsupported content central to the study. Rebuffel et al. (2022) define hallucination in data-to-text generation as content not supported by the input data. For this study, that definition is more useful than a general warning about AI. A didactic item derived from a dictionary entry fails when it introduces a relation, restriction, example or interpretation that the entry does not license. JSON input can reduce this risk by making fields explicit. It cannot remove the need for evaluation. For this reason, the evaluation must consider whether each generated item remains licensed by the entry.

Educational generation raises the same issue in another form. Recent work on LLM-generated learning lessons and assessment items shows that educational value cannot be inferred from generated interaction alone (Lin et al., 2025; Meissner et al., 2024). In the present study, the executable form of the HTML module is secondary. The relevant question is whether the module regulates a pedagogical transformation of lexicographic data.

This is the frame for the experiment on PORTLEX, a multilingual and cross-lingual valency-oriented dictionary of the noun phrase (Domínguez Vázquez & Valcárcel Riveiro, 2020). The study asks whether structured PORTLEX entries can become evaluative HTML didactic modules for learners. A module that loads is not yet a module that teaches. The substantive issue is whether the passage from entry to activity preserves nominal valency analysis, avoids unsupported reformulations and makes curated lexicographic information pedagogically usable.

2.2. PORTLEX and valency-oriented didactic potential

PORTLEX (2018) treats nouns as combinatory units. Its entries describe how a nominal head selects, organises and constrains its arguments within an online, multilingual and cross-lingual database whose language-specific modules are linked through a common semantic index (Domínguez Vázquez & Valcárcel Riveiro, 2020). A PORTLEX entry separates the elements needed for this experiment: argument structure, formal realisation and semantic restriction. It identifies arguments, roles and syntactic functions; records prepositions, other links, realisations and order; and specifies constraints on possible fillers. Examples show these relations in use. The entry is therefore not a flat list of equivalents, but a compact description of how a noun builds a nominal construction.

The four activities evaluated in this study follow the structure of the entry. Activity 1 targets formal links. Activity 2 works on the order of complete syntactic blocks. Activity 3 focuses on complement function. Activity 4 tests semantic compatibility. This transformation is also possible at the level of the interface: PORTLEX entries can be edited as HTML, which allows bounded didactic modules to be embedded in the entry itself. The experiment therefore asks whether LLMs can convert the valency-oriented microstructure of PORTLEX into evaluative learning activities without flattening, distorting or over-extending the underlying lexicographic analysis. The didactic relevance follows from the nature of learner difficulty in the three languages retained for the experiment. A learner may know the meaning of *conseil*, *Rat* or *consejo* and still hesitate over the preposition introducing an argument, the order of the noun phrase, the interpretation of a complement or the semantic compatibility of a possible filler. The problem is not vocabulary recognition. It is controlled production and interpretation inside the noun phrase. PORTLEX makes that problem explicit.

3 Methodology

3.1. Dataset and experimental design

The dataset was built from PORTLEX because its entries already separate the elements that the experiment needed to transform and evaluate: nominal heads, argumental structure, formal realisations and semantic restrictions. This structure placed the models under a constrained transformation task rather than an open generation task. Three languages were retained: German, French and Spanish. This choice introduced enough formal variation to test the transformation of nominal valency across different grammatical systems, while keeping the expert evaluation feasible within the limits of the study. The noun set was restricted to four

semantically aligned entries per language: German *Rat*, *Anwesenheit*, *Geruch* and *Sieg*; French *conseil*, *présence*, *odeur* and *victoire*; and Spanish *consejo*, *presencia*, *olor* and *victoria*. These complete, reviewed entries were selected because each cross-linguistic lexical set represented the same meaning and shared the same semantic valency schema and semantic roles across the three languages, while formal realisations remained language-specific. This purposive selection allowed the models' transformation of equivalent semantic-valency structures to be compared across different formal encodings; the resulting dataset was designed for controlled comparison rather than to represent the full lexical range of PORTLEX.

Each entry was processed by two configured LLM environments: a custom GPT running GPT-5.5 Thinking (hereafter GPT) and a Gemini Gem running Gemini 3.1 Pro (hereafter Gemini). Model names refer to the versions available in the experimental environment at the time of module generation. Both systems used the same task-specific prompting architecture and the same structured input format. The ChatGPT and Gemini ecosystems were chosen based on their widespread use and their ability to support task-specific configuration via their paid subscription interfaces. Restricting the comparison to these two systems prioritised depth of expert evaluation over breadth of model coverage. The design was fully crossed at the level of languages, entries and models: 3 languages × 4 nouns × 2 models = 24 HTML modules. Each module received two independent expert evaluations, yielding 48 evaluations in total. Four expert reviewers participated. The distribution was balanced at module level, but not fully crossed at reviewer level: two of the four reviewers evaluated modules in two languages, while the reviewer pairs were not identical across German, French and Spanish. The design therefore supports a controlled exploratory expert comparison of model outputs (Section 3.5). It does not support strong claims about independent language effects. Language-based results are treated later as diagnostic patterns, not as proof of language-specific model behaviour.

3.2. JSON conversion as controlled lexicographic input

The PORTLEX entries were converted into JSON because the experiment required the models to operate on the same lexicographic structure, not on model-specific reconstructions of that structure. JSON was chosen to remove this layer of variation. For each entry, the PDF version printed from the PORTLEX interface was used as the source for extraction. The initial JSON extraction was assisted by ChatGPT, using the same GPT-5.5 Thinking model later used in the custom GPT environment, and all JSON files were verified against the original entries by the authors. This preparatory stage was not part of the model comparison. After author verification, the same reviewed JSON files were provided to both systems as fixed input; neither received the original PDFs nor the extraction dialogue. The experiment processed only the lexicographic content of PORTLEX entries and involved no personal or learner data.

Using JSON as the sole input constrained the models to a transformation role. They were instructed not to supplement the entry with external knowledge or reinterpret the lexicographic analysis encoded in the JSON files. The conversion did not aim to enrich PORTLEX data, but to preserve them in a form that could be processed consistently. The relevant information was encoded in separate fields for the lemma, meaning, arguments, valency patterns, formal realisations, connectors, examples and semantic features. This separation was essential for the four activity types: formal links depend on connectors; ordering tasks depend on syntactic realisations; functional recognition depends on argument roles; semantic compatibility depends on the semantic characterisation of possible fillers.

Both configured systems received the same JSON structure for each entry. The shared input format-controlled variation arising from divergent reconstructions of the source entry;

observed differences therefore reflect the complete model-and-platform configurations under the shared prompting setup. JSON-based input therefore functioned as a control mechanism. It reduced interpretive noise before generation began.

3.3. Prompting strategy

We treated prompt design as part of the method, not as a preliminary technical step. The first version of the master prompt was drafted from the experimental specifications and open-access prompting guidelines (Gadesha, 2026). It was then refined through reciprocal AI-assisted dialogue using ChatGPT (GPT-5.5 Thinking) and Gemini (Gemini 3.1 Pro): each system commented on and proposed improvements to the other system's instructions until the authors consolidated a common version. The methodological constraints remained fixed by the authors. Pilot modules were generated in both environments only to identify general problems affecting evaluability and usability; they were excluded from the experimental dataset. The final version was not treated as an empirically optimal prompt, but was frozen as the standard prompt once it satisfied the predefined requirements summarised in Table 1 and the reciprocal revisions no longer produced substantive changes. The instruction text ultimately used in both systems was identical verbatim. Each model subsequently received the same master prompt, the same user request and one reviewed PORTLEX JSON entry, with no examples of completed modules.

TABLE 1 Main components of the master prompt

Prompt component	Function in the experiment
Role specification	Frames the model as specialist in second-language teaching and didactic lexicography
Source rules	Makes the PORTLEX JSON the only authoritative source
Interface language rule	Keeps learner-facing text in Spanish while preserving entry language in examples
Activity specification	Defines the four fixed activity types
Item generation rules	Prioritises direct JSON data and conservative recombination
Technical constraints	Requires one self-contained HTML file without dependencies
Evaluation constraints	Fixes eight items per activity and excludes session-level item randomisation
Usability constraints	Requires menu screen, activity cards and navigational controls

Two decisions made the comparison stricter. First, we incorporated JSON input into the prompt as a source rule: the reviewed PORTLEX JSON was the only authoritative lexicographic basis for the module content. We rejected any correction or intervention during generation, since it would have broken output comparability. Second, we removed session-level item randomisation from the module specification. The initial design allowed dynamic item selection, but this would have made expert evaluation unstable. The final prompt required exactly eight fixed items per activity and no session-level variation. Comparability took priority over replay value. When an entry did not contain enough explicit material to support all eight items directly, the prompt allowed only conservative recombination of information already encoded in the reviewed JSON and required an activity-level reliability warning. Language was also controlled. The prompt was written in English, the learner-facing interface was fixed in Spanish, and the nominal constructions remained in the language of the PORTLEX entry.

This separation allowed the same instruction set to operate across Spanish, French and German without translating the linguistic material under analysis.

The same instruction architecture had to operate in both environments. We therefore divided the final specification into a compact system prompt and an extended knowledge file, MASTER_PROMPT_PORTLEX.md. The compact prompt contained the non-negotiable rules, while the knowledge file preserved the full operational specification. The main components of this specification are summarised in Table 1. The final master prompt thus functioned as an experimental control, standardising source use, task structure, output format and comparability conditions across both environments. The full prompt specification is available in the Zenodo archive (Valcárcel Riveiro et al., 2026). Because both target systems participated in its reciprocal refinement and received the same final instruction text, the prompt was not optimised unilaterally through GPT. Nevertheless, an identical prompt cannot be assumed to be fully model-neutral, since prompt configurations may affect models differently (Liu et al., 2023). The results therefore compare GPT and Gemini under this specific shared prompt and generation setup, rather than establishing their general relative performance.

3.4. Module generation

Module generation was treated as the experimental execution of the prompt, not as a new design phase. For each PORTLEX JSON entry, the same user instruction was submitted once to the custom GPT and once to the Gemini Gem, using the default generation settings of each environment. All generations were performed manually, one entry at a time, through standard ChatGPT Plus and Google AI Pro subscriptions, without API access, batch automation or per-module costs beyond the subscription fees. Temperature, top-p and seed were not manually set or modified, and the equivalence of these default parameters across platforms cannot therefore be assumed. This single-run design was deliberate: the study focused on expert evaluation of the resulting modules rather than on variability across repeated generations. No module was regenerated or selectively improved before evaluation. Differences in the generated modules therefore reflect the complete model-and-platform configurations under their default settings, not variation in task formulation. Because no repeated generations were performed, the study does not assess output variability or generation-level repeatability. The standardised JSON-and-prompt workflow is procedurally extensible to additional entries, but manual generation and, especially, expert validation remain the principal constraints on scalability.

Each system had to generate one complete self-contained HTML file. The module opened with a Spanish interface, while the nominal constructions used in the exercises remained in the language of the corresponding entry. This separation made the evaluation interface comparable across German, French and Spanish without translating the linguistic material under analysis. The generated file had to contain a welcome screen, four activities, eight fixed items per activity, independent scores out of 10, feedback after correction and navigation between sections. These requirements fixed the minimum conditions under which a generated module could be evaluated as an experimental object: stable as an evaluation artefact and comparable across reviewers. The systems did not always provide a downloadable file directly. When necessary, the HTML code was copied into a local file without altering its content. The resulting 24 HTML files were preserved as experimental outputs and submitted to expert evaluation as generated. They are also included in the accompanying Zenodo archive (Valcárcel Riveiro et al., 2026). The modules were used solely as experimental outputs; no learner data were collected, and they were not deployed in classroom or online teaching.

3.5. Evaluation instrument and data analysis

The evaluation instrument was designed as an expert assessment of generated educational material, not as a measure of student learning. Its purpose was to determine whether each HTML module was linguistically accurate, didactically usable and technically viable as an extension of a valency-oriented dictionary interface. This follows recent evaluation-oriented work on LLM-generated learning lessons and assessment items, where generated artefacts require human validation before pedagogical use (Lin et al., 2025; Meissner et al., 2024).

The grid therefore combined three global criteria — technical validity, navigation and flow, and visual design — with four criteria repeated for each activity: linguistic-valency accuracy, didactic adequacy and scaffolding, interaction design, and feedback, correction and solutions. This structure separated interface viability from activity quality and enabled comparison of which activity types resisted automatic transformation more successfully (see Table 2). The feedback and correction criterion did not only record whether the module marked answers as correct or incorrect; it also assessed whether expected solutions were displayed and whether correction offered explanatory support after the learner's response.

TABLE 2 Aggregated evaluation dimensions

Aggregated dimension	Source in the questionnaire
Technical and interface validity	Global criteria C1–C3
Linguistic-valency accuracy	Criterion 1 across activities 1–4
Didactic adequacy and scaffolding	Criterion 2 across activities 1–4
Interaction design	Criterion 3 across activities 1–4
Feedback and correction	Criterion 4 across activities 1–4
Overall usability	Manual global rating
Critical problems	Multiple-choice critical error field + comments
Language/model diagnostic comparison	Manual global rating grouped by language and model

The manual global rating was kept separate from the analytic grid. It captured expert viability as a whole, because an arithmetic mean can hide decisive problems: a clear and functional module may still be unusable if it contains wrong answers, unsupported information or an incorrect interpretation of valency. The questionnaire also included critical-problem fields and qualitative comments. Each response corresponded to one HTML module. We calculated mean scores by model, aggregated dimension and activity type; counted critical problems by frequency; and used comments to identify recurrent failure modes such as ambiguous answers, weak distractors, insufficient explanatory feedback or unsupported valency interpretations. Model comparisons were conducted at entry level. The two reviewer scores for each module were averaged, yielding 12 paired GPT–Gemini observations; given the small sample and ordinal scale, differences were assessed using two-sided Wilcoxon signed-rank tests. The manual global rating was treated as the primary comparison, while the seven analytic criteria were treated as secondary comparisons and adjusted using Holm's method ($\alpha = .05$). Because each language was evaluated by a different fixed pair of reviewers, inter-rater agreement for the manual global rating was calculated separately by language using quadratically weighted Cohen's kappa, appropriate for the ordinal 1–5 scale.

The evaluators were also the four authors of the paper and members of the research team associated with PORTLEX. Each module was independently evaluated by two authors, assigned according to their knowledge of and academic specialisation in the output language.

Familiarity with PORTLEX was a deliberate selection criterion because the object of evaluation was not the quality of the dictionary entries themselves, but how accurately the models processed their reviewed JSON representations and transformed their valency and grammatical content into didactic activities. Given PORTLEX's highly specific grammatical orientation and microstructure, external evaluators would have required prior training in its analytical framework to assess these dimensions. The PORTLEX entries therefore functioned as validated reference input against which the generated modules could be assessed. The resulting assessment should be understood as internal expert validation rather than evaluation by independent external reviewers, though this dual role may also introduce a degree of familiarity bias despite the shared reference input.

Language was included only as a diagnostic variable. Since reviewer pairs were not identical across languages, language-based comparisons were not treated as independent language effects. They can suggest interactions between model behaviour, entry structure and expert perception, but they cannot isolate language as a causal factor.

4 Results

4.1. Overall performance by model

At module level, after averaging the two reviewer ratings, GPT obtained a mean manual global rating of 3.79 (SD = 0.40), while Gemini obtained 3.46 (SD = 0.78). Both values exceed the scale midpoint (3), but the paired difference was not statistically significant (two-sided Wilcoxon signed-rank test, $W = 12.5$, $p = .293$). Descriptively, GPT module ratings were more concentrated, whereas Gemini module ratings showed greater dispersion. However, this apparent difference is driven largely by the Spanish subset, where reviewer agreement was weakest (Section 4.7); it should therefore not be interpreted as robust evidence that GPT was globally superior or more stable. The comparison is therefore retained as an exploratory description of variation across the evaluated outputs, especially useful for identifying where expert validation may need to be reinforced and for informing larger, controlled comparisons of model–platform configurations.

4.2. Global technical and interface criteria

Technical validity received the highest and most symmetrical scores: GPT and Gemini both obtained 4.79. The modules generally loaded, displayed the expected structure and functioned as autonomous files. Basic HTML generation was therefore not the main point of contrast. Visual design followed the same pattern. Both models reached 4.46, and reviewers generally accepted the visual organisation of the outputs. The interfaces were not final products, but they were readable enough to support evaluation. Navigation and flow showed the largest descriptive difference among the global interface criteria: GPT reached 4.54 and Gemini 3.92. However, this difference did not remain statistically significant after Holm correction (adjusted $p = .141$). If learners cannot move through the module clearly, the evaluative structure becomes harder to interpret. The pattern is clear: the HTML and visual layer did not constitute the main obstacle. The more serious limitations appeared in pedagogical regulation, feedback and correction (Section 4.3).

4.3. Aggregated dimensions across activities

Figure 1 shows a split result. Linguistic-valency accuracy is the most robust dimension: GPT scored 4.55 and Gemini 4.38. This supports the JSON-based control mechanism described

in Section 3.2. When the relevant valency information had already been separated in the entry, both systems generally preserved it in the generated activities. The result does not mean that every item was valid. Rather, it suggests that the linguistic core of the entry resisted transformation better than the pedagogical regulation built around it.

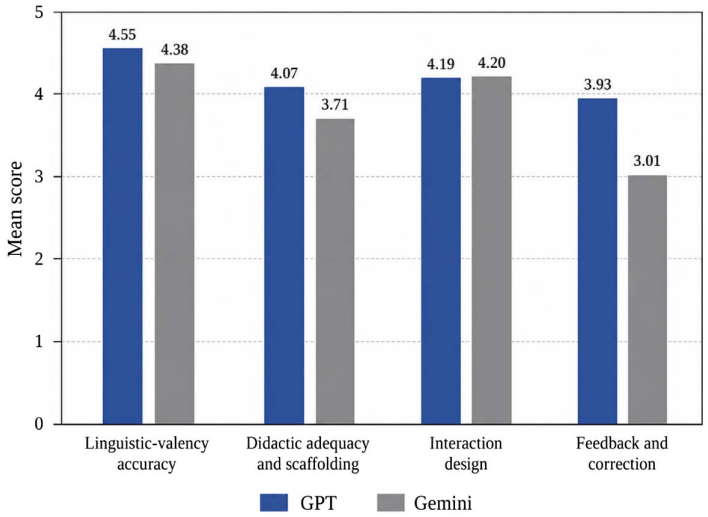


FIGURE 1 Mean scores by model and aggregated evaluation dimension

Interaction design produced almost no model difference: GPT scored 4.19 and Gemini 4.20. Once the prompt fixed the basic activity structure, both systems could generate workable interactive formats. This complicates any simple ranking, since model differences were not uniform across dimensions. The weaker dimensions are pedagogical. Didactic adequacy and scaffolding were lower than linguistic-valency accuracy for both models, with GPT at 4.07 and Gemini at 3.71. Feedback and correction produced the clearest contrast: GPT reached 3.93, while Gemini fell to 3.01. Among the seven analytic criteria, this was the only model difference that remained statistically significant after Holm correction ($W = 3.0$, adjusted $p = .017$). The modules could present tasks and calculate scores more easily than they could guide interpretation after error. Several outputs lacked stable mechanisms for showing expected answers, explaining incorrect responses or allowing learners to review them. The main limitation is therefore not the transformation of valency content itself, but the control of pedagogical regulation. Feedback and scaffolding must be specified in the prompt; they cannot be left to the model's autonomous judgement.

4.4. *Differences between activity types*

Differences between activity types were moderate. A3, focused on complement function, obtained the highest mean score (4.09), followed closely by A4 semantic compatibility (4.06) and A1 formal links (4.04). A2, which required ordering the elements of the nominal construction, obtained the lowest score (3.82). The range is too narrow to support a strong hierarchy, but it identifies ordering as the most fragile operation in this set. The exception is interaction design in A2, where Gemini scored higher than GPT (4.00 vs 3.50), suggesting that block manipulation was handled differently by the two systems. Several items involved monoargumental nominal structures, with only the head and one complement to arrange. In these cases, the answer was often too obvious. Weak disordering, capital-letter cues and unclear manipulation reinforced the same effect. The activity sometimes tested recognition of a default sequence more than control of nominal construction. The lower score for A2

therefore reflects item design and manipulation quality rather than the intrinsic unsuitability of ordering tasks. A4 shows the opposite tension: semantic compatibility is central to valency-oriented didactics, but it is also exposed to drift. Compatibility judgements may move beyond the PORTLEX entry towards world knowledge, metaphorical interpretation or contextual plausibility. Its relatively high mean therefore indicates potential, not stability. The boundary between valency-based restriction and broader semantic acceptability needs tighter control.

4.5. Exploratory comparison by language and model

Grouped by language, the model comparison becomes less linear. In German, both systems obtained the same global mean (3.88). In French, Gemini was slightly higher than GPT (3.75 vs 3.63). In Spanish, the pattern reversed more sharply: GPT reached 3.88, while Gemini fell to 2.75. The descriptive model pattern therefore differed across language subsets. These differences cannot be read as independent language effects, given the reviewer distribution described in Section 3.1. However, they are useful for describing how the observed ratings varied across language subsets. The Spanish subset shows the largest model gap, but also the weakest reviewer agreement (Section 4.7); this apparent difference should therefore not be treated as reliable evidence of either a model or a language effect. German and French point in a different direction: they show balance or even a slight Gemini advantage. The language comparison is therefore retained only as a diagnostic description of this dataset.

4.6. Critical problems and qualitative evidence

The critical-problem data confirm the pattern in Section 4.3. The main failure zone was pedagogical regulation: insufficient or incorrect feedback was reported 28 times, lack of error correction 23 times, and insufficient, excessive or inconsistent scaffolding 18 times. The modules often worked as minimal evaluative objects: they presented tasks and produced scores. They were weaker as learning-oriented tools, because they did not consistently explain errors, show expected solutions or allow learners to review their answers. A second group of problems affected item validity. Erroneous or ambiguous correct answers were identified 13 times, and inadequate distractors 7 times. These problems change what the activity measures. If two answers are possible, if the expected answer is not licensed by the PORTLEX data, or if the distractors are too easily rejected, the item no longer tests the intended valency-related competence.

The qualitative comments show the same pattern in concrete form. In semantic compatibility tasks, examples such as *Der Geruch der Freiheit* show how easily compatibility can drift from valency restriction to metaphorical or world-knowledge interpretation. In ordering activities, capital letters or overly simple structures could reveal the first block or make the task obvious. At the level of feedback, several modules remained limited to confirmatory or corrective feedback, marking answers as correct or incorrect without explaining the error or showing the expected solution. This suggests that future prompts should specify not only that feedback must be provided, but also what type of feedback is expected for each activity. These are not separate defects. They point to the same requirement: the prompt must guide scaffolding during task completion and explanatory correction after task completion, not only define activity types. Visual design problems (3) and navigation issues (5) were much less frequent, which confirms that the main limitation did not lie in the interface layer.

4.7. Reviewer convergence

The degree of reviewer convergence varied across language subsets. The mean absolute difference between the two global ratings assigned to each module was 1.00 point, and in 19 of the 24 modules the difference was equal to or lower than one point. However, quadratically weighted Cohen's kappa, calculated separately for each fixed reviewer pair, was low: $\kappa w = .17$ for German, .25 for French and .00 for Spanish. These coefficients should be interpreted cautiously because each language subset contained only eight modules and the ratings occupied a restricted range.

Convergence was strongest in the French subset (mean absolute difference: 0.38), followed by German (1.00), and weakest in the Spanish subset (1.63). The Spanish subset therefore combines the largest apparent difference between models with the weakest reviewer agreement. Its GPT–Gemini difference should not be treated as reliable evidence of either a model or a language effect. More generally, these results support only an exploratory interpretation of the tendencies observed in the dataset.

5 Discussion

The experiment shows that PORTLEX entries can be transformed into didactic modules integrable into the dictionary entry, but only when the input is structured, the prompt is constrained and the output is evaluated by experts. The relevant threshold is not whether the generated HTML runs. It is whether the module turns the lexicographic analysis into a usable learning task. The extension of dictionary interfaces through bounded, user-adapted components responds to a well-established demand in e-lexicographic research: interfaces must filter, prioritise and adapt information according to the specific needs and consultation scenarios of their users (Klosa-Kückelhaus & Michaelis, 2024; Müller-Spitzer, 2022).

5.1. Structured input reduces risk, but does not eliminate it

The strongest support for the methodology comes from linguistic-valency accuracy. As shown in Section 4.3, both systems preserved much of the information contained in the entries when it was supplied as reviewed JSON. This confirms the role assigned to the models: they were not asked to infer nominal valency from general language knowledge, but to transform curated lexicographic data into another format. That control has limits. A structured source does not decide whether a generated item is evaluatively valid. Some semantic compatibility tasks moved too easily from valency-based restriction to broader plausibility or metaphorical interpretation, as the critical problems discussed in Section 4.6 show. Other items were formally consistent with the entry but weak as test items. JSON-based prompting makes unsupported content easier to detect; it does not make it impossible.

5.2. Pedagogical regulation is the main issue

The weakest point was not interface construction. Technical validity and visual organisation were high, and both systems produced autonomous HTML modules that loaded, displayed activities and calculated scores. The difficulty appeared when the module had to behave as a pedagogically accountable evaluation tool. A score is not explanatory feedback, and interaction is not scaffolding. Confirmatory or corrective feedback may tell learners whether they were right or wrong, but several modules went no further. They marked answers without explaining the error, showing the expected solution or allowing learners to review their responses. The result was performance registration without interpretive guidance.

For didactic lexicography, that distinction is decisive. Each activity must evaluate a specific valency-related operation. Each item therefore needs a clear relation between the entry, the expected answer and the learner-facing correction. Weak distractors, ambiguous answers and unexplained corrections break that relation. The hard part is not producing interactive code, but controlling what the activity teaches, tests and explains (Heift, 2021).

5.3. Exploratory model comparison

The comparison between GPT and Gemini should not be reduced to a ranking. Although GPT ratings were less dispersed overall, this pattern was driven largely by the Spanish subset, where reviewer agreement was weakest. The observed differences therefore describe the outputs produced under this experimental configuration rather than general model performance. Their practical value lies in showing that identical structured input and prompting can produce different outcomes across model–platform configurations and that expert validation remains necessary. Larger studies with repeated generations and a fully crossed reviewer design are needed to test model-specific stability.

5.4. Implications for dictionary interfaces

The experiment supports a restrained extension of the dictionary interface. Valency-oriented entries can incorporate bounded didactic components derived from their own microstructure. The dictionary is not replaced by a learning platform; it is extended through small interactive components anchored in the entry itself (Zhang, 2025). PORTLEX is especially suited to this use because its entries combine structured valency data with an editable HTML environment.

This does not make LLM-generated modules finished pedagogical products. They are prototypes whose value depends on a controlled chain: curated entry, reviewed JSON, constrained prompt, generated HTML, expert evaluation and refinement. The results show where that chain must be strengthened. Feedback rules need to be more explicit. Distractors need stronger validation. Semantic compatibility tasks need tighter boundaries between lexicographic restriction and broader contextual plausibility. Learner-based evaluation is also necessary, since expert usability does not measure learning effect.

6 Conclusions

This study shows that reviewed PORTLEX JSON entries can be transformed into autonomous HTML didactic modules integrable into the dictionary entry. This is not evidence for autonomous pedagogical generation. Rather, it shows that curated lexicographic data can support rapid prototyping when LLMs are assigned a bounded transformation role. GPT and Gemini produced broadly usable modules, but usability was uneven. The controlled JSON input helped preserve linguistic-valency information: formal links, argument functions, syntactic realisations and semantic restrictions generally survived the passage from entry to activity. The strongest limitations appeared in the pedagogical layer. Feedback, correction, distractor quality and scaffolding remained less reliable than technical execution or visual presentation. Future integration therefore requires tighter control over what each item teaches, tests and explains.

Because reviewer agreement was weakest in the Spanish subset, which produced the largest apparent model gap, the model comparison should be understood as descriptive and exploratory rather than as evidence that GPT was globally superior or more stable. This limitation does not invalidate the module-generation experiment or the recurrent qualitative findings;

the study's main contribution lies in demonstrating the feasibility of transforming reviewed lexicographic data into didactic modules and identifying the recurring strengths and failure modes of that transformation. Future work should reinforce the generation chain at the critical points where prototypes become pedagogical objects: explanatory feedback, correction, distractor validation and semantic-compatibility control. These dimensions are more effectively handled at the prompting and generation level than in the PORTLEX-derived JSON. Student-based evaluation is also required, since expert usability does not measure learning effects. The workflow nevertheless provides a documented procedure that can be reused across entries, languages and datasets, provided that data curation, prompt constraints and expert evaluation are preserved. LLM-generated modules should therefore be treated as controlled prototypes, not as ready-made pedagogical products.

References

Arias-Arias, I., Domínguez Vázquez, M. J. & Valcárcel Riveiro, C. (2025). People, machines and dictionaries: Is artificial intelligence killing dictionaries? In I. Dias, R. H. Gouws, A. Lobenstein-Reichmann & S. J. Schierholz (Eds.), *Regarding the significance of lexicography/Zur Bedeutsamkeit der Lexikographie: Europe and Australia/Europa und Australien* (pp. 235–258). De Gruyter. https://doi.org/10.1515/9783112216644-012

De Schryver, G.-M. (2023). Generative AI and lexicography: The current state of the art using ChatGPT. *International Journal of Lexicography*, 36(4), 355–377. https://doi.org/10.1093/ijl/ecad021

Domínguez Vázquez, M. J. & Valcárcel Riveiro, C. (2020). PORTLEX as a multilingual and cross-lingual online dictionary. In M. J. Domínguez Vázquez, M. Mirazo Balsa & C. Valcárcel Riveiro (Eds.), *Studies on multilingual lexicography* (pp. 135–158). De Gruyter. https://doi.org/10.1515/9783110607659-008

Domínguez Vázquez, M. J. (Dir.), Valcárcel Riveiro, C., Mirazo Balsa, M., Sanmarco Bande, M. T., Simões, A. & Vale, M. J. (2018). *Portlex. Diccionario multilingüe de la valencia del nombre*. Universidade de Santiago de Compostela. http://portlex.usc.gal/portlex/

Fuertes-Olivera, P. A. (2024). Making lexicography sustainable: Using ChatGPT and reusing data for lexicographic purposes. *Lexikos*, 34, 1–24. https://doi.org/10.5788/34-1-1883

Gadesha, V. (2026). *The 2026 guide to prompt engineering*. IBM. https://www.ibm.com/think/prompt-engineering

Gouws, R. H. & Tarp, S. (2024). Despite current challenges: Lexicography has a bright future. *Lexicographica*, 40, 1–20. https://doi.org/10.1515/lex-2024-0009

Heift, T. (2021). Intelligent Computer Assisted Language Learning. In H. Mohebbi & C. Coombe (Eds.), *Research Questions in Language Education and Applied Linguistics* (pp. 655–658). Springer. https://doi.org/10.1007/978-3-030-79143-8_114

Klosa-Kückelhaus, A., & Michaelis, F. (2024). The design of Internet dictionaries. In A. Klosa-Kückelhaus (Ed.), *Internet lexicography* (pp. 171–190). De Gruyter. https://doi.org/10.1515/9783111233758

Klosa-Kückelhaus, A. & Tiberius, C. (2025). The lexicographic process revisited. *International Journal of Lexicography*, 38(1), 1–22. https://doi.org/10.1093/ijl/ecae016

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, 11, Article 45. https://doi.org/10.1057/s41599-024-02889-7

Lin, J., Rao, J., Zhao, S. Y., Wang, Y., Gurung, A., Barany, A., Ocumpaugh, J., Baker, R. S., & Koedinger, K. R. (2025). Automatic large language models creation of interactive learning lessons. In K. Tammets, S. Sosnovsky, R. Ferreira Mello & G. Pishtari (Eds.), *Two decades of TEL: From lessons learnt to challenges ahead. EC-TEL 2025. Lecture Notes in Computer Science, Vol. 16063* (pp. 259–274). Springer. https://doi.org/10.1007/978-3-032-03870-8_18

Liu, P., Yuan, W., Fu, J., Jiang, Z., Hayashi, H. & Neubig, G. (2023). Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language processing. *ACM Computing Surveys*, 55(9), Article 195, 1–35. https://doi.org/10.1145/3560815

Meissner, R., Pögelt, A., Ihsberner, K., Grüttmüller, M., Tornack, S., Thor, A., Pengel, N., Wollersheim, H.-W. & Hardt, W. (2024). LLM-generated competence-based e-assessment items for higher education mathematics: Methodology and evaluation. *Frontiers in Education*, 9:1427502. https://doi.org/10.3389/feduc.2024.1427502

Müller-Spitzer, C. (2022). User-adapted views in online dictionaries. In A. Klosa-Kückelhaus & F. Michaelis (Eds.), *Internet lexicography: An introduction* (pp. 191–214). De Gruyter. https://doi.org/10.1515/9783110712599-010

Rebuffel, C., Roberti, M., Soulier, L., Scoutheeten, G., Cancelliere, R. & Gallinari, P. (2022). Controlling hallucinations at word level in data-to-text generation. *Data Mining and Knowledge Discovery*, 36, 318–354. https://doi.org/10.1007/s10618-021-00801-4

Rundell, M. (2023). Automating the creation of dictionaries: Are we nearly there? In H. Kim & S. Rhee (Eds.), *Proceedings of ASIALEX 2023: Lexicography, Artificial Intelligence and Dictionary Users* (pp. 9–17). The Asian Association for Lexicography. https://asialex.org/pdf/Asialex-Proceedings-2023.pdf

Valcárcel Riveiro, C., Sanmarco Bande, M. T., Arias Arias, I. & Domínguez Vázquez, M. J. (2026). *PORTLEX_LLM_didactic_modules_EURALEX2026. XXII EURALEX International Congress 2026 (EURALEX), Vienna*. Zenodo. https://doi.org/10.5281/zenodo.20346897

Zhang, Y. (2025). Using ChatGPT in learning Chinese Mandarin: An interactive dictionary rather than a content facilitator. *L2 Journal*, 17(1), 1–28. https://doi.org/10.5070/l2.42018

Contact information

Carlos Valcárcel Riveiro
Instituto de investigación Lingua, Universidade de Vigo
carlos.valcarcel@uvigo.gal

María Teresa Sanmarco Bande
Instituto da Lingua Galega, Universidade de Santiago de Compostela
mt.sanmarco@usc.es

Iván Arias-Arias
Instituto da Lingua Galega, Universidade de Santiago de Compostela
ivanarias.arias@usc.gal

María José Domínguez Vázquez
Instituto da Lingua Galega, Universidade de Santiago de Compostela
majo.dominguez@usc.es

Acknowledgements

This contribution was developed within the framework of the research project ESMAS-ES+ (Project PID2022-137170OB-I00), funded by the Spanish Ministry of Science and Innovation, the State Research Agency (AEI)/10.13039/501100011033, and by the European Regional Development Fund (ERDF/EU).

Iván Arias-Arias also acknowledges support from grant FPU21/00188, funded by the Formación de Profesorado Universitario (FPU) programme of the Ministerio de Ciencia, Innovación y Universidades (MCIU).

Martina Vokáčová & Anna Marklová

AFFORDANCE NORMS IN CZECH AND ENGLISH

Human–LLM Misalignment in Embodied Lexical Knowledge

Abstract This paper investigates whether large language models can capture the embodied dimension of word meaning by comparing LLM-generated affordance norms with human-produced norms in Czech and English. Affordance norms are systematically collected data on the actions speakers associate with concrete objects. Human data were collected from 30 Czech native speakers in a free-production experiment using 50 nouns. The same stimuli were presented to four LLMs (Claude Haiku 4.5, Gemini 3 Flash, GPT-5, and Llama 4 Maverick) in both languages, in Czech at two temperature settings; English outputs were additionally compared against human norms from Maxwell et al. (2024). We report four main findings. First, human affordance norms are more variable in Czech. Second, LLMs consistently produce narrower affordance sets than humans, with lower lexical diversity across both languages. Third, lexical overlap between human and LLM affordances is low (shared vocabulary 15–19%; Jaccard index below 0.3 at N = 10) and does not improve substantially with higher temperature or greater language representation in the training data. Fourth, human-only affordances are more embodied and emotionally grounded, while LLM-only affordances cluster around maintenance and support. The findings suggest that human norming data may offer a valuable resource for enriching lexicographic entries.

Keywords Affordances; large language models; psycholinguistic norms; lexicography; Czech; English; cross-linguistic comparison; embodied meaning; reporting bias

1 Introduction

The advent of large language models has prompted debate within lexicography about the extent to which these systems can generate or enrich lexical entries. De Schryver (2023) critically reviewed ten studies on the use of ChatGPT for lexicographic purposes, examining performance across entry components (definitions, example sentences, collocations, and etymology, among others). The review identifies genuine practical advantages – LLMs can produce structurally plausible entries rapidly and at scale – but also recurring failures: hallucinations presented with apparent confidence, verbatim memorization of copyrighted dictionary definitions, inconsistent handling of polysemous words, and a pervasive English-centric bias that severely limits their utility for less-resourced languages. Jakubíček and Rundell (2023) illustrate the last-mentioned problem on Czech. When they asked ChatGPT in Czech for an entry for the noun *stát* 'state' (meaning 'country'), the output included a sense that exists for the English noun but not for the Czech one (*state* in the sense of *condition*). In their main experiment, they generated 99 English dictionary entries and concluded that ChatGPT can match or exceed established tools for some entry components, such as synonyms/antonyms discrimination; nevertheless, its inherent unpredictability and lack of traceability present significant methodological limitations that the lexicographic community must address before integrating this technology into substantive workflows.

A possible structural explanation for many of these limitations is reporting bias. Gordon & Van Durme (2013) articulated this phenomenon in the context of knowledge extraction: the frequency with which events, actions, or properties appear in text does not necessarily reflect their real-world frequency or typicality. In line with Gricean conversational maxims, humans tend to mention what is noteworthy, contextually relevant, or informationally novel – not what is obvious. The result is that text corpora systematically underrepresent the most prototypical

and frequent aspects of everyday interaction with the world. Gordon and Van Durme give the following example: a knowledge extraction system reading large text collections learns that a person may have eyes almost a million times, but that a person may have a spleen fewer than 1,600 times – despite both being equally universal features of human anatomy. The underrepresented knowledge is precisely the kind that is so obvious it goes without saying.

For LLMs trained exclusively on text, reporting bias is not merely a data quality problem but a structural feature of the training signal. Bisk et al. (2020) frame this as part of a broader grounding deficit: LLMs learn language from text alone, without access to the physical and social experience that makes utterances meaningful. They argue that the World Scope of current NLP – constrained to Internet-scale text – is fundamentally insufficient for the kinds of knowledge that grounded communication requires. Bender & Koller (2020) make a related argument: a system trained only on the formal properties of strings cannot in principle learn meaning as a relation between form and communicative intent, because meaning is grounded in the world and that grounding is absent from the training signal. Together, these accounts motivate a concrete prediction: LLM-generated lexical knowledge should be most reliable for dimensions well-captured by distributional text patterns, and most deficient precisely where knowledge derives from embodied, physical interaction with the world.

One way to operationalize this embodied interaction is through the concept of affordances – the potential action-based relationships between bodies and objects in the environment (Gibson, 1979). A *bed*, for instance, affords *sleeping*, *lying down*, or *jumping on*; a *conference paper* may afford *writing*, *reading*, or *submitting on time*. Empirical evidence shows that affordances are activated during sentence processing, connecting object knowledge to event expectations in comprehension (Kaschak & Glenberg, 2000; Glenberg & Kaschak, 2002; Glenberg & Gallese, 2012). This dimension of object knowledge can be captured systematically by collecting affordance norms from native speakers in a free production task, in which participants enumerate all ways they can engage with a given object. Maxwell et al. (2024) recently confirmed that affordance norms of this kind reflect a distinct aspect of word meaning. Their large-scale dataset for English concrete nouns showed only weak correlations with existing associative and semantic feature norms, which they interpret as evidence that affordances constitute a separate dimension of meaning.

In this paper, we examine precisely how well large language models capture this dimension of word meaning – one that is grounded in embodied experience rather than textual co-occurrence – by comparing LLM-generated affordance norms with those produced by human speakers of Czech and English. To address this question, we generated LLM affordance norms for 50 stimuli in these two languages and analyzed them against two human baselines: Czech norms collected from native speakers in the present study, and English norms from Maxwell et al. (2024) for translation-equivalent items. In Section 2, we describe the data collection procedure and affordance measures; results are reported and discussed in Section 3.

2 Methodology

We conducted an experiment following the design of Maxwell et al. (2024), in which Czech native speakers were asked to list activities they typically associate with a given object. The experiment was then replicated with large language models, in both Czech and English. The present paper focuses on the comparison of human and LLM-generated affordances; a full analysis of the Czech norms and their cross-linguistic comparison with English is beyond the scope of this paper and is reported in Vokáčová (in preparation). The dataset, scripts, detailed analyses and visualizations of the data are publicly available at OSF: https://osf.io/qjgby/.

2.1 Material

The stimulus set consisted of 50 concrete Czech nouns:

chléb 'bread', *voda* 'water', *večeře* 'dinner', *mléko* 'milk', *kapr* 'carp', *káva* 'coffee', *dopis* 'letter', *kniha* 'book', *esej* 'essay', *komiks* 'comic', *pohlednice* 'postcard', *encyklopedie* 'encyclopedia', *šaty* 'dress', *boty* 'shoes', *kostým* 'costume', *paruka* 'wig', *džíny* 'jeans', *náušnice* 'earrings', *auto* 'car', *nádobí* 'dishes', *krém* 'cream', *mýdlo* 'soap', *sklenice* 'glass', *parfém* 'perfume', *schody* 'stairs', *stěna* 'wall', *dům* 'house', *most* 'bridge', *stůl* 'table', *nábytek* 'furniture', *hrob* 'grave', *zařízení* 'device', *garáž* 'garage', *fotbal* 'football', *fotka* 'photo', *kostka* 'dice', *film* 'film', *sněhulák* 'snowman', *piano* 'piano', *tetování* 'tattoo', *socha* 'statue', *tělo* 'body', *žába* 'frog', *dřevo* 'wood', *pampeliška* 'dandelion', *kámen* 'stone', *hranol* 'prism', *škola* 'school', *knihovna* 'library', *sauna* 'sauna'

These nouns were distributed across eight semantic categories (food, text, clothing, everyday objects, construction/housing, entertainment, nature, and place) and three corpus frequency bands (high, mid, low).[1] English translation equivalents of 48 of these nouns correspond to items in the Maxwell et al. (2024) affordance norms.[2]

2.2 Procedure

2.2.1 Human participants

The experiment was implemented as an online study using jsPsych (v. 8.2.1) and administered via JATOS on a server hosted by the Faculty of Arts, Charles University. Participants (N = 30) received a link and were recommended to complete the study on a desktop, though the interface supported any device. After providing informed consent and completing a short demographic questionnaire, participants read task instructions with one example and two practice trials before proceeding to the main task.

In the main task, nouns were presented one at a time. Participants were asked to list all activities they or anyone else would typically perform with the given object, entering responses into dynamically generated text fields. At least one response was required to advance. Stimulus order was randomized with the constraint that at least two items from different semantic categories intervened between any two items from the same category. Upon completion, data were submitted to JATOS in CSV format.

2.2.2 Large language models

LLM data collection was carried out using a Python script that replicated the free-production task used in the human experiment, querying four models – Claude Haiku 4.5, Gemini 3 Flash, GPT-5, and Llama 4 Maverick – via the OpenRouter platform. Each of the 50 stimulus nouns was presented as a user prompt, and responses were generated across 100 iterations per noun per model; all data were collected between May 5 and 9, 2026.

[1] These particular nouns were chosen in connection with the broader aims of the underlying project on logical metonymy (complement coercion) in Czech. Candidate items were drawn from the SYN2020 corpus (Křen et al., 2020) as nominal objects of verbs that combine with phase verbs (e.g., *začít* 'begin'); the three frequency bands correspond to the first, third, and fifth quintiles of the resulting frequency distribution. Full details of the selection procedure are given in Vokáčová (in preparation).

[2] The two remaining nouns (*hranol* 'prism' and *nádobí* 'dishes') have no direct equivalent in that dataset and were therefore excluded from the English comparison.

The system prompt was designed to mirror the instructions given to human participants as closely as possible, asking the model to assume the role of a native Czech speaker and to list all activities that they or anyone else could do with the presented object, with a maximum of 20 items. To assess the effect of generation variability, data were collected at two temperature settings (T = 0.7 and T = 1.0). Temperature is a parameter controlling the randomness of the model's output, with higher values producing more diverse responses.

English data were collected for the 48 translation equivalents of Czech cues matching the Maxwell et al. (2024) stimulus set. As the participant instructions used in the present study (hereafter *authors' instructions*) differ slightly from those of Maxwell et al. (2024; hereafter *MX instructions*), data were generated under two instruction conditions – a direct translation of the Czech prompt and a prompt following the exact wording of Maxwell et al. (2024), both at T = 0.7.[3]

2.3 Data processing

Human responses were checked and lemmatized manually. Duplicate responses within a single participant's data for a given stimulus were removed. Raw model outputs were lemmatized using the Stanza NLP library (Qi et al., 2020). Responses that could not be assigned to a valid lemma were flagged as errors. These were subsequently corrected in two steps: first, by automatic matching against the human affordance data for the same stimulus; second, by manual inspection and compilation of a correction dictionary. Responses that remained uncorrected after both steps were excluded from further analysis. To ensure comparability with human data, a random subsample of model iterations was drawn for each stimulus, matched to the number of human participants and their response length distribution.

2.4 Affordance measures

Three affordance measures were computed following Maxwell et al. (2024): *affordance set size* (AFSS), *affordance strength* (AFS), *and affordance proportion* (AFP). **AFSS** is the total number of unique affordances produced for a given cue. **AFS** is defined as the proportion of all responses for a given cue that a particular affordance constitutes, i.e., its frequency divided by the sum of frequencies of all unique affordances for that cue. **AFP** addresses a limitation of AFS by dividing the frequency of a given affordance by the total number of participants (or model iterations) for that cue, thereby reducing sensitivity to individual variation in response quantity.

In addition, we introduce *affordance weighted proportion* (AFWP), which extends AFP by incorporating response order. Rather than treating all responses equally regardless of position, **AFWP** weights each response by the inverse of its position within a given iteration on the assumption that earlier-produced affordances are more salient. It is computed as the mean of inverse-position scores across all iterations for a given cue. This measure proved particularly useful, as LLM responses exhibited lower AFP variability than human data, leading to frequent ties that precluded meaningful ranking.

[3] Temperature was not manipulated for the English because a pilot analysis of the Czech data had shown rather minimal temperature sensitivity (see also Section 3, Figure 1).

3 Results and Discussion

As noted above, the present analysis focuses on the comparison of human and LLM-generated affordances in Czech and English. The cross-linguistic relationship between Czech and English human norms, which shows partial correlations across all three affordance measures, is described in detail in Vokáčová (in preparation).

3.1 *Affordance set size and LLM variability*

Figures 1 and 2 display the mean affordance set size (AFSS) for each model compared to human respondents, for Czech and English cue words respectively. The mean AFSS of Czech human participants (M = 75.3) substantially exceeds the values of all four models under both temperature conditions. Model AFSS values at T = 0.7 range from 23 to 34. Higher generation temperature leads to increased AFSS across all models, though markedly only for Claude Haiku, which reaches a mean of 43 at T = 1.0, approximately 57% of the human mean. Even at T = 1.0, however, none of the models approaches the human mean.

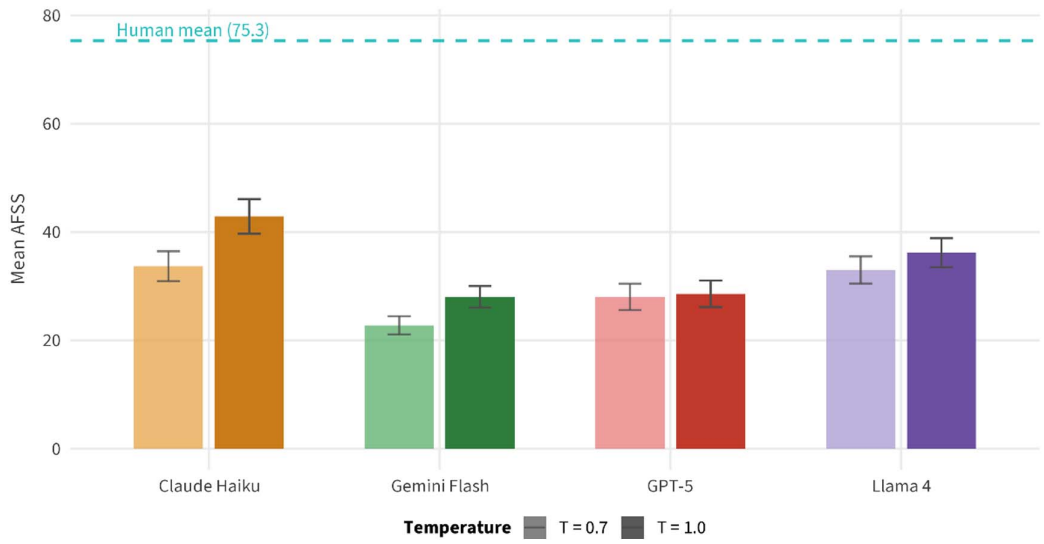


FIGURE 1 Mean affordance set size (AFSS) per model and temperature setting for Czech cue words. Human mean shown for reference

In English, data behave similarly: the mean AFSS of English human participants (M = 39.0) exceeds the values of all models under both instruction conditions. The authors' instruction yields values of 18–19 across all four models; the MX instruction produces higher AFSS (24–32). The largest difference between instructions is found for GPT-5 (19.2 vs. 32.1) and Llama 4 (18.2 vs. 32.5). Nevertheless, no model under either prompt variant reaches the English human mean. The Czech human mean (M = 75.3) is included in Figure 2 for reference and clearly indicates substantially higher AFSS in Czech. This may be largely attributable to the morphological complexity of Czech compared to English – Czech verbs differ in aspect and prefixation, such that where English uses a single verb (*wash*), Czech offers a range of distinct forms (*mýt*, *umýt*, *umývat*, *omýt* etc.).

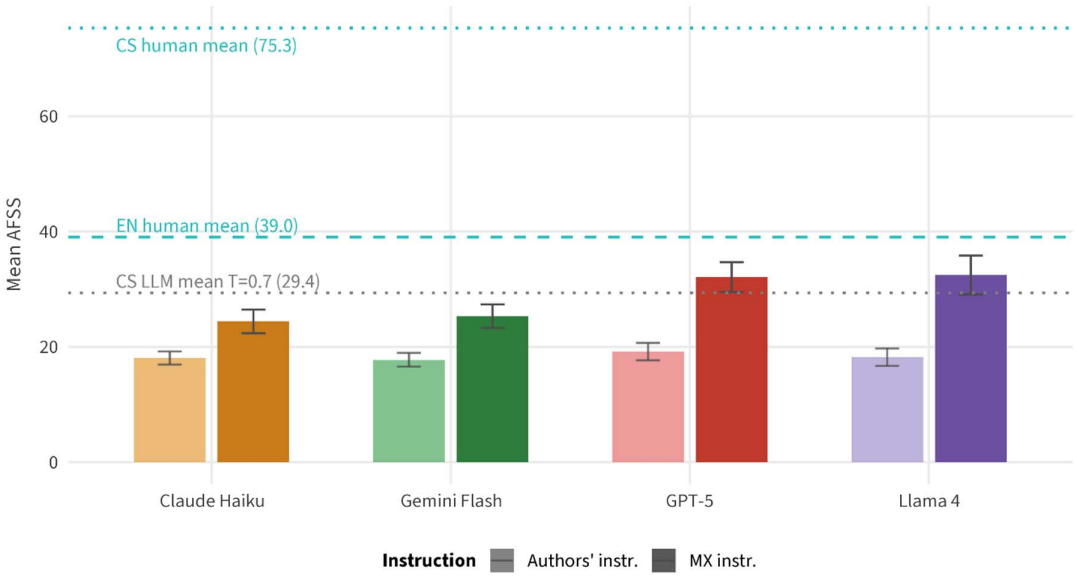


FIGURE 2 Mean affordance set size (AFSS) per model and instruction variant for English cue words. Czech human mean (M = 75.3) included for reference

A complement to AFSS is top-1 AFP – the affordance proportion of the single most frequently produced affordance. As shown by Maxwell et al. (2024) for English and confirmed for Czech (Vokáčová, in preparation), the two measures are negatively correlated: stimuli with a richer affordance repertoire inevitably yield lower AFP. Figures 3 and 4 display top-1 AFP for each model and condition and are consistent with the AFSS findings: LLMs produce smaller affordance sets, and their responses concentrate strongly on a single dominant affordance. At T = 0.7, mean top-1 AFP ranges from 0.93 to 0.97 across all four Czech models, compared to a human mean of 0.54. Higher temperature reduces this concentration somewhat, most noticeably for Claude Haiku (0.94 → 0.89) and Gemini Flash (0.97 → 0.91), while GPT-5 and Llama 4 remain largely stable.

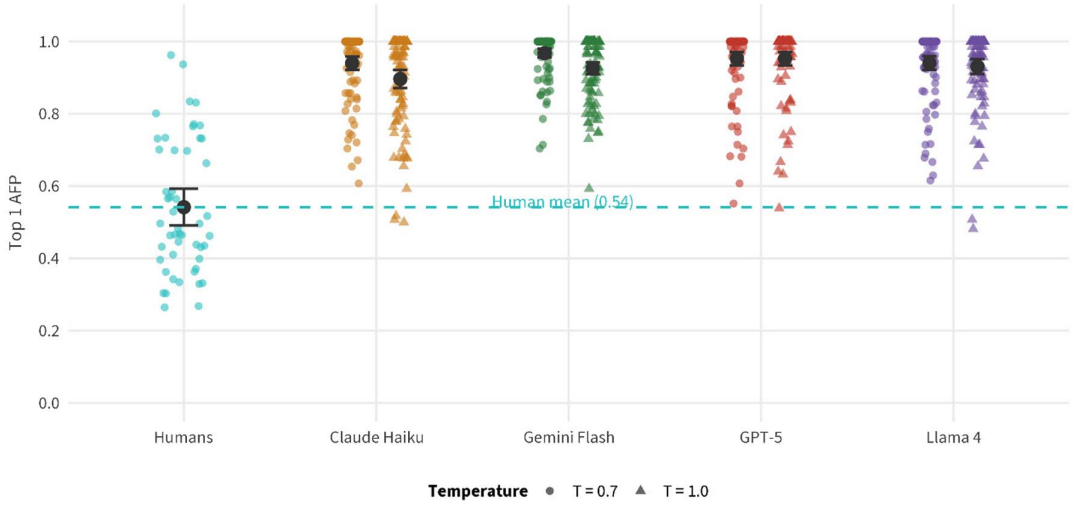


FIGURE 3 Mean top-1 AFP per model and temperature setting, compared to humans

For English, the tendency is even more pronounced under the authors' instruction: all four models approach the ceiling with a mean top-1 AFP of 0.99 or above. The MX instruction introduces somewhat more variability, particularly for GPT-5 (0.88) and Llama 4 (0.87), though values remain well above the English human mean of 0.55.

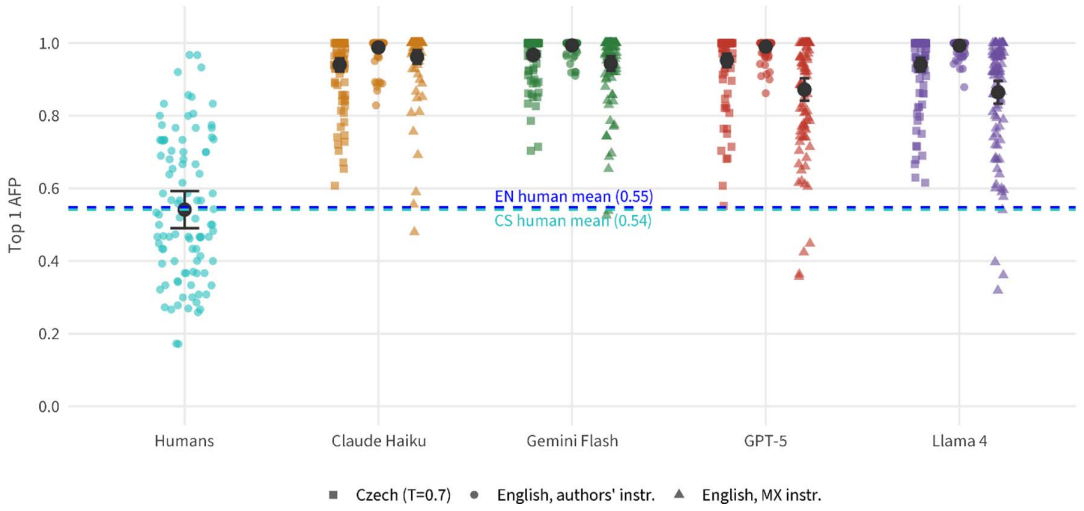


FIGURE 4 Mean top-1 AFP per model and instruction variant

3.2 *Lexical overlap between LLMs and humans*

To assess the lexical overlap between human and LLM affordance sets, Figure 5 displays the Jaccard index for Czech cue words, computed for top 1, 3, 5, and 10 affordances ranked by AFWP, which is the most appropriate metric in this context as it avoids the practical issue of ties that would arise with AFP-based ranking. The index was computed across all 50 Czech cue words for each model at two temperature settings (T = 0.7 and T = 1.0). Three out of four models (Claude Haiku, GPT-5, Llama 4) show a clear descending trend: as N increases, the Jaccard index decreases, which is to be expected – at low N, only the most salient affordances are compared, where agreement is naturally higher, whereas at higher N, less dominant items enter the comparison, which tend to diverge more between human respondents and models. The exception is Gemini Flash, whose Jaccard index shows the smallest decline as N increases (at T = 0.7: J = 0.34 → 0.27; at T = 1.0: J = 0.32 → 0.28). The highest overlap at N = 1 is found for Llama 4 (J = 0.48 at T = 0.7), followed by Claude Haiku and GPT-5 (both J = 0.42); even the best-performing model thus agrees with human respondents on the most salient affordance in fewer than half of the cases. Beyond N = 3, values stabilize at approximately 0.22–0.28 across all models. The wide confidence intervals at N = 1 reflect considerable variability across cue words: for some objects, the model's top affordance matches the human response perfectly, while for others, there is no overlap at all. The effect of generation temperature is minimal: differences between T = 0.7 and T = 1.0 do not exceed 0.06 Jaccard index points for Llama 4 (at N = 1) and 0.02 for all other models, and confidence intervals overlap in all cases.

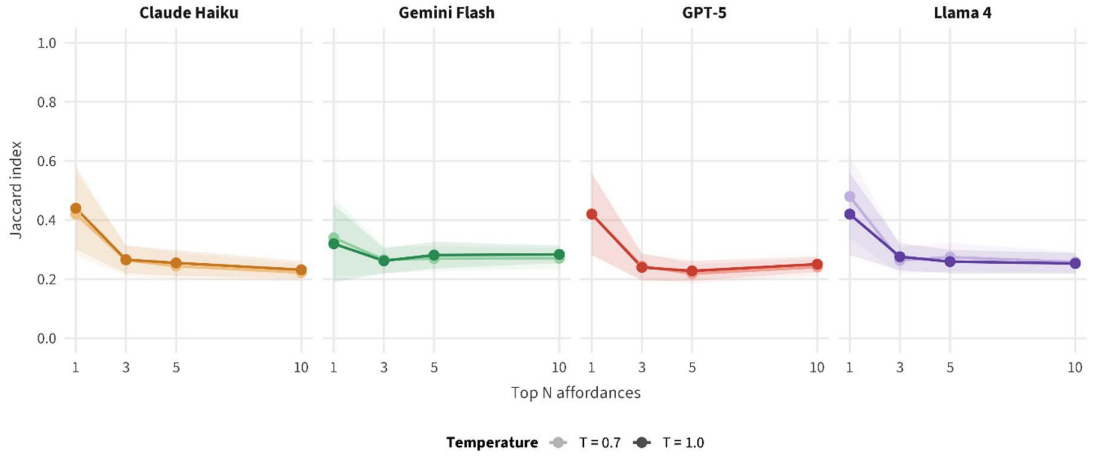


FIGURE 5 Jaccard index between human and LLM affordance sets for Czech cue words, computed for top N affordances ranked by AFWP at two temperatures

English LLM data show higher overlap with native speaker norms than Czech LLM data at $N = 1$,[4] with mean Jaccard values of $J = 0.53$ (authors' instruction) and $J = 0.54$ (MX instruction) compared to $J = 0.44$ for Czech cues. This is consistent across all four models at $N = 1$, though not uniformly so – GPT-5 shows the largest gain ($J = 0.61$ vs. $J = 0.45$ for Czech), Llama 4 the smallest ($J = 0.51$ vs. $J = 0.47$ for Czech). These results may reflect the greater representation of English in LLM training data. However, the pattern does not hold at larger affordance sets: at $N = 10$, the authors' instruction condition converges with Czech (mean $J = 0.24$ for both), and the MX instruction falls below Czech values for all four models ($J = 0.13$–0.18 vs. $J = 0.22$–0.26). The MX instruction thus appears to elicit a narrower affordance vocabulary from LLMs – one that matches the most salient human responses a bit better but diverges as the comparison broadens. Overall, even the best-performing condition does not exceed $J = 0.61$ at $N = 1$, and no model consistently surpasses $J = 0.26$ at $N = 10$, indicating that across both languages the overall level of human–LLM affordance overlap remains moderate at best.

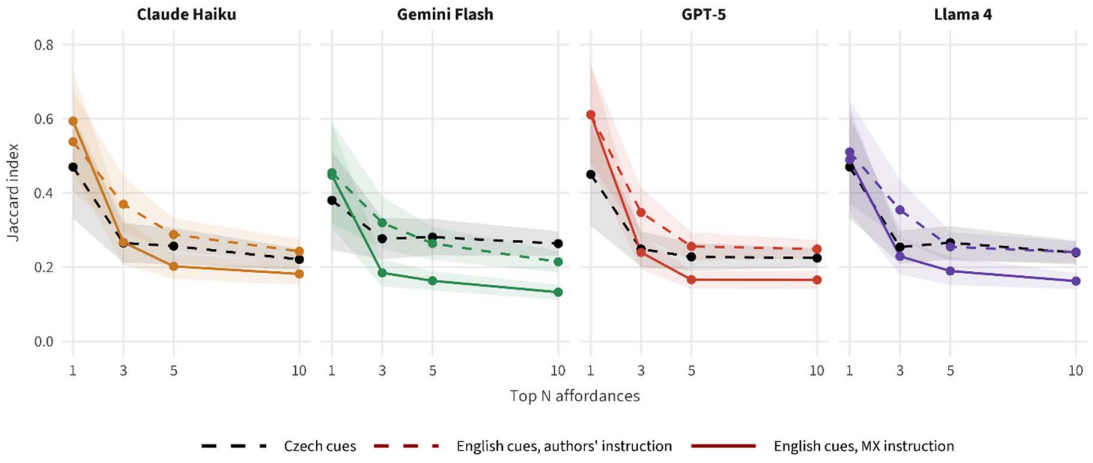


FIGURE 6 Jaccard index between LLM-generated English affordances and Maxwell et al. (2024) human norms, compared with Czech LLM–human overlap (dashed black reference line), for two instruction variants

[4] Human data in this comparison (including the Czech reference values) are ranked by AFP, as the norms of Maxwell et al. (2024) lack the response-order information needed to compute AFWP; hence the slight difference from the AFWP-based values reported for Figure 5. LLM data retain AFWP ranking.

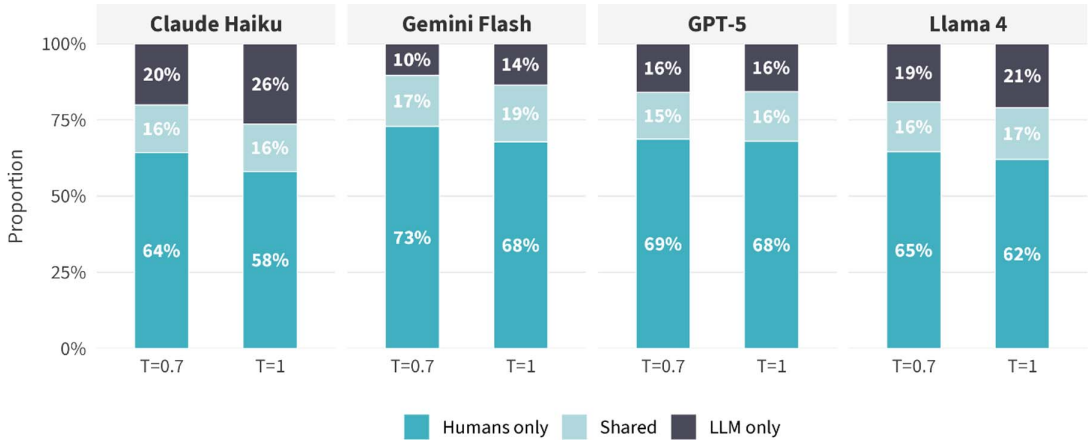


FIGURE 7 Vocabulary overlap between human and LLM affordances for Czech cue words at two temperature settings: proportions of shared, human-only, and LLM-only verbs

Figure 7 displays the proportion of lexical overlap in all responses for each model at two temperature settings. Across all conditions, the shared vocabulary (verbs elicited from both humans and models) constitutes only 15–19% of the total unique responses. Humans and LLMs thus seem to draw on substantially different lexical repertoires when generating affordances. The majority of verbs are unique to their source: humans consistently produce a set of affordances not captured by any model (58–73%), LLMs generate a smaller but non-negligible proportion of verbs absent from human responses (10–26%). Temperature appears to have only a modest effect on this composition – the human-only component remains dominant even at the increased variability induced by T = 1.0.

Figure 8 extends this analysis to English cues, showing the vocabulary overlap for two versions of the system prompt (Maxwell-like instructions vs. authors' instructions) at T = 0.7, and broadly confirms the pattern observed for Czech data (Czech results for T = 0.7 are retained here in each panel for comparison). The human-only component remains prevalent across all conditions (47–73%). The two prompt versions differ markedly in the proportion of LLM- and human-only affordances. The MX instruction yields substantially more LLM-exclusive (30–37%) and fewer human-exclusive responses (47–56%) than either the authors' instruction (18–20%; 60–63%) or the Czech reference (10–20%; 64–73%). The shared component, however, remains low for both (12–16% MX vs. 18–20% authors' instructions). The MX instruction thus appears to elicit a less human-aligned affordance vocabulary. The authors' instruction, by contrast, produces a composition that closely resembles the Czech results across all four models. These results are somewhat counterintuitive: given the far greater representation of English in LLM training data, one might expect higher shared overlap, at least under the MX instruction. The finding that the opposite holds suggests that LLMs may still be unable to generate human-comparable affordances regardless of how much data in a given language they were trained on.

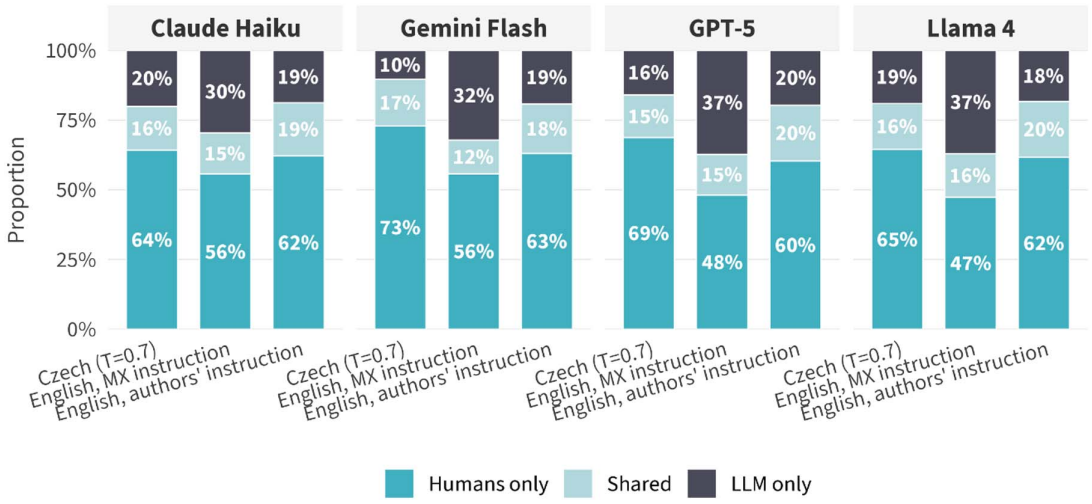


FIGURE 8 Vocabulary overlap for English cue words at $T = 0.7$ under two instruction variants, with Czech results retained for comparison

3.3 Human- and LLM-only affordances

Closer inspection of the top 20 human-only and LLM-only affordances, ranked by the number of cue words for which they occur exclusively (Table 1), reveals several systematic differences. The Czech affordances often reflect emotional or physical engagement (*milovat* 'love', *nesnášet* 'hate', *jít* 'go'), acquisition and disposal (*hledat* 'search', *vybírat* 'choose', *objednat* 'order', *prodat* 'sell', *darovat* 'gift'), and, interestingly, destruction (*zničit* 'destroy', *spálit* 'burn', *poškodit* 'damage'). LLM-only verbs, by contrast, lack emotional and destructive content and instead cluster around maintenance and enhancement of objects (*čistit* 'clean', *leštit* 'polish', *umývat* 'wash', *zdobit* 'decorate', *ozdobit* 'adorn', *malovat* 'paint', *lepit* 'glue', *opravovat* 'repair', *vyměňovat* 'replace'). Morphologically, human-only verbs are perfective more often than LLM-only verbs, which tend to be imperfective (see markings in Table 1).

Similar to Czech, English human-only verbs reflect emotional and physical engagement (*love*, *feel*, *sleep*, *eat*, *sit*), acquisition and disposal (*sell*, *buy*), and destruction (*break*, *destroy*, *burn*, *hit*, *stomp*). These verbs are absent from the LLM-only repertoire. While English LLM-only verbs are less focused on maintenance or enhancement, they contain a large number of verbs related to support (*teach*, *provide*, *serve*, *incorporate*, *practice*).

From this qualitative examination of the top 20 affordances two conclusions emerge: First, humans in English and Czech behave similarly – both show strong representation of verbs of emotion, embodiment, and destruction. Second, LLMs in Czech and English resemble each other more closely than they resemble the human-only verbs of their respective language. In both languages, LLM-only verbs include more verbs expressing maintenance, support, cooperation, or the generation of value, and they lack the human-only focus on emotions and embodiment noted above.

TABLE 1 Top 20 human-only and LLM-only affordances in Czech and English, ranked by number of cue words for which they occur exclusively. Czech perfective verbs are marked in blue

	Human-only	LLM-only
Czech	*mít* 'have'; *nemít* 'not have'; *spálit* 'burn'; *zničit* 'destroy'; *nesnášet* 'hate'; *být* 'be'; *darovat* 'gift'; *objednat* 'order'; *potřebovat* 'need'; *uklidit* 'clean'; *vybírat* 'pick'; *jít* 'go'; *najít* 'find'; *prodat* 'sell'; *uklízet* 'clean'; *vyrobit* 'create'; *hledat* 'search'; *milovat* 'love'; *poškodit* 'damage'; *vyrábět* 'manufacture'	*čistit* 'clean'; *dávat* 'give'; *zdobit* 'decorate'; *přenášet* 'carry over'; *skládat* 'assemble'; *leštit* 'polish'; *fotit* 'photograph'; *malovat* 'paint'; *vyměňovat* 'replace'; *balit* 'wrap'; *dělat* 'do'; *lepit* 'glue'; *opravovat* 'repair'; *ozdobit* 'adorn'; *přidat* 'add'; *umývat* 'wash'; *používat* 'use'; *připravit* 'prepare'; *uložit* 'store'; *dát* 'give'
English	break; throw; destroy; sell; buy; burn; look; love; feel; hit; show; warm; eat; memory; protect; run, sit; sleep; stomp; talk	use; create; place; practice; provide; make; serve; act; teach; display; put; store; take; incorporate; donate; keep; part; base; repurpose; test

4 Conclusion

In this paper, we explored affordances from two new perspectives. Firstly, we collected Czech affordance norms for 50 concrete nouns in an online study to assess the cross-linguistic transferability of affordance knowledge. Secondly, we generated a parallel dataset of Czech and English affordances for the same stimuli using four LLMs. Comparison of human-produced affordances showed that Czech speakers listed a more variable dataset than English speakers. This is due in part to the morphological complexity of Czech, where aspectual distinctions and prefixation yield multiple distinct lemmas for a single semantically similar activity.

Crucially, the comparison of humans and LLMs paints a consistent picture. LLMs generate a narrower range of affordances than human participants, show limited lexical overlap with human norms regardless of language, model, or prompt condition, and tend to produce a qualitatively distinct type of affordances – ones oriented toward maintenance and support, lacking the emotional engagement and embodied experiences characteristic of human responses. Notably, English LLM outputs overall do not appear to show substantially higher alignment with human norms than Czech LLM outputs. This suggests that the English-centric bias of current LLMs does not compensate for the grounding deficit when it comes to affordance knowledge.

These conclusions should be read in the light of several limitations. The observed differences would benefit from statistical testing on a larger sample, both of stimuli and, crucially, of participants: as in Maxwell et al. (2024), our data rest on approximately 30 respondents per cue. A larger sample would additionally enable a further step – quantifying within-human variability (for instance, via a split-half Jaccard index) so as to establish a baseline against which the human–LLM overlap can be interpreted.

Despite these limitations, our results point to one possible path lexicography could take in the age of AI – a small step back from treating the textual frequency of a word as a proxy for the real-world importance of its referent. Fine-tuned AI lexicographers, as envisaged in several of the studies discussed above, are likely to become valuable tools for drafting entries. But our data show where their drafts will be systematically thin: in capturing senses rooted in embodied experience, of which affordances are a prime example. If lexicography is to open its workflow to LLMs on one side, we suggest it should open it to psycholinguists on the other – norming studies of the kind reported here provide a useful corrective that a human lexicographer can apply to AI drafts.

Just as the corpus revolution in past decades redirected linguistics' attention to frequency, the capabilities – as well as, or even better, the limitations – of LLMs trained on those frequencies may now cause a further shift: toward normed data that recover what distributional methods miss, namely the sensorimotor knowledge speakers share about what lexemes mean without putting it into words so frequently.

References

Bender, E. M. & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In D. Jurafsky, J. Chai, N. Schluter & J. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.463

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., Lapata, M., Lazaridou, A., May, J., Nisnevich, A., Pinto, N. & Turian, J. (2020). Experience grounds language. In B. Webber, T. Cohn, Y. He & Y. Liu (Eds.), *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 8718–8735). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.703

de Schryver, G.-M. (2023). Generative AI and lexicography: The current state of the art using ChatGPT. *International Journal of Lexicography*, 36(4), 355–387. https://doi.org/10.1093/ijl/ecad021

Gibson, J. J. (1979). *The ecological approach to visual perception*. Houghton Mifflin.

Glenberg, A. M. & Gallese, V. (2012). Action-based language: A theory of language acquisition, comprehension, and production. *Cortex,* 48(7), 905–922. https://doi.org/10.1016/j.cortex.2011.04.010

Glenberg, A. M. & Kaschak, M. P. (2002). Grounding language in action. *Psychonomic Bulletin & Review,* 9(3), 558–565. https://doi.org/10.3758/BF03196313

Gordon, J. & Van Durme, B. (2013). Reporting bias and knowledge acquisition. In F. M. Suchanek, S. Riedel, S. Singh & P. P. Talukdar (Eds.), *Proceedings of the 2013 Workshop on Automated Knowledge Base Construction* (pp. 25–30). Association for Computing Machinery. https://doi.org/10.1145/2509558.2509563

Jakubíček, M. & Rundell, M. (2023). The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography? In M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček & S. Krek (Eds.), *Electronic lexicography in the 21st century (eLex 2023): Invisible lexicography. Proceedings of the eLex 2023 conference* (pp. 518–533). Lexical Computing CZ.

Kaschak, M. P. & Glenberg, A. M. (2000). Constructing meaning: The role of affordances and grammatical constructions in sentence comprehension. *Journal of Memory and Language,* 43(3), 508–529. https://doi.org/10.1006/jmla.2000.2705

Křen, M., Cvrček, V., Henyš, J., Hnátková, M., Jelínek, T., Kocek, J., Kováříková, D., Křivan, J., Milička, J., Petkevič, V., Procházka, P., Skoumalová, H., Šindlerová, J. & Škrabal, M. (2020). *SYN2020: reprezentativní korpus psané češtiny*. Institute of the Czech National Corpus, Faculty of Arts, Charles University. http://www.korpus.cz

Maxwell, N. P., Huff, M. J., Hajnal, A., Namias, J. M., Blau, J. J. C., Day, B., Marsh, K. L., Meagher, B. R., Shelley-Tremblay, J. F., Thomas, G. F. & Wagman, J. B. (2024). Affordance norms for 2825 concrete nouns. *Behavior Research Methods,* 56, 8480–8491. https://doi.org/10.3758/s13428-024-02486-1

Qi, P., Zhang, Y., Zhang, Y., Bolton, J. & Manning, C. D. (2020). Stanza: A Python natural language processing toolkit for many human languages. In A. Celikyilmaz & T.-H. Wen (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations* (pp. 101–108). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-demos.14

Vokáčová, M. (in preparation). *Logické metonymie v češtině z pohledu korpusů, lidské kognice a jazykových modelů* [Doctoral dissertation, Charles University].

Contact information

Martina Vokáčová
Faculty of Arts, Charles University
martina.vokacova@ff.cuni.cz

Anna Marklová
Faculty of Arts, Charles University
anna.marklova@ff.cuni.cz

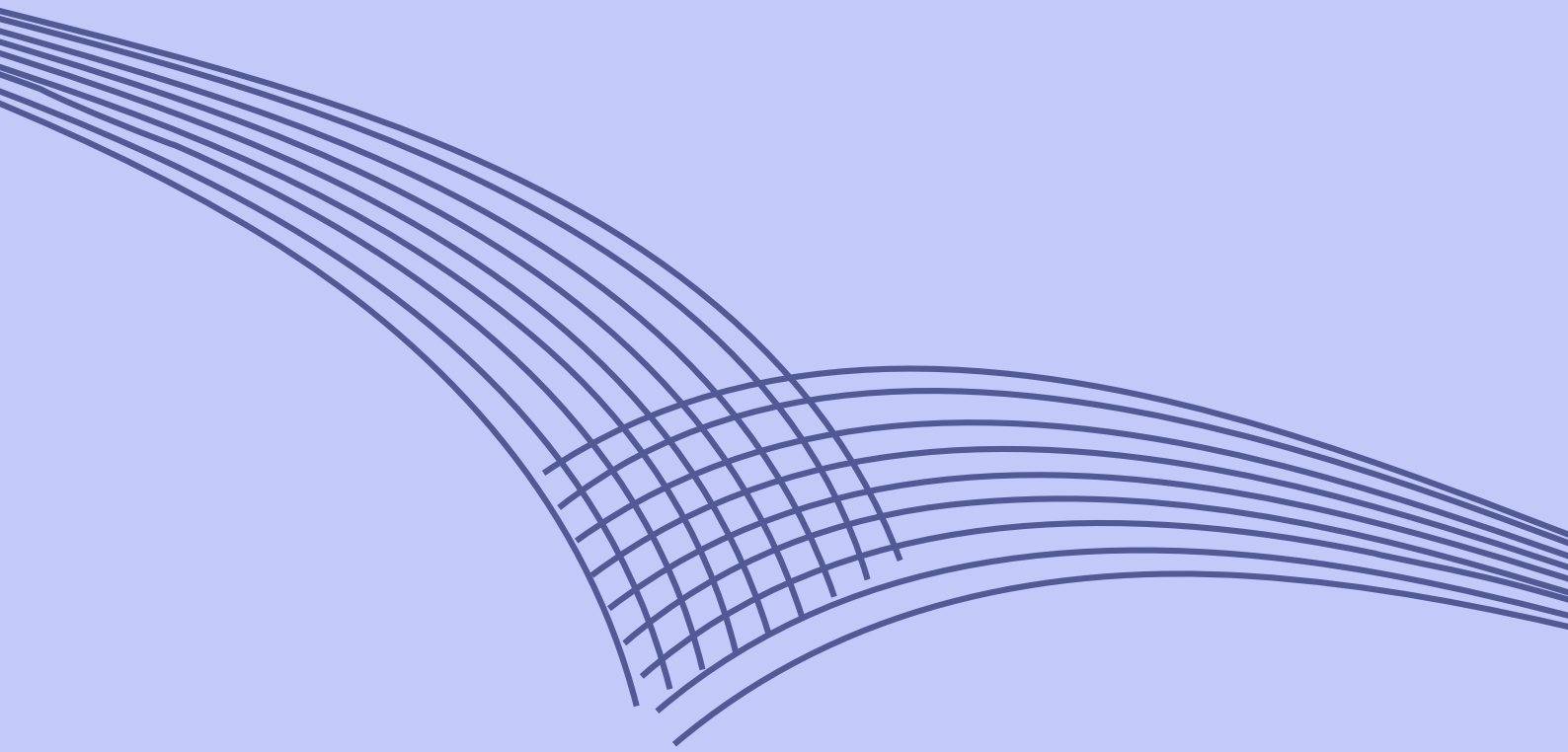
Acknowledgements

This research was supported by Primus Grant PRIMUS/25/SSH/010.

Declaration on using AI: Claude Opus 4.8 (Anthropic, 2026) was consulted during the development of the R scripts for data analysis and used for editorial assistance to improve clarity and style during the preparation of the manuscript. The authors carefully reviewed and edited all suggestions and take full responsibility for the content of the publication.

CORPUS LINGUISTICS & CORPUS-BASED RESOURCES

EURALEX 2026



Pol Garriga Martínez

THE VOCABULARY OF MENTAL HEALTH

Exploitation of Specialized Diachronic Corpora with Artificial Intelligence for Qualitative Semantic Analysis

Abstract This article explores the diachronic semantic evolution of mental health vocabulary, focusing specifically on the concept of hysteria across Spanish and German specialised corpora. Drawing on Diachronic Prototype Semantics, the paper examines how semantic changes reflect shifts in radial, hierarchical networks where meanings are updated according to communicative needs. The key methodological aspect is the use of the Gemini large language model to assist in the qualitative analysis of the MentEs and DeMens corpora contexts. This AI-supported approach enables the large-scale categorisation of contexts into six thematic areas: aetiology, clinical manifestations, differential diagnosis, treatment, conceptual debate, and non-clinical uses. The findings indicate that historical scientific discourse was overwhelmingly preoccupied with conceptual delimitation and identification of hysteria (representing 86–94% of occurrences) rather than therapeutic measures. Beyond the lexical and semantic analysis of hysteria, the validity of the methodology is presented as the primary finding. The author concludes that AI-assisted processing facilitates a comprehensive bridge between quantitative data and qualitative interpretation, allowing researchers to document prototypical modulation, for instance, how specific meanings gain or lose prominence depending on the historical and social context.

Keywords Mental health; mental illness; artificial intelligence; AI; hysteria; corpus linguistics; terminology

1 Framework

1.1 Mental health: history and terminology

The history of mental health evolved from ancient supernatural and organicist frameworks (Porter, 2002; Nutton, 2013) to medieval theological explanations in Europe, contrasted by Islamic psychological care (Dols, 1992). The Renaissance initiated gradual medicalisation (Livianos Aldana et al., 2006), followed by the seventeenth-century *grand renfermement*, which focused on social control (Foucault, 1976). The Enlightenment introduced 'moral treatment', leading to Kraepelin's nineteenth-century nosological classifications (Porter, 2002; Shorter, 1997). The twentieth century was subsequently shaped by Freud's psychoanalysis alongside invasive somatic therapies (Shorter, 1997; Oakes, 2007). During World War II, patients faced severe marginalisation, abandonment, or murder under regimes promoting so-called public protection and racial hygiene (Novella, 2023; Campos Pérez, 2014). Post-war developments were defined by the anti-psychiatry movement, the rise of the pharmaceutical industry, and a shift towards pharmacological treatment and social integration (Shorter, 1997; Mayes et al., 2005).

As argued by Guerrero Ramos (1995), specialised terminology reflects processes of institutionalisation of knowledge, where shifts in conceptual boundaries are mirrored by terminological change. This observation is particularly relevant in diachronic studies: replacement or coexistence of terms may indicate competition between theoretical models or phases of paradigmatic transition. In this regard, Gutiérrez Rodilla (2003, p. 16) emphasises that terminological consolidation is usually linked to processes of editorial standardisation, scientific translation and international diffusion of knowledge. These ideas are related to the points made in § 1.2. The way in which speakers conceptualise reality and its evolution is reflected in language through

linguistic change. This study examines how this occurs in the field of mental health through the concept of hysteria.

1.2 Diachronic Prototype Semantics

The study of lexical and semantic change operates on the premise that the lexicon constitutes the most dynamic level of the linguistic system, for it immediately reflects social, cultural and cognitive transformations of speech communities (Álvarez de Miranda, 2009, p. 135). Within the framework of cognitive linguistics, language is understood as part of general cognition and meaning is analysed in relation to mental processes and experience (Gibbs, 1996, p. 49). In this paradigm, authors such as Langacker (1987, 1991) maintain that linguistic units are cognitive schemata emergent from usage; therefore, linguistic structure does not precede communication but arises from it.

The fundamental cognitive operation in this process is categorisation, by which speakers order conceptual reality (Fernández Jaén, 2014, p. 62). Rosch (1975, 1978) demonstrates that these categories possess a graded internal structure and fuzzy boundaries; prototypical members serve as the primary representatives of their class, whilst peripheral members present graduated membership (Cuenca et al., 1999, p. 33). Lakoff (1987, p. 165) proposes that prototypes are idealised cognitive models constructed by the mind from shared cultural experience.

From a diachronic and linguistic perspective, this conception allows for the interpretation of linguistic change as a means for speakers to update language. The language's hierarchical and radial structure ensures its stability, while its dynamism allows language to evolve and adapt to the speakers' changing communicative needs. This tension constitutes the balance that enables the functionality of language (Geeraerts, 1997, p. 113).

This reasoning applies to both lexical and semantic change. In terms of semantics, polysemy is not interpreted as a flaw in language efficiency, but rather as the phenomenon through which language takes shape as an interconnected radial structure (Cuenca et al., 1999, p. 38). Following Wittgenstein's notion of family resemblance, the distinct meanings of a word are not accidental homonyms, but extensions related to each other that form a network. A central meaning functions as a core from which peripheral senses are generated through mechanisms of motivated extension (Lakoff, 1987, p. 91). Semantic change is therefore interpreted as a progressive modification in the prototypical structure of a category, observable in the redistribution of its central and peripheral uses over time (Fernández Jaén, 2014, p. 69).

The way in which these meanings relate to their signifiers determines the prototypical structure of the lexicon. Neologisms arise from the speaker's need to conceptualise a reality for which a new word is deemed necessary, thereby incorporating it into the lexicon. Similarly, lexical obsolescence represents the highest degree of peripheralization (Tichý, 2021); a word falls into disuse when it loses its functionality to categorise experience efficiently within the dynamic balance of the system (García Palacios et al., 2023).

This dynamism is driven by the speaker's communicative need to conceptualise new experiences (Geeraerts, 1997, p. 25; Blank, 2001, p. 96). Linguistic factors, such as homonymy or synonymy (Carriazo et al., 2010, p. 27), as well as extralinguistic factors, such as cultural, social or pragmatic factors (Cabré, 1999), can have a direct influence on this process (Geeraerts, 1997, pp. 32-46).

In this paper, Diachronic Prototype Semantics and the conception of language as a radial hierarchical structure are used as an interpretative framework for analysing semantic change in the field of mental health, specifically through the concept of hysteria.

1.3 Corpus linguistics: beyond frequency

Corpus linguistics has established itself as a rigorous empirical methodology for the scientific study of language, allowing for the analysis of usage patterns and frequencies on quantifiable bases (Parodi et al., 2022). However, statistical frequency does not automatically equate to meaning (Sinclair, 1991) but rather constitutes the observable manifestation of deeper cognitive processes.

To transcend simple frequency counts, it is essential to consider the history of science and various extralinguistic factors (Lara, 2004). This perspective allows researchers to situate terms within their specific discourse traditions (Martí Sánchez, 2021; Schrott, 2017), which aligns with the previously mentioned fundamental principles of diachronic terminology. This field focuses on the analysis of language in relation to the historical and epistemological shifts experienced by the scientific community (Cabré, 1999). Thus, while the construction of a corpus must be governed by criteria of representativeness and balance, prioritising the historical communicative impact of texts (McEnery et al., 2001; Müller et al., 2016), the goal is to understand the underlying conceptual changes.

As Rojo (2021) points out, corpus statistics allow for the identification of lexical transition. Biber (1988) emphasised that a complete understanding of a linguistic unit is inherently multidimensional and can be achieved through a combination of macroanalysis and microanalysis (Biber, 1988, pp. 62–63). Baker (2006), taking up these ideas, pointed out that, whilst frequencies are relevant—since language is not random—they can be overly generalising and simplistic, and that an approach combining quantitative and qualitative analysis will be more productive than one that is exclusively quantitative (Baker, 2006, p. 71).

Consequently, the data provided by quantitative analysis, such as a significant rise or fall in the frequency of a term, should be utilised as a starting point for a more profound qualitative lexical and semantic analysis. For this task, the frameworks provided by cognitive linguistics are considered especially suitable. The proposals of Geeraerts (1997, 2010), commented in §1.2, allow for corpus data to be interpreted as shifts within radial lexical and semantic networks, which are constituted by prototypical and peripheral elements. In the Hispanic context, authors such as Álvarez de Miranda (2009) and Cifuentes Honrubia (1990) have demonstrated that this conceptual reconfiguration is systematic and detectable across broad documentary series.

Ultimately, this approach seeks to bridge the gap between statistical evidence and deep semantic interpretation. By integrating the history of science with cognitive semantics, the methodology ensures that the study of specialised language is not limited to isolated signifiers. Instead, it observes the complete linguistic evolution around mental health ideas considering the historical, linguistic and social context.

2 Methodology

This study utilises the MentEs and DeMens corpora (Garriga Martínez, 2026). These resources comprise written, specialised, historical and diachronic data. MentEs corpus comprises 11,601,645 tokens from 393 files, whilst DeMens contains 8,488,770 tokens across 83 files. Each database includes a contrastive corpus of general press articles to enable terminological extraction and the calculation of statistical relevance.

The compilation of works followed four key principles: historiographical relevance, author diversity, balance in length, and thematic focus. It includes reference treatises by authors such

as Freud, Giné i Partagàs and Kraepelin, alongside official journals such as Revista Frenopática Española and *Archiv für Psychiatrie und Nervenkrankheiten*.

Terminological extraction was performed via the Sketch Engine platform (Kilgarriff et al., 2014), leveraging the concept of 'keyness' through the 'simple maths' method to identify lexical items with significantly higher frequency in medical texts compared to contemporary general language. This process generated lists of 1,000 ranked candidate terms for each language, which were manually reviewed to eliminate digitisation errors. The criteria for selecting the final terms for analysis include sufficient frequency of occurrence, the observation of marked changes in chronological frequency, and comparison with the historical theoretical framework. A pre-analysis onomasiological grouping was ruled out due to the ambiguous usage of certain terms, such as *psiquiátrico*. Instead, a semasiological approach was adopted based on the Greek root ὑστέρα (hystéra, 'uterus'), selecting for Spanish *histeria, histerismo* and *histérico* (both as an adjective and as a noun), and for German *Hysterie, hysterisch, Hysterische, Hysteriker*.

As detailed in Garriga Martínez (2026, p. 58), quantitative analysis constitutes the initial phase of the qualitative study, based on the premise that variations in frequency indicate linguistic changes (§ 1.3). However, this article focuses on the phase of semantic qualitative analysis and its execution with the support of artificial intelligence.

The specific aim of this phase is to examine the usage patterns of each term across the corpus, identifying patterns of usage that may remain constant or change over specific periods; in this regard, the approach is based on Baker's (2006, pp. 56–59) ideas on semantic analysis through the categorisation of usage patterns. This approach is in line with the research of Schuster, who analyses how the evolution of medical language reflects and influences shifts in medical paradigms and social context (Schuster, 2001, p. 157), through observable linguistic patterns such as the progressive trend towards objectification through nominalisation in the late 18th century (Schuster, 1999, p. 187), or the reinforcement of ideas of genetic pathological inheritance during National Socialism to justify policies of social exclusion and eugenics (Schuster, 1999, p. 184), amongst others.

In carrying out this semantic analysis, the size of the databases compiled—the MentEs and DeMens corpora—has been considered particularly valuable. Consequently, the extraction of a random sample for qualitative analysis was not deemed optimal due to the inherent risk of omitting significant contexts. To analyse the corpora in their entirety whilst maintaining a rigorous qualitative approach that is as detailed as possible (Baker, 2023), the Gemini Large Language Model (Google Gemini Team, 2023) was implemented as a processing aid.

The selection of Gemini, developed by Google DeepMind, arises from an institutional partnership between Pompeu Fabra University and Google Workspace for Education. As a researcher at said university, access to Gemini's tools is provided under a corporate licence that guarantees data confidentiality and unlimited processing capacity—two key features for this research. On the one hand, the institutional agreement ensures that the data provided is not used to train Google's models, thereby preventing any infringement of copyright in the works comprising the corpus. On the other hand, access to Gemini's most advanced engines and the ability to use them without input limits enables massive and systematic processing of contexts for the terms under study extracted from the corpora. Furthermore, Gemini's advanced models have demonstrated strong performance in linguistic reasoning, excelling in text classification and categorisation tasks and supporting long and complex text strings (Akter et al., 2023). Therefore, the use of Gemini for this task is based on its confidentiality, unlimited use and strong performance in linguistic tasks.

The subsequent phase of the qualitative analysis involves the categorisation of contexts based on usage patterns. Using spaCy, all contexts for each term were extracted. The result is an Excel-editable file, organised chronologically and by work due to the file naming convention, in which the full context of each occurrence of the analysed term can be found.

To establish the classification categories for the uses of both groups, the model proposed by Garriga Martínez and Campos Martín (under review) has been drawn, which qualitatively analyses the treatment of the concept of hysteria in general and specialised dictionaries through their definitions from a diachronic and interlinguistic perspective. Regarding the definition of the condition, the study identifies its cause, symptoms and treatment as central aspects. As for the adjectives relating to the condition, the study identifies references to pathological phenomena and to people as the main lines of usage. In both groups, usage labels are also identified as relevant, taking into account their specialisation or general, colloquial or pejorative use. These aspects have proved to be key in the study of the evolution of hysteria (Garriga Martínez, 2026).

Taking these ideas as a reference, the terms under study were divided into two groups: terms that refer to the disease (*histeria, histerismo, Hysterie*) and adjectives related to it (the adjectives *histérico* and *hysterisch* and their substantivised forms). Following this, a basic classification structure has been established for each group, taking into account the aspects indicated previously. For reasons of conciseness, this article will only cover the terms from the first group. Beneath are the classification categories for the usage contexts of the terms under study, together with examples of the term *histeria* in the MentEs corpus (Garriga Martínez, 2026).

1. Aetiology. It encompasses the various theories that discuss the origin or triggers of hysteria, which are typically hereditary, traumatic, sexual or constitutional:

   > Así, pues, la etiología específica de la histeria está constituida por una experiencia de pasividad sexual anterior a la pubertad (*Obras completas del profesor S. Freud xi*, 1927. Translation by Luis López-Ballesteros y de Torres) [MentEs. Garriga Martínez, 2026].

2. Clinical manifestations. It includes descriptions of how hysteria manifests itself, its symptoms (such as convulsions or paralysis) and its various forms:

   > En la histeria son frecuentes las descripciones del dolor como un clavo, o sea de un dolor penetrante que taladra un pequeño punto circunscrito de las sienes o del occipucio (*Tratado de diagnóstico diferencial - neurología*, Honigmann, 1932. Translation by Antonio Vallejo Nágera and Francisco Piñero) [MentEs. Garriga Martínez, 2026].

3. Differential diagnosis and nosology. It groups together instances where hysteria is compared or contrasted with other conditions, such as epilepsy, neurasthenia or schizophrenia, in order to distinguish it, classify it or highlight the difficulty in doing so:

   > El diagnóstico diferencial con la histeria puede ofrecerse en casi todas las enfermedades organicas (*Tratado de diagnóstico diferencial - neurología*, Honigmann, 1932. Translated by Antonio Vallejo Nágera and Francisco Piñero) [MentEs. Garriga Martínez, 2026].

4. Treatment. It includes references to measures for curing, relieving or controlling hysteria, which typically refer to pharmacology, hydrotherapy, psychotherapy and psychoanalysis:

   > Nuestra técnica se ha desarrollado en el tratamiento de la histeria y permanece aún orientada hacia esta afección (*Obras completas del profesor S. Freud xiv,* 1930. Translation by Luis López-Ballesteros y de Torres) [MentEs. Garriga Martínez, 2026].

5. Conceptual debate and redefinition. It focuses on the theoretical debate surrounding the nature of hysteria. It covers the various definitions put forward by leading figures such as Charcot, Janet, Freud and Bumke, critiques of earlier concepts, and proposals for new paradigms:

> En el concepto de la histeria reina modernamente un verdadero caos: tan distintos son los puntos de partida (*Tratado de psiquiatría*, 1944) [MentEs. Garriga Martínez, 2026].

6. Non-clinical uses. It covers instances where hysteria is mentioned in contexts outside the clinical sphere, such as social, political or legal contexts.

> Cuando el historiador de arte GUILLERMO WORRINGER habla de «la sublime histeria del gótico», no hace una metáfora ingeniosa, sino una comparación exacta (*Tratado de diagnóstico diferencial - neurología*, Honigmann, 1932. Translation by Antonio Vallejo Nágera and Francisco Piñero) [MentEs. Garriga Martínez, 2026].

These classification structures have been developed by combining the key aspects identified in Garriga Martínez & Campos Martín (under review) with the analytical requirements of each term and its specific characteristics; for this reason, these classifications have been developed, refined and adapted to each term during its qualitative analysis, so that there are no empty categories, nor any context that cannot be classified in the established categories. For further details on the methodology and qualitative classification, see Garriga Martínez (2026).

Following these categories, Gemini was provided with the research objectives and instructed to classify the selected terms preserving the original format ("filename"; "context"; "classification"); to ensure stringent quality control, processing was conducted in batches of approximately 50 contexts, allowing for manual count and verification of ambiguous cases (see Garriga Martínez, 2026, pp. 99, 161). This has enabled the creation of a new CSV file, whose contexts can be classified by work, year and usage category of the term under study.

This initial classification of contexts according to usage patterns enables the calculation of statistical data which, straddling the line between quantitative and qualitative analysis, serves to observe the relative weight of certain usage patterns over others; this provides an indication of usage patterns that can be analysed in greater detail in subsequent phases.

3 Results and interpretations

The systematic categorisation of all contexts within the MentEs and DeMens corpora for the selected terms yields the following data.

TABLE 1 Absolute and relative frequency of each usage pattern for *Hysterie* (Garriga Martínez, 2026)

usage pattern	*histerismo*	*histeria*	*Hysterie*
3. differential diagnosis and nosology	119	273	418
5. conceptual debate and redefinition	103	241	204
2. clinical manifestations	99	239	154
1. aetiology	62	184	125
4. treatment	36	99	53
6. non-clinical	22	8	2
	441	1044	956

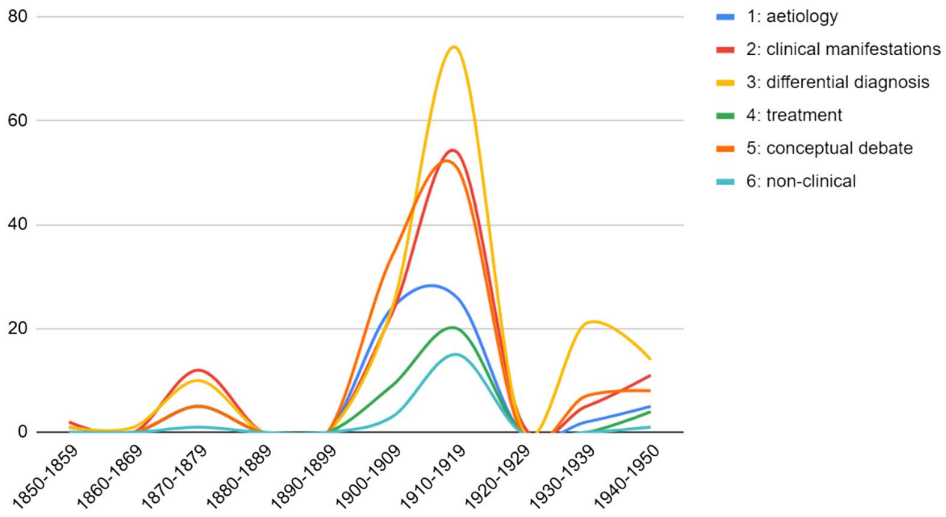


FIGURE 1 Timeline of categorised uses of *histerismo*

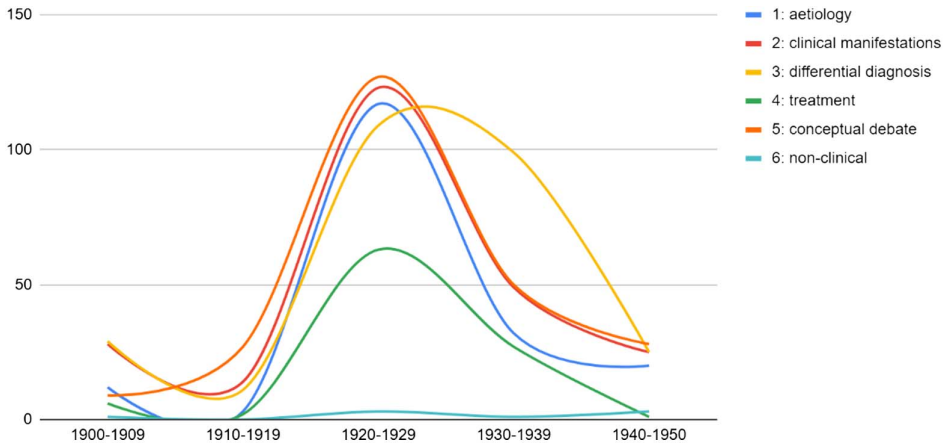


FIGURE 2 Timeline of categorised uses of *histeria*

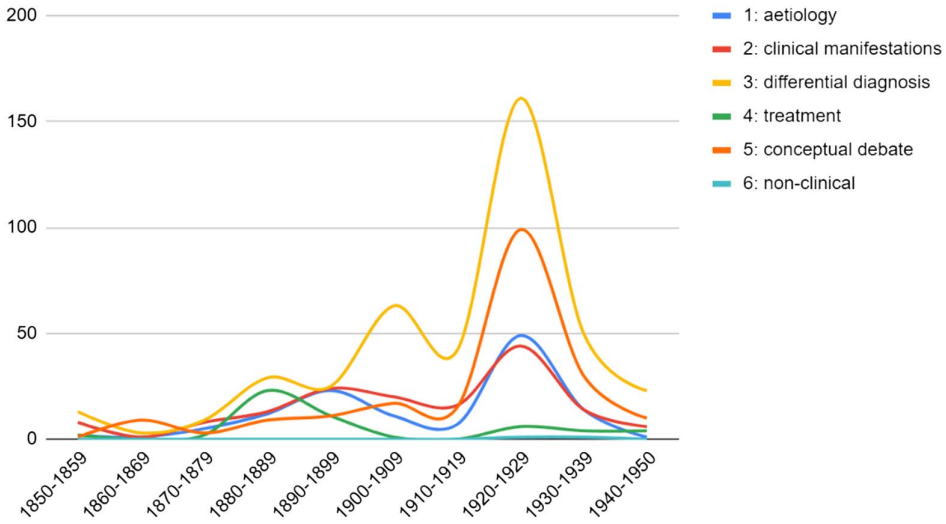


FIGURE 3 Timeline of categorised uses of *Hysterie*

From a lexical perspective, in Spanish, the MentEs corpus shows a discernible decline in the frequency of *histerismo* in the 1920s, the decade in which *histeria* attained its peak frequency. This transition indicates a prototypical modulation within the Spanish lexicon regarding the designation of the condition. In contrast, the German data shows that *Hysterie* remained the sole dominant term throughout the same period. Whilst this article primarily addresses the semantic perspective, these lexical shifts provide essential context for the underlying conceptual evolution; for further information on lexical change, see Garriga Martínez (2026, p. 227).

An analysis of the usage patterns for *histeria, histerismo*, and *Hysterie* reveals a consistent and shared trend towards the identification and conceptual **delimitation** of these terms within specialised discourse. In all three cases, categories associated with theoretical and clinical characterisation (categories 1, 2, 3, and 5) constitute the vast majority of occurrences. These categories account for 89.7% for *histeria*, 86.85% for *histerismo*, and 94.25% for the German *Hysterie*. This major thematic block encompasses the tracing of successive causal theories, the definition of symptoms, and a marked inclination towards epistemological reflection. A notable nuance within the German data is the preponderance of differential diagnosis (category 3), which accounts for nearly half of all occurrences, thereby underlining a superior analytical effort to compare and delineate the boundaries of *Hysterie* against other affections.

In sharp contrast, interest in treatment (category 4) occupies a secondary position across all datasets, representing 9.5% for *histeria*, 8.16% for *histerismo*, and falling to 5.54% for *Hysterie*. This relative paucity suggests that scientific discourse was primarily preoccupied with the theoretical understanding and nosological classification of the pathology, rather than detailing clinical cures. It may also indicate that therapeutic proposals (such as hydrotherapy or psychoanalysis) were derived directly from the prevailing medical doctrine of each period.

Finally, category 6, comprising usages outside the clinical sphere, exhibits the lowest frequencies, an expected result given the specialised nature of the analysed texts. However, qualitative nuances emerge depending on the term (Garriga Martínez, 2026): whilst the presence of *Hysterie* and *histeria* remains marginal and limited to figurative or metaphorical applications to social phenomena, constituting an incipient sign of its subsequent popularisation, *histerismo* reaches 4.99% and acquires a formal institutional character, frequently appearing in discourses concerning legal medicine, criminality and forensic responsibility.

Despite slight fluctuations in the prevalence of these categories during certain periods, the aforementioned trend towards the identification and definition of hysteria remains consistent.

4 Conclusions

The proposed methodology, involving large-scale semantic categorisation assisted by Gemini, provides interpretable and representative data within a line of research that bridges quantitative and qualitative analysis. By integrating frequency data and a semasiological perspective with semantic and onomasiological interpretation, this approach aligns with the tenets of Diachronic Prototype Semantics (Geeraerts, 1997), which argues that meaning is a dynamic conceptualisation process structured in radial, hierarchical and dynamic networks. By classifying 100% of the extracted contexts, this research transcends the analysis of isolated examples to examine the comprehensive semantic architecture of the terms *histeria, histerismo* and *Hysterie*.

The selection of usage categories for classification, developed from the key aspects identified by Garriga Martínez and Campos Martín (under review) and aligned with the research objectives of Garriga Martínez (2026), is fundamental to this methodology. This classification framework remains open to academic debate, particularly concerning its level of granularity. For instance,

categories 1, 2, 3 and 5 could be aggregated into a simplified super-category related to the identification of hysteria, contrasted with categories 4 and 6, depending on specific research requirements. Conversely, the granular breakdown into several subcategories occasionally results in overlap, that is, there are contexts that could be grouped into more than one category. This is particularly evident in categories 3 and 5, as the difficulty in distinguishing hysteria from other illnesses often triggers conceptual debates regarding its very nature. This reflects Ullmann's (1962, p. 219) observation that conceptual vagueness and blurred boundaries in conceptual categorisation are precisely what allow for semantic change.

In conclusion, this semantic analysis—which straddles quantitative and qualitative approaches—enables the derivation of detailed conclusions whilst working with large-scale, curated, and representative data. Specifically, in this case, it highlights the historical preponderance of scientific efforts directed towards identifying and defining hysteria rather than its clinical treatment. Nevertheless, as mentioned in § 1.3, it is advisable that this data be utilised as a departure point for further exploration through more granular, qualitative analysis of individual contexts, as is done in Garriga Martínez (2026). The nuances within category 6, which encompasses diverse domains and meanings for each term, exemplify the depth of information that figures alone may obscure (see the final paragraph of § 3). In this regard, diachronically, systematic classification across these six enables researchers to document how certain meanings gain prominence depending on the historical context whilst others shift to a peripheral position, providing empirical evidence of prototypical modulation (Garriga Martínez, 2026, pp. 136, 225).

References

Álvarez de Miranda, P. (2009). Neología y pérdida léxica. In E. de Miguel (Ed.), *Panorama de la lexicología* (pp. 133-156). Ariel.

Baker, P. (2006). *Using corpora in discourse analysis*. Continuum.

Biber, D. (1988). *Variation across Speech and Writing*. Cambridge University Press.

Blank, A. (2001). *Words and concepts in time: Towards diachronic cognitive semantics*. Mouton de Gruyter.

Cabré, M. T. (1999). *La terminología: representación y comunicación*. Universitat Pompeu Fabra.

Campos Pérez, M. P. (2014). *"Eutanasia" y nazismo*. Universidad Pública de Navarra.

Carriazo Ruiz, J. R. & Giménez Eguíbar, P. (2010). Procesos de sustitución léxica en el tecnolecto naval del Siglo de Oro: neología frente a obsolescencia. In R. Verdonk & M. J. Mancho Duque (Eds.), *Aspectos de la Neología en el Siglo de Oro: lengua general y lenguajes especializados* (pp. 15-40). Rodopi Brill.

Cifuentes Honrubia, J. L. (1994). *Gramática cognitiva. Fundamentos críticos*. Eudema.

Cuenca, M. J. & Hilferty, J. (1999). *Introducción a la lingüística cognitiva*. Ariel.

Dols, M. (1992). *Majnun: The Madman in Medieval Islamic Society*. Clarendon Press.

Fernández Jaén, J. (2014). Modelos cognitivos. In de E. de Miguel (Ed.), *Panorama de la lexicología* (pp. 281-298). Ariel.

Foucault, M. (1976). *Historia de la locura en la época clásica*. Fondo de Cultura Económica.

García Palacios, J., Sánchez Ibáñez, M. & Maroto García, N. (2023). La neología: encrucijada entre estabilidad y renovación. *CLINA: an interdisciplinary journal of translation, interpreting and intercultural communication*, 9(1), 9-11.

Garriga Martínez, P. (2026). *Estudio del cambio léxico y semántico en el lenguaje de la salud mental entre 1850 y 1950 en castellano y alemán: elaboración y explotación de dos corpus especializados a través del concepto de histeria* [Doctoral dissertation, Universitat Pompeu Fabra, Universität Potsdam].

Garriga Martínez, P. & Campos Martín, N. (in revision). *El concepto de histeria en francés a lo largo de la historia, a través de los diccionarios*.

Geeraerts, D. (1997). *Diachronic Prototype Semantics. A Contribution to Historical Lexicology*. Clarendon Press.

Geeraerts, D. (2010). *Theories of lexical semantics*. Oxford University Press.

Gibbs, R. (1996). Why many concepts are metaphorical. *Cognition*, 61(3), 309-319.

Guerrero Ramos, G. (1995). *Neologismos en el español actual*. Arco Libros.

Gutiérrez Rodilla, B. (2003). La historia del lenguaje científico como parte de la historia de la ciencia. *Asclepio*, 55(2), 7-25.

Lakoff, G. (1987). *Women, fire and dangerous things: What categories reveal about the mind*. University of Chicago Press.

Langacker Ronald, W. (1987). *Foundation of Cognitive Grammar. Vol. 1. Theoretical Prerequisites*. Stanford University Press.

Langacker Ronald, W. (1991). *Foundation of Cognitive Grammar. Vol. 2. Descriptive Application*. Stanford University Press.

Lara, L. F. (2004). *Lengua histórica y normatividad*. El Colegio de México.

Livianos Aldana, L. & Ciscar Vilata, C. (2006). *El manicomio de Valencia del siglo xv al xx. Del Spital dels Fols, Orats e Ignocents al Convento de Jesús*. Ajuntament de València.

Martí Sánchez, M. (2021). La ciencia y su discurso: una cala en la medicina española del s. XVIII. In E. Carpi & J. L. Ramírez Luengo (Eds.), *Estudios sobre historia del léxico de especialidad* (pp. 13-29). Editorial Universidad de Sevilla.

Mayes, R., Horwitz, A. (2005). DSM-iii and the revolution in the classification of mental illness. *Journal of the History of the Behavioral Sciences*, 41, 249-267.

McEnery, T., Wilson, A. (2001). *Corpus Linguistics: An Introduction*. Edinburgh University Press.

Müller, E. & Schmieder, F. (2016). Historische Lexikologie, Korpuslinguistik, Fachsprachenforschung. In E. Müller & F. Schmieder (Eds.), *Begriffsgeschichte und historische Semantik. Ein kritisches Kompendium*. Suhrkamp Verlag.

Novella, E. (2023). *Las políticas de la locura. Psiquiatría y sociedad en la España de Franco*. Universitat de València.

Nutton, V. (2013). *Ancient Medicine*. Routledge.

Parodi, G., Cantos-Gómez, P. & Howe, C. (2022). Introducción. In *Lingüística de corpus en español* (pp. 1-9). Routledge.

Porter, R. (2002). *Madness: A Brief History*. Oxford University Press.

Rojo, G. (2021). *Introducción a la lingüística de corpus en español*. Síntesis.

Rosch, E. (1975). Cognitive representations of semantic categories. *Journal of Experimental Psychology: General*, 104, 192-233.

Rosch, E. (1978). Principles of categorization. In E. Rosch & B. Lloyd (Eds.), *Cognition and Categorization* (pp. 27-48). Routledge.

Schrott, Á. (2017). Las tradiciones discursivas, la pragmalingüística y la lingüística del discurso. *Revista de la Academia Nacional de Letras*, 13, 25-57.

Shorter, E. (1997). *A history of psichiatry: from the era of the asylum to the age of Prozac*. John Wiley & Sons.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.

Tichý, O. (2021). Corpus driven identification of lexical bundle obsolescence in Late Modern English. In S. Kranich & T. Breban (Eds.), *Lost in Change. Causes and processes in the loss of grammatical elements and constructions* (pp. 101-129). John Benjamins Publishing Company.

Contact information

Pol Garriga Martínez
Universitat Pompeu Fabra, Universität Potsdam
pol.garriga@upf.edu

Mariia Kopylova

AI-ASSISTED DETECTION AND MITIGATION OF LEXICOGRAPHIC BIAS

Semantic Reduction in Definitions of Five Socio-Legal Terms

Abstract When non-specialist users consult general dictionaries to understand legally significant terms, they risk encountering definitions that are fundamentally incomplete. This study introduces a reproducible computational methodology to quantify the level of semantic reduction and explore AI tools to resolve it. Five socio-legal terms were examined from the Cambridge Dictionary, Oxford English Dictionary, Merriam-Webster (general) and Black's Law Dictionary, Oxford Dictionary of Law, Merriam-Webster's Dictionary of Law (legal): *harassment, stalking, discrimination, abuse*, and *neglect*. General and legal definitions (35 and 15 respectively) were vectorised using Sentence-BERT and compared via cosine similarity. None of the five terms reached the 0.70 similarity threshold. Mean similarity ranges from 0.491 (*abuse*) to 0.571 (*stalking*); 92.4% of all pairwise comparisons fall below this threshold. Statistical significance was confirmed by one-sample t-tests ($p < 0.01$). The Qwen2.5-3B-Instruct model extracted key legal components for each term, which can serve as editorial checklists for lexicographers reviewing dictionary entries. The findings demonstrate that semantic reduction is systematic, measurable, and varies across terms. AI tools can assist lexicographers in detecting and addressing such gaps, though final decisions remain with the human expert.

Keywords Lexicographic bias; semantic reduction; Sentence-BERT; cosine similarity; NLP; legal lexicography

1 Introduction

Dictionaries are traditionally regarded as authoritative and neutral sources of information. However, lexicographic decisions can inadvertently reinforce social and cultural biases. As Kaplan (2020, p. iv) observes, "a problem in the current lexicographic practice is a lack of inclusive and bias-free expression dictionaries"; general dictionaries "do not do an adequate job of alerting users to this [biased and exclusive] usage" (Kaplan, 2020, p. 65). When a dictionary reduces a legally significant term to a simple everyday concept, users risk misunderstanding their rights and legal status.

In this paper, *bias* refers to the systematic, uncompensated omission of legally salient meaning from general-purpose definitions – not merely to any difference between general and specialised sources, which is expected given their different communicative aims (see 2.1). In the present study *semantic reduction* denotes the specific, measurable form this omission takes, i.e., a low cosine-similarity score between corresponding general and legal definitions. Cosine similarity is a mathematical measure that quantifies the semantic closeness between two texts; values range from 0 (no similarity) to 1 (identical meaning).

The word *harassment* serves as a pertinent example. The Cambridge Advanced Learner's Dictionary defines it as "behaviour that annoys or upsets someone." In contrast, the Oxford Dictionary of Law interprets *harassment* as a form of prohibited discrimination involving unwanted conduct that violates dignity and creates a hostile environment. As Tobia (2020) demonstrated, "neither the dictionary nor legal corpus linguistics methods reliably track ordinary people's judgments about meaning", and that "dictionary and legal corpus linguistics verdicts diverge dramatically from each other" (p. 727). This discrepancy is not unique; it is typical of many socio-legal concepts.

The present study analyses five concepts: *harassment*, *stalking*, *discrimination*, *abuse*, and *neglect*. For each term, a corpus of general definitions was compiled from the Cambridge Dictionary (including the Advanced Learner's, Academic Content, and Business English editions), Oxford English Dictionary, and Merriam-Webster, alongside a corpus of legal definitions sourced from Black's Law Dictionary, the Oxford Dictionary of Law, and Merriam-Webster's Dictionary of Law. The primary objective is to quantify the semantic gap between these two types of sources using computational methods and to explore the potential of AI in detecting and bridging such discrepancies. The specific tasks are: (1) measuring the semantic gap for the five terms; (2) applying Transformer-based models – Sentence-BERT to compare definitions and a generative model to extract key components; and (3) demonstrating the practical role of AI as an auxiliary tool for lexicographers in the digital age (Lew, 2024).

2 Background

2.1 General Language vs. Language for Special Purposes (LSP) Dictionaries

General-purpose dictionaries describe everyday language, whereas legal dictionaries prioritize terminological accuracy within the legal system. As Bajčić (2017, p. 7) notes, "law is not a discipline of words, but of concepts" with meaning shaped by an extralinguistic context. General dictionaries are designed to describe how a word is actually used across a speech community, prioritising frequency and accessibility while still aiming for accuracy appropriate to general use. Their primary goal is descriptive coverage, which does not necessarily require the same level of terminological precision as a legal dictionary. Legal dictionaries, by contrast, encode terminology whose meaning has often been fixed by statute or case law, so that precision – not usage frequency – is the guiding criterion. This functional difference is not a flaw in either type of source; it reflects genuinely different target users and communicative goals (cf. Bajčić, 2017). The problem this study addresses is not that these differences exist, but that general dictionaries rarely signal when a term also carries a distinct, legally binding meaning elsewhere. Tiersma (1999, p. 111) refers to words that "seem like ordinary words but are not" such as legal homonyms, which constitute the focus of this study.

The existence of multiple meanings for the same word is a common phenomenon in English. The problem arises when dictionaries fail to indicate these differences or warn users about the limitations of their definitions in legal contexts. This implies that users consulting a general dictionary for the definition of a legally significant term risk receiving fundamentally incomplete information without being aware of it. This study argues that general English dictionaries should either include the legal meanings of socially significant terms or explicitly direct users to specialized sources.

2.2 Quantitative Measurement of the Semantic Gap

If there is a discrepancy between the general and legal meanings of a word that has social consequences, such as misidentification of legal rights or underestimation of harm, the question arises: can it be measured? Research in the field of language models shows that semantic differences can indeed be quantified. Durrheim et al. (2022, p. 617) demonstrate that word embeddings "provide quantitative representations of word semantics ... and they can identify subtle biases that may not be communicated explicitly." Puttick et al. (2025) further confirm that such biases are systematically reflected in language models. If bias in language can be mathematically encoded and measured, then the semantic gap between general and legal

definitions can be quantified in the same way – this is the methodological premise of the present study. Specific analytical tools and procedures are described in Sections 3 and 4.

3 Methodology

Five terms were selected as examples of what Tiersma (1999, p. 111) calls *legal homonyms* – words that "seem like ordinary words but are not" and "can easily engender confusion" among non-specialists. Unlike strictly technical legal terms, *legal homonyms* appear in everyday language and are therefore highly likely to be searched for in general dictionaries, making the accuracy of such definitions particularly important. *Harassment*, *stalking*, *discrimination*, *abuse*, and *neglect* meet this criterion: all five are common in everyday speech, have well-established legal definitions, and carry significant social or ethical implications.

Two corpora were constructed for the analysis. The selection of dictionaries encompassed both American and British lexicographic and legal traditions; the chronological range (1996–2026) allowed for a comparison of current editions.

The total corpus comprises 35 definitions from three dictionaries (Table 1). The Cambridge Dictionary is presented in three editions: Advanced Learner's, Academic Content, and Business English, allowing coverage of definitions across different registers within the same publisher. The Oxford English Dictionary offers a historical and descriptive perspective, while Merriam-Webster provides a modern American norm. Between four and ten definitions were selected for each term; historical, metaphorical, and non-legal meanings were excluded (for example, the hunting meaning of *stalking*), as were usage examples, etymological information, and meta-comments. These were excluded to ensure that only core definitional content was compared, isolating the semantic information a user typically receives when looking up a term, rather than contextual or historical annotations.

TABLE 1 Composition of the general dictionary corpus

Dictionary	Type	Edition
Cambridge Dictionary	Learner's / Academic / Business	2025–2026 online
Oxford English Dictionary	Historical / descriptive	Online (revised entries 2003–2023)
Merriam-Webster Dictionary	Contemporary descriptive	2025–2026 online

The legal corpus comprises 15 definitions sourced from three reputable legal dictionaries (Table 2): the American Black's Law Dictionary and Merriam-Webster's Dictionary of Law, as well as the British Oxford Dictionary of Law. For each term, the definitions were included in full; only meta-information (e.g., [Cases: . . .]) and usage examples were excluded. As with the general corpus, only the core definitional wording was retained, so that the comparison reflects the substantive legal content of each entry rather than procedural annotations or case citations.

TABLE 2 Composition of the legal dictionary corpus

Dictionary	Type	Edition
Black's Law Dictionary	American Legal Handbook	9th Edition, abridged (2010)
Merriam-Webster's Dictionary of Law	American Legal Handbook	1st Edition (1996)
Oxford Dictionary of Law	British Legal Handbook	10th Edition (2022)

The final corpus consisted of 105 pairwise comparisons: 35 general definitions and 15 legal definitions for five terms. The number of common definitions ranged from 4 (*stalking*) to

10 (*abuse*), reflecting the varying degrees of polysemy of these terms in everyday language. For each term, three legal definitions representing different jurisdictions and traditions were used.

4 Results

To assess the degree of semantic discrepancy between general and legal dictionaries, two complementary methods were employed: a qualitative analysis using the Voyant Tools platform and a quantitative analysis based on vector representations (Sentence-BERT). To illustrate the methodology in detail, the term *harassment* is examined first; the same procedure was then applied to the other four terms, with the results summarized at the end of the section.

4.1 Qualitative Analysis: Voyant Tools

Voyant Tools is a web-based text corpus analysis tool widely used in the humanities (Alhudithi, 2021). It enables users to visualize word frequency and compare the lexical profiles of different text collections. The tool is commonly employed in lexicography, discourse analysis, and digital humanities for primary corpus exploration. Two corpora were uploaded separately to Voyant Tools: a general corpus of *harassment* definitions (comprising seven definitions from the Cambridge Dictionary, Oxford English Dictionary, and Merriam-Webster) and a legal corpus (containing three definitions from Black's Law Dictionary, Oxford Dictionary of Law, and Merriam-Webster's Dictionary of Law). The most frequent lexical units in each corpus were then compared.

The results were clear: the general corpus is dominated by words associated with emotional reactions – such as *behaviour* (4 occurrences), *annoy* (3) and *upset* (2). In the legal corpus, the picture is different: normative and behavioural categories take precedence – *conduct* (3), *hostile environment* (3 occurrences), *sexual* (4), *offense* (2) (Figure 1). This difference is not surprising, as common and legal languages serve different communicative purposes. However, both cases describe the same type of interpersonal conduct: unwanted behaviour directed at another person. The fact that definitions of the same term differ so substantially in meaning that the model perceives them almost as descriptions of different phenomena is no longer a matter of style but one of information completeness.



FIGURE 1 Keyword frequency visualisation for the legal corpus of *harassment* (Voyant Tools) [Source: generated by the author using Voyant Tools]

4.2 Quantitative Analysis: Sentence-BERT and Cosine Similarity

A quantitative analysis was conducted to confirm and formalize this observation. Sentence-BERT ('all-MiniLM-L6-v2') (Reimers & Gurevych, 2019) operates as follows: each definition

is input into the model, which returns a vector of 384 numbers – a kind of "semantic fingerprint" of the text. Definitions that describe the same concept yield similar vectors, while those describing different concepts produce dissimilar ones. Unlike methods that analyse individual words, Sentence-BERT captures the meaning of an entire sentence, which is especially important for comparing detailed dictionary definitions. This model is widely used for tasks such as semantic search, duplicate text detection, and measurement of semantic similarity.

To compare two such semantic fingerprints, cosine similarity – a standard mathematical measure – is used to determine how closely two numeric vectors "point in the same direction." In simple terms, cosine similarity measures the angle between two vectors: the smaller the angle, the more similar the meanings. A value of 1 means the vectors point in exactly the same direction (identical meaning), while 0 means they are orthogonal (completely unrelated). The result is a value between 0 and 1: the closer to 1, the greater the semantic similarity between the two definitions. In this study, a threshold of 0.70 is adopted as an exploratory operational benchmark rather than a universal semantic boundary; it is not intended to represent an absolute criterion of semantic equivalence, and future research may calibrate this value empirically for specific term types (Patel & Joshi, 2025).

Each of the seven general and three legal definitions of *harassment* was transformed into a vector. Then, for all 21 possible "general–legal" pairs, the cosine similarity was calculated. The results are presented as a heat map (Figure 2), with the three legal definitions arranged horizontally and the seven general definitions vertically. The colour of each cell reflects the degree of similarity: red indicates a low value, while green indicates a high value. The pattern is clear: 20 out of 21 pairs (95.2%) fell below the threshold of 0.70. The average similarity was 0.547 ± 0.092.

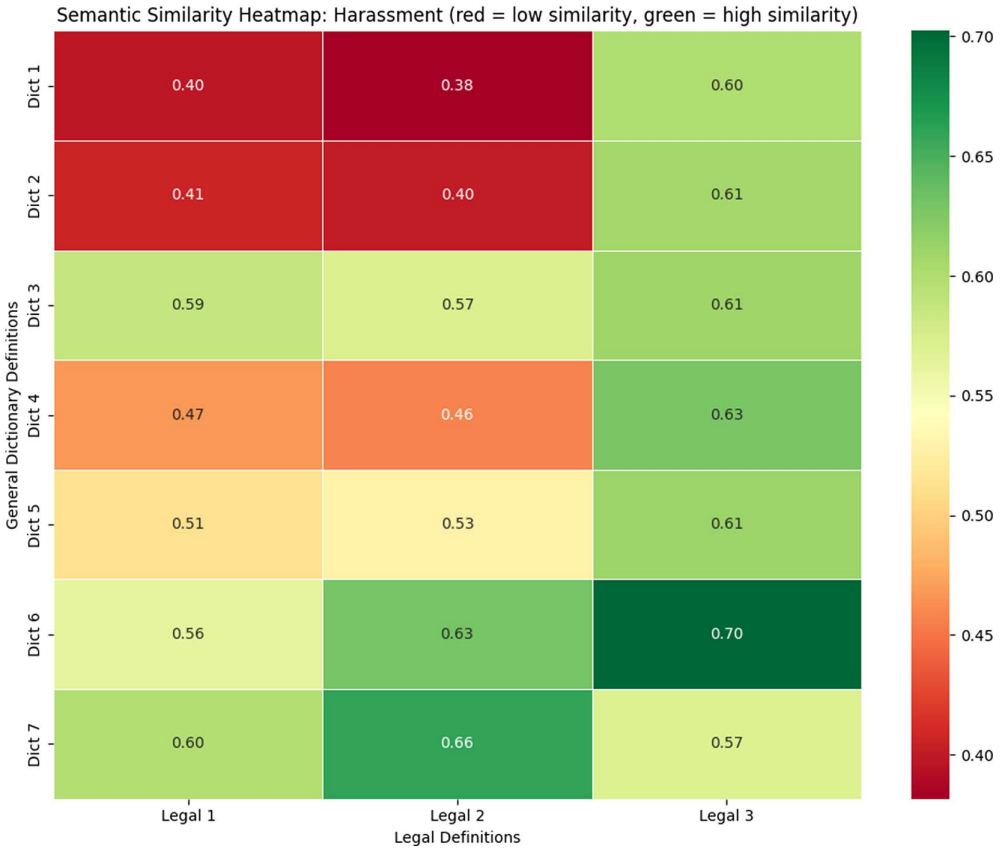


FIGURE 2 Semantic similarity heatmap for *harassment* (cosine similarity between general and legal definitions) [Source: generated by the author using Python]

The lowest similarity value (0.382) was recorded between the Cambridge Advanced Learner's Dictionary definition – "behaviour that annoys or upsets someone" – and the Oxford Dictionary of Law definition, which includes elements such as *unwanted conduct*, *violation of dignity*, and *hostile environment*. The Cambridge definition proved to be the most reductive overall, with an average similarity of only 0.459 when compared to all three legal definitions. Notably, this is the definition users encounter first when opening the Cambridge Dictionary. It is acknowledged that the Cambridge Advanced Learner's Dictionary is designed with pedagogical simplification in mind. This does not, however, diminish the relevance of its definitions for the present study: learner's dictionaries are among the most widely consulted general references, and it is precisely these definitions that non-specialist users encounter first.

Thus, the quantitative analysis confirms what the qualitative analysis suggested: the discrepancy between the general and legal understandings of *harassment* is neither accidental nor isolated but systematic and measurable.

4.3 Extracting Semantic Components Using AI

Having identified the problem, a step was taken toward its mitigation. To achieve this, the Qwen-2.5-3B-Instruct model was used, a compact, open-source generative language model optimized for instruction-following tasks. It was selected for its reproducibility and sufficient capacity to extract structured information from text without requiring access to closed APIs. The model was tasked with identifying the key semantic components of the term *harassment* according to legal dictionaries and presenting them as short noun phrases.

The model identified the following components: *offense, conduct, unwanted, repeated, intimidating, threatening, humiliating, causing fear, persistent* – among others. All these elements are absent from the most reductive general definition. Such a list can serve as an editorial checklist for lexicographers reviewing dictionary entries: if none of these components are reflected in a definition, it signals a potential semantic reduction. The same extraction was performed for all five terms. The resulting components, presented in Table 3, can serve as reference checklists for lexicographic revision.

TABLE 3 Key semantic components extracted by Qwen2.5-3B-Instruct

Term	Extracted Semantic Components
Harassment	offense; conduct; unwanted; repeated; intimidating; offensive; threatening; humiliating; interfering; bullying; distressing; causing fear; unwelcome; persistent; harassing; offensive behavior; abusive; vexatious; oppressive; offensive conduct; offensive acts; offensive treatment
Stalking	individuals; pursuit; harassment; monitoring; unwanted; presence
Discrimination	gender; sex; race; color; religion; nationality; age; disability; sexual orientation; disability status; marital status; family status; genetic information; protected classes; unlawful discrimination; protected characteristic; protected group; protected status
Abuse	abusive behavior; abuse of authority; abuse of process; emotional abuse; physical abuse; sexual abuse; neglect; exploitation; financial abuse; psychological abuse
Neglect	children; parents; care; supervision; responsibilities; neglectful; caregiver; omission; duty; failure; abuse; deprivation; harm; neglecting; mistreatment; provide; adequate; attention

4.4 Results for the Remaining Four Terms

The same procedure – qualitative analysis using Voyant Tools, vectorization with Sentence-BERT, calculation of cosine similarity for all pairwise comparisons, and component extraction by the Qwen model – was applied to the terms *stalking, discrimination, abuse*, and *neglect*. The summary results are presented in Table 4.

TABLE 4 Semantic similarity between general and legal definitions: summary results for five terms

Term	Average Similarity	Standard Deviation	% of Pairs < 0.70	95% Confidence Interval
Harassment	0.547	0.092	95.2%	[0.504, 0.590]
Stalking	0.571	0.136	91.7%	[0.480, 0.661]
Discrimination	0.532	0.049	100%	[0.507, 0.557]
Abuse	0.491	0.197	83.3%	[0.416, 0.566]
Neglect	0.551	0.128	95.8%	[0.496, 0.606]

The results for all five terms show a consistent pattern: none reached the threshold value of 0.70. The average similarity scores range from 0.491 (*abuse*) to 0.571 (*stalking*). However, the nature of the discrepancies varies. *Discrimination* exhibits the most consistent gap: all 18 pairs (100%) fell below the threshold, with a standard deviation of only 0.049. This indicates that the discrepancy between the general and legal understandings of discrimination is uniformly large, regardless of which pair of definitions is considered. In contrast, *abuse* shows the greatest variation (standard deviation = 0.197): some general definitions – such as "rude and offensive words" (average similarity 0.264) – are significantly distant from the legal meaning, while others – like "physical maltreatment" – are considerably closer. This heterogeneity may create a misleading impression for users that the dictionary fully covers the legal meaning of the term, when in practice this holds true only for certain definitions. Table 5 presents the most reductive general definition identified for each term, together with its average similarity score across all three legal definitions.

TABLE 5 Most reductive general definitions by term

Term	The Most Reductive Definition	Average Similarity
Harassment	"behaviour that annoys or upsets someone" (Cambridge Learner's)	0.459
Stalking	"the act of following a person or animal as closely as possible without being seen or heard" (Cambridge Learner's)	0.446
Discrimination	"the treating of some people better than others without any fair or proper reason" (Merriam-Webster)	0.478
Abuse	"rude and offensive words said to another person" (Cambridge Learner's)	0.264
Neglect	"Disregard of or indifference towards a thing. Usually with of." (OED)	0.431

To visually compare the distributions for all five terms, a box plot was constructed (Figure 3). A box plot is a way to display not just a single number but the entire distribution at once. The horizontal line inside the rectangle represents the median, which is the value exactly in the middle of all pairs. The rectangle itself covers the interquartile range, representing the middle 50% of all values. The whiskers indicate the range of the extreme values, and the individual

points beyond the whiskers are outliers – pairs that differ significantly from the majority. The wide box for *abuse* suggests that some of its general definitions differ substantially from the legal meaning, while others are unexpectedly close. Extreme outliers are particularly notable: for *neglect*, one pair reaches a value of 0.951 – the Merriam-Webster definition almost exactly matches the legal one; for *stalking*, the value is 0.792. However, these exceptions only emphasize the general pattern: they are isolated, while the vast majority of pairs remain well below the threshold.

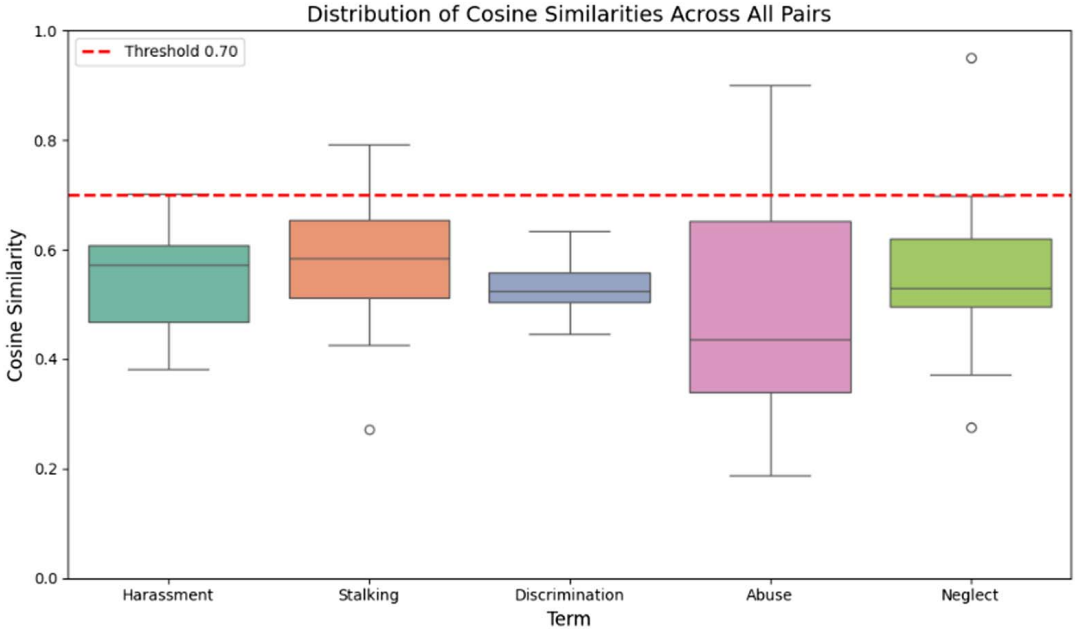


FIGURE 3 Distribution of cosine similarity for all term pairs (boxplot) [Source: generated by the author using Python]

To ensure that the results obtained were not due to chance, a one-sample t-test was conducted for each term. The one-sample test was selected because a directional hypothesis was being evaluated: that the mean similarity is significantly lower than 0.70, rather than simply different from it in either direction. The null hypothesis stated that the true mean similarity was 0.70. For all five terms, the p-values were below 0.01, ranging from 0.0094 for *stalking* to 8.12×10^{-11} for *discrimination*. This indicates that the likelihood of obtaining the observed results by chance is negligible (in statistical terms, a p-value below 0.05 means the result is unlikely to be random), confirming that the semantic reduction is statistically significant.

The qualitative analysis conducted using Voyant Tools aligns with these findings. To illustrate, consider two contrasting cases. The term *stalking* is notable because one of its general definitions, sourced from the OED, reaches a value of 0.792 – nearly exceeding the threshold. However, other definitions of the same term diverge significantly from legal interpretations. In the general corpus, *stalking* is described in terms of action and duration (e.g., *following, period, act*), whereas in legal contexts, it is characterized by the consequences for the victim (e.g., *fear, victim, safety, crime*). Conversely, *discrimination* presents a more homogeneous picture: all 18 pairs fall below the threshold, with minimal variation. In general dictionaries, discrimination is defined as unequal treatment (e.g., *treating, people, way*); in legal dictionaries, it is framed as a violation of rights established by law (e.g., *rights, unlawful, law*). These two cases represent the extremes of the observed spectrum and best illustrate that semantic reduction is not a uniform phenomenon; rather, it manifests differently depending on the nature of the term.

5 Discussion

The results provide quantitative confirmation of the lexical discrepancies identified in the qualitative analysis: general dictionaries tend to simplify socio-legal concepts. Across the five terms examined, not a single average cosine similarity value reached the 0.70 benchmark, and 92.4% of pairwise comparisons fell below it – a pattern that, while consistent within this corpus, should be read as indicative rather than conclusive given the limited number of terms (see Section 6). The term "semantic reduction" is used here descriptively, to denote the systematic omission of legally salient components from general definitions, without implying any evaluative judgement about the adequacy of general dictionaries for their intended communicative purposes.

The results confirm that the nature of the semantic gap varies by term (Section 4.4). This variability is of greatest analytical interest: *discrimination's* uniformly low similarity suggests that no general definition approaches the legal meaning, whereas *abuse's* high variance means some general definitions mislead users more than others – creating an inconsistent, and therefore harder-to-detect, gap. This heterogeneity is particularly problematic, as it can create a false impression among users that the dictionary covers the legal meaning of the term, although in practice this is true only for certain meanings. *Stalking* and *neglect* each have single upward outliers (0.792 and 0.951, respectively), where the general definition nearly coincides with the legal one. However, these exceptions only underscore the overall pattern: they are isolated cases, while most definitions still omit critical elements – such as victim's fear and criminal liability for *stalking*, and duty breach and harm for *neglect*.

The key issue is not that general and legal dictionaries differ – this is a natural consequence of their distinct communication goals and, in itself, is not problematic. The problem arises from communication failures: when a non-specialist user consults a general dictionary as a sufficient source of information about a concept that has a fundamentally different meaning in a legal context – and receives no indication of this difference. As Kaplan notes (2020, p. 65), general dictionaries inadequately warn users about potentially misleading usages. A user who looks up the term *harassment* to understand its meaning does not know beforehand that a legal dictionary is necessary; they only know they need to find the word's definition. Upon receiving a definition such as "behaviour that annoys or upsets someone," the user has no reason to assume it is incomplete, as the dictionary does not signal the existence of a more precise legal meaning. It is this lack of meta-information – not the discrepancy in definitions itself – that poses potential harm. A person who interprets *harassment* as merely "behaviour that annoys" may underestimate the legal implications of such behaviour, while the victim may underestimate the seriousness of their situation.

The study demonstrates that AI tools can play a practical role in addressing this problem. Sentence-BERT reliably measures the semantic gap in a reproducible and scalable manner, while Qwen2.5-3B-Instruct extracts editorial checklists that can serve as a starting point for lexicographers reviewing dictionary entries. However, AI remains an auxiliary tool rather than a substitute for professional judgment. As Lew (2024) notes, the role of the lexicographer in the digital age is shifting toward critical analysis, ethical evaluation, and the improvement of automated systems – it is ultimately the person who makes the final decision on how a definition should be formulated. The contributions of this work are summarised in the Conclusion.

6 Limitations and Directions for Future Research

The present study has several limitations that suggest directions for future research. The corpus includes only five terms and 105 pairwise comparisons; the findings should therefore be

read as a proof of concept demonstrating that the method can detect and quantify semantic reduction, rather than as evidence of its prevalence across socio-legal vocabulary in general. Expanding the corpus to 20–30 terms, including medical, financial, and environmental vocabulary, would enable verification of the universality of the observed pattern. The analysis is limited to English legal terminology. Cross-linguistic studies involving Russian, German, French, and Spanish could determine whether the phenomenon is specific to English or universal. The cross-sectional design is static, so diachronic changes were not examined; comparing different editions of the same dictionaries could reveal trends toward convergence of values over time. A user study involving law students, translators, and native speakers would help assess the practical consequences of semantic reduction. Finally, the threshold value of 0.70 is accepted as a provisional standard and can be refined in future studies – different types of terms may require different thresholds.

7 Conclusion

The present study demonstrates that the semantic gap between general and legal definitions of socio-legal terms is systemic, measurable, and statistically significant. Out of 105 pairwise comparisons across five terms, 92.4% fell below the cosine similarity threshold of 0.70, and the t-test confirmed significance for all terms ($p < 0.01$). The issue is not the discrepancy itself – which naturally arises from differing communication goals – but rather the lack of any indication to users that a more precise legal meaning exists.

This work makes three key contributions: it quantifies semantic reduction using a reproducible, NLP-based methodology; demonstrates that embeddings and generative models can be effectively combined in lexicographic analysis; and shows that AI can serve as a practical auxiliary tool for lexicographers, while final decisions remain with humans. Future research should expand the corpus of terms, conduct cross-linguistic and diachronic analyses, and develop a prototype tool for identifying and correcting semantic reduction in real time.

References

Alhudithi, E. (2021). Review of Voyant Tools: See through your text. *Language Learning & Technology*, 25(3), 43–50.

Bajčić, M. (2017). *New insights into the semantics of legal concepts and the legal dictionary*. John Benjamins Publishing Company.

Black, H. C. (2010). *Black's law dictionary* (Abridged 9th ed., prepared by B. A. Garner). West.

Cambridge Dictionary (2025–2026). Cambridge University Press. https://dictionary.cambridge.org

Durrheim, K., Schuld, M., Mafunda, M., & Mazibuko, S. (2022). Using word embeddings to investigate cultural biases. *British Journal of Social Psychology*, 62(1), 617–629. https://doi.org/10.1111/bjso.12560

Kaplan, D. (2020). *A theoretical model for the preparation of an inclusive and bias-free expression dictionary* [Doctoral dissertation, New York University]. https://hal.science/tel-02458629/file/KAPLAN_BIAS-FREE_2020%20_.pdf

Kopylova, M. (2026). *AI-assisted lexicographic bias detection for five socio-legal terms* [Google Colab notebooks]. Retrieved from https://drive.google.com/drive/folders/12Dctjw9xJakDRpQ9JxZzhncASOFNs8DR?usp=sharing

Law, J. (Ed.). (2022). *A dictionary of law* (10th ed.). Oxford University Press.

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, 11(1). https://doi.org/10.1057/s41599-024-02889-7

Merriam-Webster Dictionary (2025–2026). Merriam-Webster. https://www.merriam-webster.com/dictionary/

Merriam-Webster's dictionary of law (1996). Merriam-Webster.

Oxford English Dictionary (2003–2023). Oxford University Press. https://www.oed.com/dictionary/

Patel, A., & Joshi, H. (2025). DebugProGrade: Improving automated assessment of coding assignments with a focus on debugging. *Journal of Informatics and Web Engineering*, 4(1). https://journals.mmupress.com/index.php/jiwe/article/view/1266

Puttick, A. R., Ikae, C., Rigotti, C., Fosch-Villaronga, E., Kharas, M. W., Søraa, R. A., & Kurpicz-Briki, M. (2025). A systematic review of bias detection methods for non-English word embeddings and language models. *Artificial Intelligence Review*, 58(12), 370. https://doi.org/10.1007/s10462-025-11375-8

Reimers, N., & Gurevych, I. (2019). Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In S. Padó & R. Huang (Eds.), *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing* (EMNLP-IJCNLP) (pp. 3980–3990). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1410

Tiersma, P. M. (1999). *Legal language*. University of Chicago Press.

Tobia, K. P. (2020). Testing ordinary meaning. *Harvard Law Review*, 134(2), 726–806. https://harvardlawreview.org/print/vol-134/testing-ordinary-meaning/

Voyant Tools (2026). *Voyant Tools* [Web-based text analysis suite]. https://voyant-tools.org

Contact information

Mariia Kopylova
Roma Tre University
mashulya.kopylova@list.ru

Aleksandra Marković, Cvetana Krstev,
Milica Dinić Marinković & Ranka Stanković

TOWARDS A REPOSITORY OF SIMILES IN CONTEMPORARY SERBIAN

Abstract This paper presents the methodological framework for building the *Similes Repository for Contemporary Serbian* (Simili-SR), a corpus-based linguistic resource designed for the collection, annotation, and analysis of similes in contemporary Serbian. The study presents a multi-stage extraction and validation pipeline based on the genre-balanced corpus SrpKor2021+ of the *Reference Corpus of Contemporary Serbian*, combining CQL queries, local grammars implemented in UNITEX, and LLM-assisted semantic annotation. The initial extraction phase produced 574,195 candidate sentences across six major simile patterns, confirming both the productivity and structural diversity of simile constructions in Serbian. Subsequent rule-based filtering and LLM-assisted validation substantially reduced noise while preserving a large pool of linguistically relevant figurative comparisons. Manual evaluation performed on a stratified sample additionally revealed that simile interpretation frequently involves distributed semantic structure, and complex interactions between grammatical realization and figurative meaning, and highlighted the importance of linguistic post-processing, as well as several challenges related to semantic annotation in morphologically rich languages such as Serbian. By combining rule-based methods, LLMs, and detailed linguistic analysis, the paper contributes to both the development of computational resources for Serbian and broader research on figurative language annotation and representation.

Keywords Similes; repository; Serbian language; corpus linguistics; semantic annotation; lexical resources; UNITEX; LLM; NLP

> "Similes play a central role in enabling people to come to terms with the infinite variety of new phenomena encountered in daily life"
> (Hanks, 2005, p. 4)

1 Introduction

Similes represent one of the most productive (Hanks, 2005) and semantically diverse forms of figurative language, occupying an important place in linguistic description, lexicography, cognitive semantics, literary studies, and natural language processing. At the same time, their structural variability, possible occurrence in different word orders (which holds for multiword expressions in general), semantic and pragmatic gradience, and frequent occurrence in creative and context-dependent forms make them particularly challenging for automatic extraction, annotation, and semantic interpretation. These challenges become even more pronounced in morphologically rich languages such as Serbian, where figurative meaning is often distributed across larger syntactic structures and cannot always be reduced to fixed lexical patterns. The development of large-scale, linguistically informed simile resources is therefore important both for the study of figurative language and for the advancement of computational approaches to semantic annotation and interpretation.

In response to these challenges, the aim of this paper is to present the methodological framework for building the Similes Repository for Contemporary Serbian (Simili-SR), a corpus-based linguistic resource designed for the collection, annotation, exploration, and analysis of similes in contemporary Serbian. The repository is conceived as a dynamic and extensible resource intended for multiple types of linguistic, lexicographic, literary, and NLP-oriented research. In addition to serving as a searchable inventory of conventionalised and creative similes,

the repository is designed to support the study of figurative language variation, semantic annotation, lexical creativity, stylistic analysis, automatic figurative language processing, and the development of future computational models for Serbian. Instead of focusing exclusively on fixed or canonical similes, highly conventionalized (one of the criteria for inclusion in the simile category is conventionalization; Mršević, 1987, p. 93; Vulović, 2015, pp. 187–188; Marjanović, 2017, p. 62), Simili-SR aims to preserve the full variability of authentic simile usage, including elaborated, idiosyncratic, event-centred, and pragmatically complex figurative comparisons extracted from corpus data.

Within this broader framework, the present study focuses on constructing the repository using the genre-balanced corpus SrpKor2021+ (Vitas et al., 2025) and presents the complete extraction and validation pipeline that combines corpus queries, local grammars implemented in UNITEX, and LLM-assisted semantic annotation. In addition to evaluating the effectiveness of the extraction model, the paper places particular emphasis on manual linguistic evaluation and qualitative analysis of annotated simile structures. Special attention is devoted to the relationship between surface extraction patterns and linguistic simile structure, as well as to the challenges posed by semantically distributed and pragmatically complex figurative comparisons. By combining rule-based methods, large language models, and detailed linguistic analysis, the paper aims to contribute both to the development of computational resources for Serbian and to broader research on figurative language annotation and representation.

Since similes are multiword units, it is important to mention their main components, namely *tenor*, *property*, *vehicle*, and *marker*. The terminology differs in different linguistic traditions. Tenor is also known as the subject of comparison, topic, or target; property is also known as ground or, in Latin, *tertium comparationis*; vehicle is known as source or object of comparison.

1.1. Related work

Research on similes spans several complementary directions, including linguistic description, comparative research, multilingual resources, annotation methodology, computational processing, and corpus-based studies. From a theoretical perspective, an important line of research concerns the distinction between similes and metaphors. Hanks (2005) argues that similes foreground a salient property conventionally associated with another entity rather than establishing a holistic conceptual mapping, as in metaphor. Annotation-oriented research has further highlighted the semantic and syntactic complexity of similes. Mpouli (2017) proposed a framework for modeling similes as multi-component structures (tenor, ground, marker, vehicle), while emphasising annotation at phrase and word levels, reflecting the fact that similes often rely on larger syntactic units to create meaning. Related observations have been made by Parizoska & Filipović Petrović (2017), who demonstrate the variation in the adjective slot in similes with the pattern *adjective + kao ('as') + noun*, and argue that this variation reflects different ways of conceptualising the same semantic content as a property (adjectives), a manner (adverbs), or a process/event (verbs), highlighting the structural flexibility of similes.

In computational linguistics, research has focused both on distinguishing figurative from literal comparisons and on modeling the interpretation of figurative language. Niculae & Danescu-Niculescu-Mizil (2014) demonstrated that figurative and literal comparisons can be distinguished using semantic and contextual cues, while more recent work (Khoshtab et al., 2025) has investigated the ability of multilingual LLMs to interpret similes and other

figurative expressions across languages. Their findings indicate that performance varies significantly across figurative types, languages, and model architectures, with idiom interpretation generally achieving higher accuracy than simile interpretation. These studies highlight both the feasibility of automatic simile processing and the continuing challenges posed by semantic interpretation.

Similes have also attracted increasing attention in multilingual and translation-oriented research. Comparative studies have examined fixed similes across languages and their cultural variation (Mitrović et al., 2020), while parallel corpora have been used to investigate simile translation and cross-linguistic correspondences (Krstev et al., 2023). The analyses of the Serbian extension of the ELEXIS parallel corpus have shown that the distribution of similes varies considerably across genres, with literary texts exhibiting substantially higher frequencies than scientific and publicistic genres (Krstev et al., 2024).

For Serbian, Milošević & Nenadić (2016; 2018) developed a corpus of contemporary and traditional similes using web data and crowdsourcing. Their work highlighted a persistent challenge in simile extraction: distinguishing figurative similes from structurally similar literal comparisons, a task that required extensive manual curation. In Krstev (2021), the author compared the use of conventional and novel similes in SrpELTeC, a Serbian subcollection of ELTeC, with their use in contemporary Serbian novels. It is suggested that further research is needed – a systematic study of simile use in contemporary novels and in general corpora. The present study directly responds to this need by combining large-scale corpus extraction, semantic annotation, linguistic validation, and repository-based exploration within a unified framework.

There is a long line of linguistic research on similes in Serbian, and different authors use different terminology (a survey of this terminology can be found in Marjanović 2017, pp. 1–2). Due to space constraints in this paper, we focused only on those which were NLP-oriented.

2 Methodology

Figure 1 illustrates the complete pipeline for simile extraction and annotation. First, candidate sentences are retrieved from the corpus using CQL patterns. Next, UNITEX regular expressions and finite-state transducers supported by Serbian electronic dictionaries (Krstev, 2008) are used to filter and classify candidates into three match classes: reliable, possible, and rejected. The remaining candidates are then processed by GPT-5.4 LLMs, which should determine whether the sentence contains a simile and which annotates its main components, including tenor, property, vehicle, and marker. After manual validation, the confirmed similes are lemmatized, aligned with corpus sentences, and exported as XML resources together with user-friendly HTML representations.

2.1. Initial CQL-based candidate extraction

The initial candidate extraction was based on six CQL patterns designed to capture frequent Serbian simile constructions with comparative markers such as *kao*, *kô*, *k'o*,[1] *ka*, *kano*,[2] *nalik*, *poput*. The patterns targeted several syntactic configurations, including adjective phrase–comparison marker–noun phrase structures (*AkaoN*),[3] comparison marker-noun

[1] *Kao*, *kô*, *ka* and *k'o* are four variants of the marker *kao* ('as') in Serbian.

[2] This marker is obsolete in contemporary Serbian.

[3] We don't use usual syntactic notation in brackets; here N stands for NP, A for AP, Adv for AdvP, V for VP, and PrepN for PP.

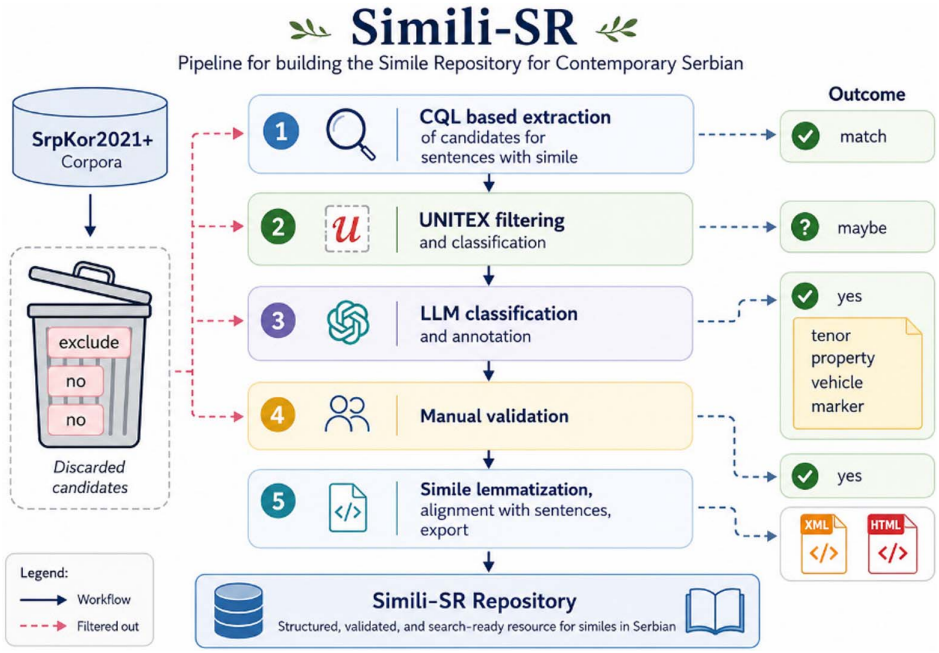


FIGURE 1 The methodology overview

phrase-adjective phrase structures (*kaoNA*), verb phrases followed by prepositional phrases (*VkaoPrepN*), verb phrases with noun vehicles (*VkaoN*), and adverbial phrases followed by a comparison marker and a prepositional phrase (*AdvkaoPrepN*). These queries were intentionally broad to maximize recall and retrieve a large pool of potential simile candidates from the corpus. Table 1 presents patterns and CQL queries used in the first phase for candidate extraction.

TABLE 1 CQL queries to retrieve sentences with simile occurrences

ID	Pattern	CQL query
1	AkaoN[a]	[tag="A" & lemma="^[a-zčćžšđ].*"] [tag="V" & lemma="jesam|biti"]? [word="kao|ko|k'o|ka'ka|ka|nalik|poput|kano"] [tag="A" & lemma="^[a-zčćžšđ].*"]? [tag_ud="NOUN"]
2	kaoNA	[word="kao|ko|k'o|ka'|ka|nalik|poput|kano"] [tag_ud="NOUN"] [tag="V" & lemma="jesam|biti"]? [tag="A" & lemma="^[a-zčćžšđ].*"]
3	VkaoPrepN	[tag="V"] [word="se" | tag="V" & lemma="jesam|biti"]? [word="kao|ko|k'o|ka'|ka|nalik|poput|kano"] [tag="PREP.*"][tag="A" & lemma="^[a-zčćžšđ].*"]? [tag_ud="NOUN"]
4	kaoPrepNV	[word="kao|ko|k'o|ka'|ka|nalik|poput|kano"] [tag="PREP.*"][tag="A" & lemma="^[a-zčćžšđ].*"]?[tag_ud="NOUN"] [word="se" | tag="V" & lemma="jesam|biti"]? [tag="V"]
5	VkaoN	[tag_ud="VERB"] [tag="V" & lemma="jesam|biti" | word="se"]? [tag_ud="NOUN"]? [word="kao|ko|k'o|ka'|nalik|poput|kano"] [tag="A" & lemma="^[a-zčćžšđ].*"]? [tag_ud="NOUN"]
6	AdvkaoPrepN	[tag_ud="ADV"] [tag="V" & lemma="jesam|biti" | word="se"]? [word="kao|ko|k'o|ka'|ka|nalik|poput|kano"] [tag_ud="ADP"] [tag="A" & lemma="^[a-zčćžšđ].*"]? [tag_ud="NOUN"]

[a]Here we offer an example for each pattern: AkaoN – *hladan kao led*; kaoNA – *kao sneg beo*; VkaoPrepN – *odjuriti kao bez duše*; kaoPrepNV – *kao bez duše stići*; VkaoN – *drhtati kao prut*; AdvkaoPrepN – *jasno kao na dlanu*.

2.2. Rule-based filtering and finite-state refinement

In the second step, the candidate list was further refined using finite-state transducers in UNITEX, together with rich Serbian electronic dictionaries (Krstev, 2008). This rule-based filtering enabled a more precise classification of extracted examples into three groups: reliable simile candidates ("match"), possible candidates ("maybe"), both requiring manual inspection, and non-candidates. In this way, the workflow combines broad CQL-based retrieval with linguistically informed FST-based refinement, reducing noise while preserving coverage for subsequent annotation.

Additional lexical and morphosyntactic constraints were introduced for adjectival patterns (*AkaoN* and *kaoNA*), including restriction to positive-degree adjectives, exclusion of relational, possessive, deverbal and proper-name-derived adjectives, ordinal numerals and filtering based on lowercase initial position. These refinements substantially reduced noise, although some valid similes may have been excluded, such as *zaštićen kao beli medved* (lit. 'protected like a polar bear'). The variant of the marker 'ka' was also excluded, since it often introduced prepositional phrases rather than comparative markers.

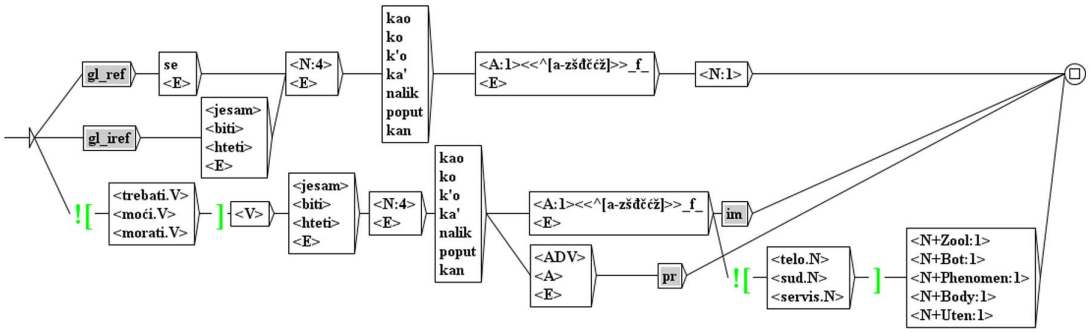


FIGURE 2 FST for simile pattern *VkaoN*: verbal constructions with noun vehicles

For the *VkaoN* pattern, regular expressions were not sufficient because the number of matches was too large. Therefore, a UNITEX graph was used instead (Figure 2). The graph combined verbs already attested in confirmed verbal similes with arbitrary nouns, while arbitrary verbs were combined only with nouns already found in identified similes. This set was further expanded to include selected semantic noun classes from the electronic dictionary (Krstev, 2008), including nouns denoting entities related to zoology, botany, natural phenomena, body parts, and utensils.[4] Because the filtering was stricter than in the adjectival patterns, rejected examples were labeled as 'probably not match' rather than simply 'not match'.

For the *kaoPrepNV* and *VkaoPrepN* patterns, no graphs of fully reliable cases have been constructed yet. Therefore, the current workflow separates the original corpus output into sentences marked for further processing, possible matches identified by a graph, and examples in which no candidate was found. Also, the marker "ko" was excluded because it was frequently confused with the pronoun *ko* ('who'). Preliminary inspection suggests that these patterns contain some useful examples, but also many multiword adverbial expressions that can combine with a wide range of verbs. For this reason, the interpretation of these candidates should also take into account previous outputs, especially whether a given *kaoPrepN* construction occurs with many different verbs. The *ADVkaoPrepN* pattern was processed

[4] Of course, there are also other semantic classes, but they were not included in this research, since it was corpus-based, and only the most frequent were taken into account.

in the same way as the verbal-prepositional patterns. However, initial inspection suggests that this group contains very few useful simile candidates.

2.3. LLM-assisted semantic annotation

The third part of the pipeline was the automatic annotation stage, where we designed a prompt-based LLM workflow specialised for simile identification and component extraction in Serbian. The system prompt defined the task as the recognition of figurative comparisons introduced by explicit comparison markers, like in previous phases. The prompt (see Appendix) explicitly distinguished figurative similes from literal comparisons or role expressions (e.g., *radi kao učitelj* – 'works as a teacher'), emphasising that the comparison must involve sharing of a salient property between entities.

For each sentence, the model was instructed to decide whether the sentence contains a simile and to return a structured JSON representation of the analysis. If a simile was detected, the model extracted its main semantic components: the *tenor* (subject of comparison), the *vehicle* (object of comparison), the *marker* (comparison conjunction), the *property* (lexical realisation in the text), and the *ground* (shared semantic property). While previous literature frequently uses the terms *ground*, *property*, and *tertium comparationis* interchangeably, our annotation process revealed the practical usefulness of distinguishing between the broader semantic basis of comparison and its explicit lexical realisation. Previous studies have already noted that the *tertium comparationis* may remain implicit, distributed, or inferentially reconstructed rather than overtly expressed (Parizoska & Filipović Petrović, 2017), while recent work on figurative language processing additionally suggests that figurative interpretation often extends beyond locally identifiable lexical correspondences toward more globally distributed interpretative structures (Mathisen et al., 2025). The distinction also aligns with previous NLP-oriented approaches (e.g. Qadir et al., 2016), as well as with recent computational models of figurative language (e.g. Chakrabarty et al., 2021; Niculae & Danescu-Niculescu-Mizil, 2014).

The model generated a canonical/lemmatised representation of the simile expression. Each prediction also included a confidence score and a short linguistic explanation in Serbian. The output format was strictly constrained to valid JSON in order to facilitate automatic post-processing and integration into downstream annotation pipelines. Sentences were processed in batches through dynamically generated user prompts containing numbered examples. This batching strategy reduced API overhead and enabled large-scale annotation of candidate similes extracted from the corpus. The structured prompting approach combines linguistic instructions with machine-readable output constraints, allowing efficient semi-automatic construction of a Serbian simile dataset suitable for further corpus analysis, evaluation, and model training.

In the preparation phase, both OpenAI GPT-5.4 and GPT-5.4-mini were evaluated on a manually inspected test set to assess their suitability for large-scale simile annotation. Although GPT-5.4 produced slightly more stable interpretations in ambiguous cases, GPT-5.4-mini achieved sufficiently reliable results and was therefore selected for large-scale annotation as a better compromise between annotation quality, processing speed, and cost.

2.4. Manual evaluation and annotation validation

The final stage of the pipeline involved manual validation and correction of a stratified sample of extracted candidates. To ensure reliable evaluation while maintaining a feasible annotation

workload, a sample of 900 sentences was selected across extraction patterns (it was an equal-sized random sample from each class) and automatically assigned confidence groups ("match" and "maybe"). The sampled instances were independently annotated by two linguistically trained annotators following detailed annotation guidelines. In addition to binary classification (simile vs. non-simile), annotators evaluated the correctness of extracted simile components, including tenor, vehicle, marker, property, and lemmatised form. For the evaluation, precision was estimated separately for each extraction pattern and globally across the dataset. Inter-annotator agreement was calculated using Cohen's κ coefficient, while system performance was evaluated against a gold standard using accuracy, precision, recall, and F1-score. Qualitative error analysis was additionally performed to identify recurrent sources of false positives, ambiguities, and annotation challenges.

This sampling-based framework enabled reliable assessment of the extraction methodology while keeping the manual annotation effort manageable. It also provided insights into the effectiveness of individual CQL patterns, UNITEX graphs, and LLM-based filtering strategies, thereby guiding future refinement of the Serbian simile extraction pipeline.

3 Results

3.1. Corpus extraction and candidate retrieval

The initial extraction phase relied on six CQL patterns to retrieve a total of 574.195 candidate sentences. The most productive pattern was *VkaoN*, which models verbal simile constructions, producing 396,319 concordances, 285,523 distinct surface forms, and 232,462 distinct lemmas. This confirms that verbal comparative constructions are highly frequent and lexically diverse in the corpus. The second largest group was *kaoNA*, representing comparison-initial structures, with 99,732 concordances and 57,327 distinct lemmas, followed by the adjectival pattern *AkaoN* with 46,849 concordances and 32,861 lemmas. Prepositional constructions generated smaller but still substantial candidate sets: *VkaoPrepN* produced 22,143 concordances and 14,547 distinct lemmas, while *kaoPrepNV* resulted in 4,938 concordances and 3,793 lemmas. The least productive pattern was *AdvkaoPrepN*, with 4,214 concordances and 3,452 lemmas. Overall, the large number of concordances and distinct lemmas demonstrates both the richness of comparative constructions in Serbian and the necessity of subsequent filtering and classification stages in order to distinguish genuine figurative similes from literal or non-figurative formulaic expressions.

3.2. UNITEX filtering and candidate classification

UNITEX filtering and classification stage, which significantly reduced the number of candidate sentences by separating them into reliable matches, excluded cases, and uncertain candidates requiring further inspection. The *AkaoN* pattern produced the highest number of reliable matches among adjectival constructions, with 2,551 confirmed candidates, alongside 15,654 possible matches and 28,644 excluded cases. In contrast, the *kaoNA* pattern generated only 428 reliable matches but a very large number of uncertain candidates (37,385), indicating that this structure is considerably noisier and more ambiguous. The verbal-prepositional patterns, *VkaoPrepN* and *kaoPrepNV*, did not yield directly confirmed matches at this stage, producing mostly excluded or uncertain examples, which suggests that these constructions are dominated by literal or formulaic expressions rather than figurative similes. The largest candidate pool was generated by *VkaoN*, yielding 1,302 reliable matches, 75,298 uncertain candidates, and more than 305,000 excluded cases, reflecting both the productivity and the high ambiguity of verbal comparison structures in Serbian. Finally, *AdvkaoPrepN* produced only uncertain

and excluded candidates, further confirming that adverbial-prepositional structures rarely correspond to genuine similes. Overall, the filtering stage dramatically reduced noise while preserving a substantial pool of potentially valuable simile candidates for subsequent LLM-based annotation and manual validation.

3.3. LLM-based validation and annotation

For the next stage, match and maybe subsets were processed. However, for three subsets with more than 15,000 samples, only a sample of 3,000 candidates was processed. The results revealed a strong agreement between the rule-based FST filtering and the LLM-based validation for candidates classified as matched (Figure 3). All three principal patterns (*AkaoN*, *kaoNA*, and *VkaoN*) achieved very high confirmation rates, ranging from 0.97 to 0.99, indicating that the LLM confirmed that the UNITEX filtering stage successfully identified highly reliable simile candidates. In total, 3,995 out of 4,055 matched candidates were confirmed as genuine similes by the LLM-based annotation pipeline. In contrast, the "maybe" category exhibited substantially lower confirmation rates and much larger variation across extraction patterns. The *kaoPrepNV* pattern achieved the highest validation rate among uncertain candidates (0.80), followed by *VkaoPrepN* and *AdvkaoPrepN* (0.64). Overall, only 44% of candidates from the "maybe" category were validated as genuine similes, confirming that this group contains substantially more noise and ambiguity than the matched category.

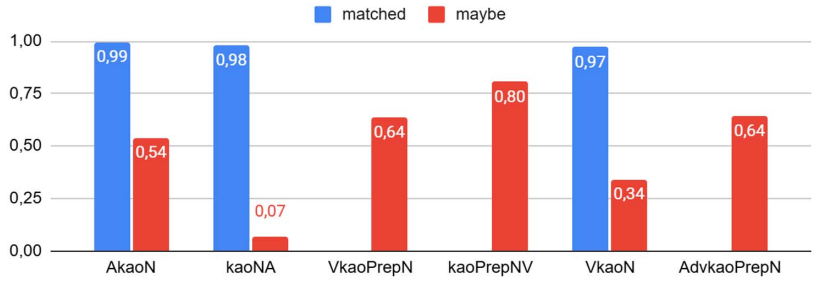


FIGURE 3 Result of LLM classification and annotation

3.3.1. Inter-annotator agreement

Inter-annotator agreement between the two human annotators was evaluated using Cohen's κ under flexible and strict evaluation settings. Human agreement was high in both cases, reaching $\kappa = 0.830$ (agreement 0.916) in the flexible setting, where only a simile was required, and $\kappa = 0.792$ (agreement 0.896) in the strict setting, which additionally required agreement on the annotation of simile components, indicating almost perfect and substantial agreement, respectively. Agreement between the annotators and the combined FST+LLM system was lower but remained substantial in the flexible setting ($\kappa = 0.669$ and 0.662; agreement > 0.83) and moderate in the strict setting ($\kappa = 0.532$ and 0.546; agreement ≈ 0.76). Overall, the results confirm the reliability of the annotation process while indicating that the automatic system captures most relevant simile annotations, although differences remain in stricter interpretation criteria.

3.3.2. Gold-standard evaluation

The evaluation results indicate that both approaches achieve very high precision, demonstrating that the identified similes are highly reliable and that false positives are relatively rare (Table 2). These findings indicate that the extraction and validation pipeline is particularly

effective at identifying reliable simile candidates, while differences between the two evaluation settings primarily reflect variation in annotation granularity and matching criteria rather than systematic classification errors.

TABLE 2 CQL queries to retrieve sentences with simile occurrences

Metric	Strict	Flex
Accuracy	0.7667	0.8445
Precision	0.9800	0.9790
Recall	0.6801	0.7944
F1-score	0.8029	0.8771

3.4. Manual evaluation and linguistic analysis of simile structures

Manual evaluation revealed that extraction patterns designed for broad corpus retrieval do not necessarily correspond to distinct linguistic simile types. Rather, several extraction patterns captured different surface realisations of the same underlying simile construction. This was particularly evident in the relationship between *AkaoN* and *kaoNA* patterns, which frequently produced alternative linear configurations of equivalent similes (examples 1a-b). These findings confirm that retrieval-oriented extraction patterns do not necessarily correspond to stable semantic simile categories.

(1a) da se žedan napije vode$_{\text{TENOR: ENTITY}}$ **bistre**$_{\text{PROPERTY: ADJ}}$ **kao**$_{\text{marker}}$ **kristal**$_{\text{VEHICLE: N}}$
lit. 'a thirsty man to drink water$_{\text{TENOR: ENTITY}}$ **clear**$_{\text{PROPERTY: ADJ}}$ **as**$_{\text{marker}}$ **crystal**'$_{\text{VEHICLE: N}}$

(1b) da se napijemo **kao**$_{\text{marker}}$ **kristal**$_{\text{VEHICLE: N}}$ **bistre**$_{\text{PROPERTY: ADJ}}$ vode$_{\text{TENOR: ENTITY}}$
lit. 'to drink **as**$_{\text{marker}}$**crystal**$_{\text{VEHICLE: N}}$ **clear**$_{\text{PROPERTY: ADJ}}$ **water**'$_{\text{TENOR: ENTITY}}$

The relatively high validation rate observed among uncertain kaoPrepNV candidates, mentioned in the subsection 2.2, warrants further linguistic interpretation. Manual inspection showed that a substantial proportion of these examples consist of lexicalised or semi-idiomatic comparative expressions, such as *kao bez duše* ('as if chased by the devil', 'breathlessly'), *kao iz vedra neba* ('like a bolt from the blue'), *kao od šale* (lit. 'like from a joke', effortlessly), and *kao na traci* ('as on an assembly line'). These comparative units occur productively across a wide range of verbal contexts and frequently combine with different verbs while preserving a relatively stable figurative meaning, e.g. *stići/trčati/dotrčati/razleteti se/poleteti kao bez duše* ('to arrive/run/rush/scatter/fly breathlessly'), or *menjati/isporučiti/menjati se/defilovati/ponavljati se kao na traci* ('to replace /deliver/change/parade/repeat oneself as on an assembly line'). From the perspective of surface extraction, such examples were retrieved through the kaoPrepNV pattern because the comparative phrase precedes the verb in the linear sequence. However, manual analysis showed that these constructions are more adequately interpreted as realisations of broader verbal simile structures, especially VkaoPrepN constructions, in which the verbal event constitutes the semantic centre of comparison, while the comparative phrase functions as a relatively stable idiomatized comparative unit.

AdvkaoPrepN candidates revealed an additional layer of structural-semantic mismatch, as most of these constructions function not as genuinely adverbial similes, but as parts of broader verbal or event-centred simile structures. Most frequently, the comparison modifies the event expressed by the verb as a whole and extends over the broader experiential or narrative configuration of the utterance. For example, in *vraćao se odatle kao iz duboke vode u koju je bio zagnjuren i u njoj zadržan do granice izdržljivosti* ('he returned from there as if emerging from deep water into which he had been submerged and held to the limits of endurance'), the comparison

unfolds across the entire experiential structure of the event, developing into an elaborated figurative scene that evokes psychological pressure, exhaustion, and release. The analysis further showed that genuinely adverbial comparisons occur primarily in cases where the comparative construction refers to a property, as in *Okidoh to sigurno kao iz puške* ('I answered that instantly, like a shot from a rifle'). In such examples, the comparative phrase functions more closely as an idiomatized adverbial modifier expressing manner. Overall, *AdvkaoPrepN* constructions appear to occupy a heterogeneous area between verbal similes, event-centred figurative comparisons, and lexicalised adverbial expressions, further confirming that extraction patterns based on surface structure do not necessarily correspond to stable semantic simile categories.

Another notable finding was the presence of numerous elaborated and highly creative similes that were not directly anticipated in the initial retrieval design. Due to space limitations, detailed analyses of these, together with idiosyncratic and ironic similes, are deferred to future work.

3.5. Repository interface and annotation visualisation

In Simili-SR[5], similes are organised by their lemmatised canonical forms[6] and displayed as collapsible sections that facilitate navigation through large collections of examples. Each section contains corpus attestations together with structured semantic annotations generated during the extraction and validation process. The annotation layer highlights the main simile components through colour-coded labels, including tenor, ground/property, marker, and vehicle, while corpus identifiers and sentence references support traceability and corpus alignment. The interface combines machine-readable annotations with a user-friendly HTML visualisation suitable for lexicographic work, linguistic research, and qualitative validation of extracted similes.

4 Discussion

As shown in Section 3.4, syntactic extraction categories and semantic simile categories do not always align straightforwardly. During annotation, boundaries between grammatical realisation and semantic interpretation frequently proved difficult to separate. This observation is consistent with previous cognitive-linguistic analyses showing that simile structures may alternate across adjectival, adverbial, and verbal realisations while preserving closely related conceptual content (Parizoska & Filipović Petrović, 2017). In our material, this became particularly evident in elaborated idiosyncratic similes, where semantic components are often distributed across larger syntactic structures rather than corresponding directly to individual lexical units, which also aligns with previous observations that simile components are frequently realised at the phrase level rather than through isolated lexical elements (Mpouli, 2018). These findings further suggest that similes operate simultaneously on multiple representational levels and that simile annotation cannot always rely on a simple one-to-one alignment between grammatical categories and semantic components. This opens important questions for future research on the interaction between syntactic structure, semantic organisation, and figurative language annotation. From the perspective of computational annotation, these observations indicate that simile interpretation frequently depends on distributed semantic inference rather than on locally identifiable lexical correspondences alone.

[5] https://llod.jerteh.rs/simili-sr/

[6] As the canonical form, we consider the form with the unmarked word order, unifying different possible variants in one entry. For example, *beo kao sneg* ('white as snow') and *kao sneg beo* (lit. 'as snow white') both represent one simile: *beo kao sneg*. Similes with lexical variations (e.g., *žut kao zlato* 'yellow as gold', *žut kao dukat* 'yellow as gold coin') are treated as separate entries.

4.1. Challenges of LLM-assisted annotation

The manual evaluation additionally revealed several tendencies related to tenor annotation that extend beyond the scope of the present paper and therefore could not be systematically analysed in full detail within the current study. Nevertheless, these observations proved sufficiently recurrent and linguistically relevant to indicate potentially important directions for future research, particularly regarding simile annotation in morphologically rich and pro-drop languages such as Serbian. Overall, the evaluation suggests that tenor identification was shaped by an unstable interaction of local linear proximity, verbal morphology, and partial reconstruction of predicational structure, frequently resulting in fluid or inconsistently delimited annotation spans.

One important difficulty concerns the identification of implicit tenors in Serbian as a pro-drop language. In examples such as *Tresem se kao prut* ('I am trembling like a twig'), the model was often capable of reconstructing the unexpressed participant from verbal agreement morphology. This suggests that the annotation process did not rely solely on surface lexical extraction, but also involved grammatical inference based on morphosyntactic information. At the same time, reconstructed tenors were frequently represented through inferred pronominal forms or morphologically inflected surface realisations instead of canonical lexical forms required by the annotation scheme. The evaluation additionally indicated that tenor identification was frequently influenced by local linear structure and broader predicational context. In examples such as *među sobom grizemo kao zubat bezubog* (lit. 'among ourselves we bite like the toothed bites the toothless') the model selected nearby leftward lexical material preceding the comparative construction, instead of the implied participant encoded through verbal morphology ('we'). Or, in some cases like "Mi imamo ovu vodu, možemo od toga svi da *živimo kao grofovi*" ('We have this water; we could all live like counts from it'), the model expanded the tenor span toward larger predicational segments "svi da živimo" ('for all of us to live') instead of isolating the participant itself. Nevertheless, the model occasionally identified the correct tenor even in non-canonical word orders. For example, the model was capable of "catching" the tenor even in word orders different from canonical, like in: *Crn, i lep i lak kao pero podizao se i spuštao* ***jedrenjak*** *u svetlucavoj vodi luke . . .* ('Black, and beautiful, and light as a feather, the **sailing ship** rose and fell in the shimmering water of the harbor').

Despite formal inconsistencies in case marking, normalisation, or span boundaries, the model was often semantically accurate. Examples such as *leže kraj Džoa koji je spavao kao zaklan* ('he lay beside Joe, who slept like the dead') demonstrate that the model frequently captured the correct semantic referent while preserving morphologically inflected (the genitive form of proper noun, *Džoa*) or contextually embedded surface realisations instead of normalised canonical forms (e.g., Džo).

Property identification proved more challenging than tenor annotation. In adjectival similes, explicitly expressed properties were generally recognized correctly (we miss the statistics for this at the moment), whereas coordinated properties were often only partially identified. The model struggled with cases with two coordinated properties (*. . . grof slučajno dodirnu tananu maramicu na kojoj su kapi krvi bile* ***vlažne i crvene*** *kao karanfil na snegu* ' . . . the Count accidentally touched the delicate handkerchief, where the drops of blood were **moist and red** as carnations on snow'); in such cases, the model only recognized one adjective, *red*, as a property. Greater difficulties emerged in verbal similes, particularly when the property was implicit and contained in the verb, like in: *pilići* ***se žute*** *kao smilje i kao suvo zlato*, lit. 'the chicks are yellowing like immortelle and pure gold', 'the chicks are as yellow as immortelle and as pure gold', where the word *žute*, as subject complement, was recognized as a property. The model also occasionally confused vehicle modifiers with grounds, suggesting uncertainty regarding annotation

boundaries. In example . . . *predstave su se ocrtavale pred mojim očima kao žive slike*. ('The scenes took shape before my eyes like living pictures.') the model incorrectly chose the adjective *žive*, 'living', as the ground, although it is a modifier of the noun *picture* and thus part of the vehicle. Additional challenges arose in constructions containing multiple similes, where only one comparison was usually recognized.

Taken together, these observations suggest that simile annotation in morphologically rich languages cannot be reduced to simple lexical extraction alone, but instead involves interactions between grammatical inference, referential interpretation, local linear structure, semantic scope, and decisions regarding annotation boundaries. A more systematic qualitative and quantitative investigation of these phenomena, including the annotation of implicit and distributed simile components, verbal event structure, semantic inference, and annotation granularity, remains an important task for future work.

5 Conclusions

This paper presented the methodological framework for building the *Similes Repository for Contemporary Serbian* (SSR), a corpus-based resource designed for the collection, annotation, and exploration of similes. To our knowledge, existing work has largely focused on simile collections, dictionaries, annotated datasets, or task-specific NLP resources, whereas SSR brings together corpus evidence, semantic annotation, linguistic validation, and repository-based exploration within a unified framework.

The results demonstrate that a combined pipeline based on corpus queries, FST, and LLM-assisted annotation can successfully support the large-scale construction of a simile repository while maintaining high precision and annotation reliability. At the same time, the manual evaluation revealed several linguistically relevant phenomena, including the mismatch between surface extraction patterns and semantic simile categories, the importance of distributed and event-centred comparison, and challenges related to tenor and ground identification and pragmatic interpretation. These findings highlight the importance of integrating linguistic expertise into the development and evaluation of figurative language resources. These findings also confirm broader conclusions that AI-assisted research demands linguistic evaluation. Future work will focus on extending the repository with additional corpora, refining the annotation model, further investigating phenomena such as tenor and property identification, their span, irony, idiosyncratic similes, and the interaction between grammatical structure and figurative meaning.[7]

This research has additional value for lexicography, knowing that traditional dictionaries include similes rather unsystematically (both in the sense of their inclusion and inclusion under one of their components). Since it is corpus-based, the frequency of each simile can be traced and used as a criterion for inclusion/exclusion. The repository might serve as a basis for other dictionaries, monolingual, bilingual, phraseological. Since the repertoire of similes used in a language changes over time, the presence of some similes and the absence of others may indicate cultural and historical changes, and serve as a basis for different ethnolinguistic research.

References

Bergenholtz, H. & Johnsen, M. (2013). User Research in the Field of Electronic Dictionaries: Methods, First Results, Proposals. In R. H. Gouws, U. Heid, W. Schweickard & H. E. Wiegand (Eds.), *Dictionaries.*

[7] The paper is to be published soon: Dinić Marinković et al., 2026, forthcoming.

An International Encyclopedia of Lexicography: Supplementary Volume: Recent Developments with Focus on Electronic and Computational Lexicography (pp. 556–568). Walter de Gruyter.

Chakrabarty, T., Zhang, X., Muresan, S. & Peng, N. (2021). MERMAID: Metaphor generation with symbolism and discriminative decoding. In *Proceedings of the 2021 conference of the North American chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 4250-4261). https://doi.org/10.48550/arXiv.2103.06779

Dinić Marinković, M., Krstev, C., Marković, A. & Stanković, R. (2026). *Repozitorijum simila savremenog srpskog jezika – izgradnja i evaluacija mašinski potpomognute anotacije*. In print.

Hanks, P. (2005). Similes and Sets. In R. Blatná & V. Petkevič (Eds.), *Jazyky a jazykověda (Languages and Linguistics: Festschrift for Professor Fr. Čermák)* (pp. 1–15). Philosophy Faculty of Charles University.

Khoshtab, P., Namazifard, D., Masoudi, M., Akhgary, A., Mahdizadeh Sani, S. & Yaghoobzadeh, Y. (2025). Comparative study of multilingual idioms and similes in large language models. In O. Rambow, L. Wanner, M. Apidianaki, H. Al-Khalifa, B. Di Eugenio & S. Schockaert (Eds.), *Proceedings of the 31st International Conference on Computational Linguistics* (pp. 8680-8698).

Krstev, C. (2008). *Processing of Serbian – Automata, Texts and Electronic Dictionaries*. Faculty of Philology, University of Belgrade.

Krstev, C., Jaćimović, J. & Vitas, D. (2020). *Analysis of Similes in Serbian Literary Texts (1860-1920) using computational methods*. CLIB.

Krstev, C. (2021). White as Snow, Black as Night – Similes in Old Serbian Literary Texts. *Infotheca*, 21(2), 119–36. https://doi.org/10.18485/infotheca.2021.21.2.6

Krstev, C., Stanković, R. & Marković, A. (2023). Multiword Expressions - Comparative Analysis Based on Aligned Corpora. In *Book of Abstracts of the UniDive 1st general meeting*. Paris-Saclay University.

Krstev, C., Ranković, S., Marković, A. & Mihajlov, T. (2024). Towards the semantic annotation of SR-ELEXIS corpus: Insights into Multiword Expressions and Named Entities. In A. Bhatia, G. Bouma, A. S. Dogruöz, K. Evang, M. García, V. Giouli, L. Han, J. Nivre & A. Rademaker (Eds.), *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies* (MWE-UD) @ LREC-COLING 2024 (pp. 106–114).

Marjanović, S. (2017). *Poredbene frazeme s komponentom* comme/kao *u francuskom i srpskom jeziku*. Filološki fakultet Univerziteta u Beogradu, rukopis neobjavljene doktorske disertacije.

Mathisen, E. K., Allott, N. & Ronderos, C. R. (2025). Cognitive mechanisms in simile and metaphor comprehension. *Language and Cognition*, 17, e89. https://doi.org/10.1017/langcog.2025.10038

Milošević, N. & Nenadić, G. (2016). As Cool as a Cucumber: Towards a Corpus of Contemporary Similes in Serbian. *ArXiv*. https://arxiv.org/abs/1605.06319

Milošević, N. & Nenadić, G. (2018). Creating a contemporary corpus of similes in Serbian by using natural language processing. *ArXiv*. https://arxiv.org/abs/1811.10422

Mitrović, J., Markantonatou, S. & Krstev, C. (2020). A cross-linguistic study on Greek and Serbian fixed similes and enrichment of lexical resources via crowdsourcing. In S. Markantonatou & A. Christofidou (Eds.), *Multiword Expressions: Drawing on Data from Modern Greek and Other Languages* (pp. 241-262). Research Center for Scientific Terms and Neologisms.

Mpouli, S. (2017). Annotating similes in literary texts. In H. Bunt (Ed.), *Proceedings of the 13th Joint ISO-ACL Workshop on Interoperable Semantic Annotation (ISA-13)* (pp. 26-36).

Mršević, D. (1987). *Frazeološke glagolsko-imeničke sintagme u savremenom srpskohrvatskom jeziku*. Filološki fakultet Beogradskog univerziteta. Frazeološke glagolsko-imeničke sintagme u savremenom srpskohrvatskom jeziku. Knj. XI.

Niculae, V. & Danescu-Niculescu-Mizil, C. (2014). Brighter than Gold: Figurative Language in User-Generated Comparisons. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 2008–2018). Association for Computational Linguistics.

Parizoska, J. & Filipović Petrović, I. (2017). Variation of Adjectival Slots in kao ('as') Similes in Croatian: A Cognitive Linguistic Account. In R. Mitkov (Ed.), *International Conference on Computational and Corpus-Based Phraseology* (pp. 348-362). Springer International Publishing.

Qadir, A., Riloff, E. & Walker, M. A. (2016). Automatically Inferring Implicit Properties in Similes. In K. Knight, A. Nenkova & O. Rambow (Eds.), *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (pp. 1223–1232). Association for Computational Linguistics.

Vitas, D., Stanković, R. & Krstev, C. (2025). O familiji korpusa savremenog srpskog jezika SrpKor. In J. Moskovljević Popović & R. Stanković (Eds.) *Proceedings of the International Conference South Slavic Languages in the Digital Environment JuDig: Thematic Collection of Papers*, Vol. 1 (pp. 3–27).

Vulović, N. (2015). *Srpska frazeologija i religija. Lingvokulturološka istraživanja*. Monografije 23, Institut za srpski jezik SANU.

Contact information

Aleksandra Marković
Institute for Serbian Language, Serbian Academy of Sciences and Arts
aleksandra.markovic@isj.sanu.ac.rs

Cvetana Krstev
JeRTeh Society
CvetanaJK@gmail.com

Milica Dinić Marinković
University of Belgrade, Faculty of Philology
milica.dinic.marinkovic@fil.bg.ac.rs

Ranka Stanković
University of Belgrade, Faculty of Geology and Mining
ranka.stankovic@rgf.bg.ac.rs

Appendix

System prompt:

```
You are a linguist specializing in figurative language.

Task:
Determine whether the given sentence contains a SIMILE (figurative
comparison).

Definition:
A simile is a figurative comparison between two entities using an
explicit comparison marker
such as "kao", "poput", "k'o", "nalik", "poput", etc., where one
entity is described by comparing it to another (e.g., "bled kao krpa",
"kao krpa bled, "sija kao zlato").

Important:
   - The comparison must be figurative, not literal.
   - Expressions like "radi kao u□itelj" (works as a teacher) are NOT
     similes.
   - Focus on whether a property is transferred (figurative meaning).
```

```
Instructions:
1. Decide if the sentence contains a simile (YES or NO).
2. If YES, extract as found in text (in Serbian):
   - tenor (what is being described)
   - ground (shared property in English)
   - property (lexical property - word in text)
   - marker (comparison word)
   - vehicle (what it is compared to)
   - lemmatized (canonical) form of simile
3. Provide a short explanation.

You will receive a numbered list of sentences.

Return ONLY a valid JSON array — one object per sentence, in the
same order.
No markdown, no extra text.

Each object must follow this schema exactly:
{
   "id": <integer, matches input number>,
   "sentence": "<original, unchanged sentence>",
   "is_simile": "YES" or "NO",
   "tenor": "" or extracted value,
   "ground": "" or extracted value,
   "property": "" or extracted value,
   "marker": "" or extracted value,
   "vehicle": "" or extracted value,
   "simile": "" or extracted value,
   "confidence": <float 0.0-1.0>,
   "explanation": "<one sentence in Serbian>"
}
```

Sanni Nimb, Ida Flörke, Agnes Aggergaard Mikkelsen,
Nathalie Hau Norman & Nicolai Hartvig Sørensen

AI USED TO DETECT PROBLEMATIC CONTENT IN CORPUS EXAMPLES

Abstract Based on OpenAI's moderation model, we describe a method developed to automatically identify unwanted discriminatory content in the citations found in the printed edition of *Den Danske Ordbog* (DDO, 2003–2005). Today, the DDO is published online and is continuously revised and expanded with new headwords. We begin by testing whether the model can identify 234 citations that have been removed since 2015 because the editorial team deemed them discriminatory. The results are promising, so we apply the model to 71,000 citations from headwords that have not yet been revised. The model classifies 6.3% of these as harmful. In addition to identifying the expected types of citations, a large proportion are flagged because they involve violence or self-harm—a type of problematic content that had previously been overlooked. Of the 4,800 flagged instances, 37% have been manually validated. Regarding violent content, the precision is only 10%—which nevertheless yields a list of approximately 350 problematic cases. The editorial team intends to discuss this type of content and establish new guidelines. For discriminatory content, the precision is higher—33%—which likewise results in a total of approximately 350 citations that are candidates for removal from the DDO and replacement with others.

Keywords Moderation model; harmful content; corpus citations; Danish; DDO

1 Introduction

In recent years, there has been growing interest within the lexicographical community in reviewing and revising the content of older dictionary entries published in todays' online versions, driven by society's increased focus on gender, minorities, and equality. There is a desire to avoid content that is conceived as being harmful, offensive, or stereotypical, and at the same time, this effort must be balanced with the aim of practicing descriptive lexicography. Among the dictionaries that have been revised to address problematic content are the Oxford Dictionaries, Dictionary.com (The Guardian, 2020 a and b), the German Duden (MailOnline, 2021), the Norwegian dictionaries NOB and BOB (Lyse & Svardal, 2020), and the Swedish dictionary SO (Petersson & Sköldberg, 2020).

This article focuses on the task of identifying problematic content of this type within the citations of the corpus based *Den Danske Ordbog* (DDO). The dictionary was originally published in print between 2003 and 2005 but has been available online since 2009. Since the first digital edition was released, it has been updated approximately twice a year with new headwords, fixed expressions, and new meanings based on a corpus that is supplemented with new texts every month and currently contains more than 1.4 billion tokens. The editorial team comprises ten lexicographers and computer scientists.

DDO currently contains 134,861 definitions illustrated by 171,803 corpus citations. More than 71,000 of these come from texts written between 1983 and 1992 and appear as the sole example of a meaning defined for a headword in the 1990s that has not been revised since. In recent years, many inquiries from users to the editorial team have concerned discriminatory corpus examples in the DDO, and both Fjeld (2015) and Lazić (2024) point out that the dictionary contains gender- and age-stereotypical examples. The problematic citations are very difficult to locate systematically; they are hidden among far more unproblematic citations in the dictionary.

Unlike the printed edition, which typically provided only a single corpus citation, the digital edition displays two visible citations for every headword and sense added or revised since 2010. In many of these newly edited entries, users can also choose to view additional corpus citations. The corpus examples in DDO are never altered; they appear in their original form and serve to illustrate the headword in authentic use. A good corpus example complements the definition without drawing attention to something entirely different—which can happen, for instance, if the example is biased regarding gender, minorities, and equality.

However, when the DDO was originally compiled in the 1990s, the focus was not on whether the corpus examples contained stereotypes, but rather on ensuring they reflected authentic language use—and, by extension, authentic attitudes. Electronically edited, corpus-based dictionaries were a relatively new phenomenon. The numerous examples were selected from a 40-million-word (token) corpus comprising texts from the period 1983–1992; a large portion of the material came from newspapers (including tabloid newspapers), a genre that often contains stereotypes according to Müller-Spitzer & Rüdiger (2022). Since 2015, the editorial team has been involved in revising an ever-increasing number of corpus examples. However, there are still many examples awaiting reviews and updating.

1.1 Revision of examples in DDO

Reviewing corpus examples is a high-priority task, as we aim to keep the dictionary free of offensive and outdated content (Jensen et al., 2018, p. 142). Identifying the cases that require updating is a time-consuming task. One approach has been to review specific text sources likely to contain problematic content, such as the women's magazines, erotic magazines, and tabloids. Another has been to search for specific text strings within the corpus examples—such as *kvinder* ('women')—to identify gender stereotypes. User feedback has also proven valuable. Problematic citations have been removed from the directly visible examples in DDO, but in some cases, they have been retained among the additional citations that can be accessed with a click.

To accelerate the editorial process, ensure that all examples have been systematically checked, and avoid this type of user feedback in the future, this article explores the possibility of using AI-based tools to identify any remaining problematic content in the old examples.

2 OpenAI's moderation model

We use the "omni-moderation-2024-09-26" moderation model developed by OpenAI which is designed to identify potentially harmful content in text and images, for example on social media. It is designed to identify not only the types of harmful discriminating content that have until now been focused upon in the DDO revisions—such as biased attitudes towards gender, minorities, and equality—but also to identify content regarding self-harm, violence and illegal actions. The model used in our experiments accepts text, i.e. in our case citations from the DDO, as input and provides a moderation assessment consisting of a binary flag (TRUE/FALSE) indicating whether the input is classified as harmful, as well as category-specific scores for a range of predefined harm categories established by OpenAI. These categories encompass not only the types of harmful or discriminatory content previously focused on by the DDO editors—such as biases regarding gender, minorities, and equality—but also the identification of content related to self-harm, violence, and illegal acts.

This type of moderation model differs from traditional LLMs (such as GPT models, Claude, and Llama) in that it is not trained to engage in conversation or follow instructions; instead, it is

specifically trained as a content moderation classifier. Consequently, it is more difficult to simply swap the model for another available one for comparison purposes, as the expected inputs and outputs, category definitions, and scoring methods are typically determined by the individual developers behind each moderation system.

The model calculates automatically for a given text input the degree of 'harassment', 'hate', 'self_harm', 'sexual', 'illicit' and 'violence' content, as well as more specific subtypes (harassment_threatening, hate_threatening, illicit_violent, self_harm_instructions, self_harm_intent, sexual_minors, violence_graphic). If the sum of the calculated degrees for each type exceeds a certain, predetermined value (0.5), the text is assigned the value 'TRUE' by the model, i.e. judged as problematic. If it is below 0.5, it is instead assigned the value 'FALSE' (i.e. judged as non-problematic). Table 1 shows an example.

TABLE 1 OpenAI's Moderation Model rates the corpus example to the left illustrating the verb *ævle* ('to talk nonsense') in DDO. The sum is above 0.5 ('TRUE')

Example	Harassment?	Harassment-threatening	Hate	Hate-threatening	Illicit
du godeste, hvor det kvindemenneske kunne ævle ('Oh my god, how could that woman talk such nonsense')	0.5267	0.003	0.1819	0.0034	0.0038

3 Test of OpenAI's moderation model

Initially, we tested the moderation model to determine whether it would be useful at all for identifying the type of content in the older citations that we wish to remove from the dictionary. For the test, we draw on the results of the manual DDO revisions conducted since 2015 (Jensen et al., 2018, p. 142). We created two datasets.

3.1 Datasets

Dataset 1 consists of 234 manually identified corpus examples agreed upon by two editors to be either harmful, offensive and/or stereotypical to such an extent that they had to be removed from the dictionary. We divide the 234 citations into categories depending on the type of discriminative content:

1. Ethnicity:
   adj. *plat* ('flat'): *De indfødte er et besynderligt skue. De har platte næser og store øreflige som er gennemhullede* ('The natives are a strange sight. They have flat noses and large earlobes that are pierced').

2. Antifeminism (i.e., misogyny and denial of discrimination):
   adj. *tantet* ('aunty'): *de gamle kvindesagskvinder – var de ikke lidt tantede, måske endda lidt pebermøagtige*? ('the old women's rights activists - weren't they a little aunty, maybe even a little sissy-like?').

3. Gender stereotypes:
   adj. *modsigelsesfyldt* ('fickleness'*): Med kvindeligt vægelsind bevæger hun sig ud i den ene modsigelsesfyldte argumentation efter den anden* ('With feminine fickleness, she moves into one contradictory argument after another')

noun *velbefindende* ('well-being'): *For en mands velbefindende er beundrende kvindeblikke vigtigere end kalorier og medikamenter* ('For a man's well-being, admiring female glances are more important than calories and medication').

4. Woman as sexual object:
   adj. *herlig* ('lovely'): *Mette er en herlig pige på 24 år, ufattelig køn og velskabt og ofte fuld af frække indfald* ('Mette is a lovely girl of 24 years old, incredibly pretty and well-shaped and often full of naughty ideas')
   adj. *gennemsigtig* ('transparent'): *han rækker ud efter sygeplejerskens ben, som han skimter under den gennemsigtige kittel* ('he reaches out for the nurse's legs, which he glimpses under the transparent coat').

5. Homosexuality:
   noun *mudder* ('mud', figurative use): *Det dér kønsløse mudder jeg som lesbisk feminist havde trampet rundt i i 70'erne, havde gjort begæret noget kedsommeligt* ('The genderless mud I had trampled through as a lesbian feminist in the 70's had made desire somewhat boring').

6. Physical disabilities/appearances:
   adj. *abnorm* ('abnormal'): *Den grimme søster med hareskåret var ikke bare åndssvag, hun var også i videste forstand anatomisk abnorm* ('The ugly sister with harelip was not only mentally retarded, she was also anatomically abnormal in the broadest sense')
   adj. *fed* ('fat'): *hun gemte sin fede krop i en badedragt, hvor dellerne pressede sig som dejpølser ud over dragtens kanter* ('she hid her fat body in a swimsuit, the rolls of fat pressing out like dough sausages over the edges of the suit').

Dataset 2 used in the initial test consists of 198 citation examples approved by two DDO editors to replace the problematic citations that were removed. We initially regard these examples as unproblematic—that is, neither harmful nor offensive, nor characterized by stereotypes.

3.2 Results

We are primarily interested in how many of the 'TRUE' values are correctly identified as problematic—in accordance with the two editors' assessment—when the examples from the two datasets are used as input. The results of the test show that:

- The model identifies 50% of the categorized examples as 'TRUE' in dataset 1 (the problematic and therefore removed examples) based on content typically identified as harassment, hate, and sexual, or a combination.
- It performs best in the cases of '5. Homosexuality' (100% accuracy), '1. Ethnicity ´ (75% accuracy), and '2. Anti-feminism' (also 75% accuracy).
- It performs worse in the cases of '3. Gender stereotypes' (17% accuracy), 6. Physical disabilities/appearances' (28% accuracy), and '4. Woman as sex objects' (31% accuracy).

In dataset 2 (the dataset consisting of unproblematic citations that were inserted by DDO editors to replace the problematic ones), 18% (35) are identified as 'TRUE' by the model, and not zero as expected. Among them are also citations illustrating offensive or stereotypical headwords, i.e. citations that are inevitably harmful and problematic, even though they are less harmful than the ones they replace. If the citations had instead been included in dataset 1 from the beginning, OpenAI's moderation model would be able to identify 56% of the harmful, offensive or stereotypical citations in the dataset.

The conclusion of our initial test of OpenAI's Moderation Model is therefore that if we implement it on all citations in the non-revised old DDO entries, more than half of the examples marked as 'TRUE' in the output are most likely to be removed from the dictionary, making it worth the effort to carry out the implementation. However, we are also aware that problematic examples reflecting gender stereotypes, physical disabilities/appearances and 'woman as sex object' are rarely identified by the model, and that we will have to find these by other means.

4 Implementing OpenAI's moderation model

In implementing the model, we use 71,095 citation examples from the first printed edition as input. They all appear under headwords that have not yet been reviewed and revised. Among the 71,095 citations, 6.3% (4,800 of which 400 are used in more than one DDO entry) are identified as 'TRUE' by the model. To our surprise, almost four out of five citations marked as 'TRUE' by the model, are true not due to stereotypical or offensive content as expected, but due to 'violent' or 'self harm' content (78% of the 4,800 identified 'TRUE' cases).

A bit more than a third of the 4,800 citations (37%) were validated, distributed among three DDO editors. The validation shows that only 25% of the identified 'TRUE' examples are likely to be removed from the dictionary according to the (today updated) editorial guidelines. In this number, we include also more unclear cases and cases where the citation is outdated for other reasons (2,7%)).

4.1 Violent content

The manual validation is a bit time-consuming since 90% of the 'violent' cases identified by OpenAI's Moderation Model are wrongly interpreted, typically due to figurative use of fixed expressions and highly polysemous verbs of which the concrete sense is violent:

> Idiom: *et skud fra hoften* (litt. 'a shot from the hip', 'a quick guess'): *jeg kan ikke huske det, jeg forsøger et skud fra hoften, hans violinkoncert* (litt.'I can't remember it, I'm trying a shot from the hip, his violin concerto'/'I can't remember it, I'll try a quick guess, his violin concerto')

> Polysemous verb: *slå* ('to hit; to beat'): *Han så sig tilbage over skulderen. En mand kom gående med kraven slået op om ørerne* (' litt. 'He looked back over his shoulder. A man was walking by with his collar hit up over his ears'/'He looked back over his shoulder. A man was walking by with his collar turned up over his ears').

The model also wrongly identifies violent content in cases of meat preparation.

> *Vildænder skal plukkes, ikke flås* ('Wild ducks should be plucked, not skinned').

This means that only 10% of the 'violent'/'self_harm' cases are in fact candidates to be removed from the DDO. However, we find it highly interesting that they are pointed out by the model. We were not aware of this type of problematic content in the old entries.

> Verb *tidsindstille* ('to time'): *Han tidsindstiller sit hævn-selvmordsforsøg til konens sædvanligt præcise hjemkomsttidspunkt* (He times his revenge-suicide attempt to coincide with his wife's habitually punctual arrival home)

> Noun *møgunge* ('little brat'): *Forbandede møgunger. Hvis I ikke er stille, skal jeg slå jeres hoveder ind i væggen* ('Damn little brats. If you don't shut up, I'll smash your heads against the wall')

adj. *maniodepressiv* ('manic-depressive'*): Lisa spørger, om hun er maniodepressiv, fordi hun tænker på selvmord* ('Lisa asks if she is manic-depressive because she is thinking about suicide')

The editorial team will discuss the extent to which this type of content should be removed from the dictionary. We expect to end up with 300–350 problematic citations expressing violence and self-harm.

4.2 Offensive and stereotypical content

In the remaining fifth of the cases that the model identifies as 'TRUE' (22% of the 4,800 cases, i. e. 1,056 citations), it does so due to content estimated to be 'harassment', 'hate', 'sexual', or a combination. According to the validators, every third of these 'TRUE' cases resemble the type of citations that were in the first place manually removed from the dictionary over the last years and replaced by other citations. Some examples are:

verb *abstrahere* ('to abstract'): *Mænd kan måske ikke abstrahere fra kvinder som kønsobjekt* ('Men may not be able to abstract from women as sex objects')

adv. *også ('*also*'): hvorfor skal I kvinder også altid være hysteriske og på tværs* ('Why do you women always have to be hysterical and confrontational?)

noun *mandfolk* ('men'): *for hende er alle mandfolk de samme bæster, den ene ikke bedre end den anden* ('For her, all men are the same beasts, one no better than the other)

idiom *gå i arv* ('to be inherited'): *Homoseksualitet kan gå i arv* ('Homosexuality can be inherited')

adj. *vommet* ('pot-bellied'): *du ligner præcis det du er. En overvægtig, vommet, hjerneskadet malermester* ('You look exactly like what you are. An overweight, pot-bellied, brain-damaged master painter')

noun *ulykkesfugl* ('bird of misfortune'): *Jeg er en ulykkesfugl, en nigger der driver folk ud i selvmord og religiøs fanatisme* ('I'm a bird of misfortune, a nigger who drives people to suicide and religious fanaticism')

Even though this is a lower percentage than expected from the test results, we estimate to find 350 problematic citations of this type among the identified 'TRUE' citations. These will be handed over to the editors of the DDO dictionary. Subsequently, two DDO editors will be involved in the final review of the citations as part of the DDO editing process.

Since we have not categorized them according to the six groups used in the initial test, we do not know exactly which type of problematic content is most prevalent. A large number appear to reflect gender stereotypes, even though—based on the results of the initial test—we had expected the model to struggle with identifying such instances. This likely implies that a significant number of stereotypical citations remain hidden within the data that was not identified as 'TRUE' by the moderation model. It may also be useful to examine the borderline cases and lower the threshold value slightly.

5 Conclusion and future work

A highly useful outcome of the implementation is that we have identified up to 350 citations containing discriminatory content, especially considering that, since 2015, only 234 citations of

this type had been identified based on user feedback and by using the manual methods described by Jensen et al. (2018, p. 142). We hope that by revising the DDO entries containing these citations, we can avoid some of the future user feedback regarding discriminatory content in the dictionary.

Furthermore, we have uncovered a previously overlooked type of problematic content in the old DDO citations—namely, citations describing violence and self-harm (e.g., suicide). Even though nine out of ten cases are erroneously identified by the model as relating to violence or self-harm—and reviewing the list is time-consuming due to the high number of false positives—we still consider the process valuable. We expect to end up with 300–350 problematic citations concerning violence and self-harm. This type of content was apparently more acceptable in Danish society in the 1980s and 1990s than it is today, now that the right of parents to hit their children has been abolished (1997) and official recommendations regarding the public discussion of suicide have been issued. The editorial team will discuss how this type of content should be handled in the dictionary.

It is interesting to compare the proportion of flagged examples among the older headwords in the printed edition—examples selected in the 1990s based on texts from 1983-1992—with the proportion among headwords described after 2010 based on more recent corpus texts. The proportion has fallen by 28% (from 6.3% to 4.52%). Although there are fewer examples of potential "harassment" among headwords edited in recent years, and although we know that the model incorrectly identifies many instances, it may still be a good idea to use the moderation model regularly to identify potentially problematic content.

In our future work, we will explore ways to identify the remaining problematic citations that the moderation model does not find—especially gender stereotypes, negative descriptions of physical disabilities or appearance, and the portrayal of women as sex objects. It may be appropriate to first examine some borderline cases and potentially lower the threshold value (0.5) slightly. One way to improve the model would be to use the citations in dataset 1 as training data. It might also be useful to study 'FALSE' cases close to the threshold value (0.5).

References

DDO: Hjort & Kristensen (Eds.) (2003-2005). *Den Danske Ordbog. Gyldendal og Det Danske Sprog- og Litteraturselskab;* since 2009: ordnet.dk/ddo. Det Danske Sprog- og Litteraturselskab.

Fjeld, R. Vatvedt (2015). Om ordbokseksempler og stereotypisering av kjønn i noen nordiske ordboker. In C. Sandström, I. Cantell, E.-R. Grönros, P. Nuolijärvi & E. Sommardahl (Eds.), *Särtryck ur Perspektiv på lexikografi, grammatik och språkpolitik i Norden* (pp. 35-65).

The Guardian (2020 a). *No more 'nagging wives': how Oxford Dictionaries is cleaning up sexist language*. Retrieved May 25, 2026, from https://www.theguardian.com/books/2020/mar/06/no-more-nagging-wives-how-oxford-dictionaries-is-cleaning-up-sexist-language

The Guardian (2020 b). *Dictionary.com revises definitions to eliminate prejudiced language*. Retrieved May 25, 2026, from https://www.theguardian.com/books/2020/sep/01/dictionarycom-revises-definitions-eliminate-prejudiced-language

Jensen, J., Lorentzen, H., Nimb, S., Svendsen, M.-M. Møller & Trap-Jensen, L. (2018). Thaipiger, muskelhunde og fulde svenskere: nedsættende ord, stereotyper og ligestilling i Den Danske Ordbog. In *Nordiske Studier i Leksikografi 14* (pp. 141-151).

Lazić, D. (2024). Lexicographic Treatment of Vocabulary Related to Age – The Example of Croatian and Danish. In K. Š. Despot, A. Ostroški Anić & I. Brač (Eds.), *Lexicography and Semantics - Proceedings of the XXI EURALEX International Congress* (pp. 633-647).

Lyse, G. I. & Svardal, T. (2020). "Den raude rase" og "negerfolk": kontroversielle ord i Bokmålsordboka og Nynorskordboka. *LexicoNordica*, 27, 77-96.

MailOnline (2021). *Influential German dictionary axes general masculine job titles in move to make the language more gender neutral – to the fury of traditionalists who say it is imposing 'politically correct expressions' on society*. Retrieved May 25, 2026, from https://www.dailymail.co.uk/news/article-9260687/Influential-German-dictionary-axes-general-masculine-job-titles-make-language-gender-neutral.html

Müller-Spitzer, C. & Rüdiger, J. O. (2022). The influence of the corpus on the representation of gender stereotypes in the dictionary. A case study of corpus-based dictionaries of German. Dictionaries and Society. In A. Klosa-Kückelhaus, S. Engelberg, C. Möhrs & P. Storjohann (Eds.), *Proceedings of the XX EURALEX International Congress* (pp. 129-141).

Petersson, S. & Sköldberg, E. (2020). To discriminate between discrimination and inclusion: a lexicographer's dilemma. In Z. Gavriilidou, M. Mitsiaki & A. Fliatouras (Eds.), *Proceedings of the XIX EURALEX International Congress* (pp. 381-386).

OpenAI's Moderation Model, omni-moderation-2024-09-26. Retrieved May 25, 2026, from https://developers.openai.com/api/docs/guides/moderation

Contact information

Sanni Nimb
Society of Danish Language and Literature
sn@dsl.dk

Ida Flörke
Society of Danish Language and Literature
if@dsl.dk

Agnes Aggergaard Mikkelsen
Society of Danish Language and Literature
aam@dsl.dk

Nathalie Hau Norman
University of Copenhagen
naha@hum.ku.dk

Nicolai Hartvig Sørensen
Society of Danish Language and Literature
nhs@dsl.dk

Dušica Terzić, Dejana Mirković-Birtašić & Saša Marjanović

TOWARDS A FRENCH–SERBIAN DICTIONARY OF SPORTS TERMS

Corpus-Based or AI-Assisted Methodology?

Abstract Earlier multilingual dictionaries of sports terminology that include French and Serbian were largely based on simple translations from foreign sources, without systematic terminological analysis whereas more recent works do not include French. To be able to address this gap and compile a reference dictionary, we first need to identify the optimal methodology for the compilation of this type of dictionary entries. This paper presents the evaluation of two different possible approaches — corpus-based and AI-assisted generation. First, we selected ten polysemous French sports terms with multiple Serbian equivalents across different contexts and generated dictionary entries using both corpus analysis and AI models. The corpus used in the process is a multilingual parallel corpus ParCoLab containing aligned English, French, and Serbian texts from official sports rulebooks of 14 sports, providing a solid basis for terminological extraction, whereas the AI models evaluated were GPT-5.5, Sonnet 4.6 and Gemini 3.6. The generated entries were then evaluated by human experts for accuracy, consistency, and practical usefulness. The results show that while AI significantly accelerates the compilation process, corpus-based methods remain more reliable. Combining both approaches, thus, proves to be most effective: corpus ensures data quality, while AI supports efficiency. However, final validation by a human terminologist remains essential for refining meanings, standardizing terminology, and ensuring long-term usability.

Keywords Dictionary project; corpus-based dictionary; AI generated entries; lexicographic processing; French-Serbian dictionary; sporting terminology

1 Introduction

The rapid development of corpus linguistics and the emergence of artificial intelligence have revolutionized the processes involved in dictionary compilation, profoundly transforming traditional lexicographic methodologies. This paper explores and evaluates two contemporary approaches — corpus-based and AI-based — that could be used in order to facilitate the compilation of a new French-Serbian Dictionary of Sports Terms. This dictionary is primarily conceived as a specialized terminological resource for advanced users, such as translators, journalists, and professionals working in sports-related fields whose native language is Serbian. From the perspective of function theory (*cf.* Tarp, 2008), the dictionary is intended to fulfil a cognitive function by highlighting both the similarities and the differences between the French and Serbian terminological systems. Furthermore, through the planned presentation of clearly defined and contrasted headwords, equivalents, and other microstructural elements, it aims to support the receptive function by facilitating the comprehensive understanding of sports terminology, as well as the productive function by enabling its accurate and appropriate use in both French and Serbian. In addition, by assisting users in selecting the most appropriate translation equivalents, the dictionary is also intended to fulfil a mediative function. The motivation for this new dictionary project arises from the recognition that existing French–Serbian lexicographic resources in the domain of sports terminology are no longer sufficient for contemporary use. As we will detail in the following sections, these resources exhibit significant limitations.

The limited adequacy of available lexicographic resources, combined with the innovative potential and significant influence of corpus linguistics and AI on lexicographic practice, prompts critical reflection on the optimal method for compiling a French-Serbian Dictionary

of Sports Terms. Our research, therefore, raises several key questions. Can AI accelerate the lexicographic process despite its demand for repeated fine-tuning and editorial oversight? Does a corpus-based methodology provide more precise, domain-specific terminology, whereas AI tends to generate more general or colloquial terms? What level of lexicographic intervention remains necessary for each method? Moreover, we investigate the potential synergy between these methods, considering whether AI-generated definitional templates could facilitate and speed up subsequent corpus-based editing, thereby reducing editorial workload without compromising quality.

To address the research questions, we structured this paper as follows. First, in Section 2, we provide an overview of the current lexicographic resources for sports terminology in both French and Serbian, highlighting their limitations. Next, we describe the methodology (Section 3) and evaluation metrics (Section 4) employed in this study. This is followed by a presentation and discussion of the results from both quantitative evaluations and qualitative analyses. Finally, we conclude with a summary of key findings and implications drawn from our research (Section 5).

2 Related work

According to the available bibliographies (i. e. Šipka, 2000), only two dictionaries of sports terminology include both French and Serbian. Both are multilingual dictionaries: in addition to these two languages, the first dictionary (Janković & Janković, 1979) also contains terminology in English and German, whereas the second (Beusan, 1984) includes terms in English, German, and Russian. The two dictionaries are not directly comparable, however, as they cover different sports disciplines and were produced for different sporting events.

The first dictionary was compiled for the VIII Mediterranean Games, held in Split (Croatia, Yugoslavia) in 1979. As stated in the preface (Janković & Janković, 1979, p. 5), it contains terminology relating to as many as twenty-five sports: archery, athletics, basketball, boxing, canoeing, cycling, diving, equestrian sports, fencing, football, gymnastics, handball, hockey, judo, rowing, rugby, sailing, shooting, swimming, table tennis, tennis, volleyball, water polo, weightlifting, and wrestling. An examination of the dictionary's macrostructure nevertheless reveals that the terminological coverage of these sports is neither equal nor balanced. Certain sports, particularly ball sports, receive considerably more extensive treatment than others. The second dictionary (Beusan, 1984) was produced for the 1984 Winter Olympic Games held in Sarajevo (Bosnia and Herzegovina, Yugoslavia) and therefore focuses exclusively on winter sports terminology. The disciplines covered include Alpine and Nordic skiing, cross-country skiing, ski jumping, biathlon, bobsleigh, luge, skating, speed skating, figure skating, and ice hockey.

The organization of the macrostructure differs significantly between the two dictionaries. In the first dictionary (Janković & Janković, 1979), entries are arranged alphabetically within a single sequence containing headwords from all included languages except Serbian. Consequently, a French entry may be preceded by an English term and followed by a German one. Although each headword is marked with a language label, this method of organization considerably reduces the transparency and usability of the macrostructure. Moreover, such an arrangement makes it impossible to determine either the total number of entries or the precise number of French headwords. In addition, the dictionary does not contain Serbian-language indexes that would enable reverse searches, for example from Serbian into French. The second dictionary (Beusan, 1984) adopts a different organizational principle. Its terminology is grouped according to individual sports disciplines, which allows the macrostructure to be characterized as predominantly onomasiological. Within each sports section, however, the entries

remain alphabetically organized in Serbian. Unlike the first dictionary, this one also contains indexes in which the terms are presented alphabetically in French, English, German, and Russian, each accompanied by cross-references to the corresponding Serbian terms. In this respect, both dictionaries may be considered to function as French–Serbian dictionaries.

The microstructure of both dictionaries is relatively limited. In the first dictionary (Janković & Janković, 1979), each entry consists of a headword followed by grammatical information (word class, gender, and number), a language label, and the corresponding equivalent(s). In most cases, a single equivalent is provided, although some entries contain two or more equivalents without any discriminating glosses or explanatory notes. Furthermore, only certain entries include an indication of the sport to which the term belongs. In the second dictionary (Beusan, 1984), the microstructural elements consist exclusively of equivalents in French, English, German, and Russian. No explicit indication of the relevant sport appears within the microstructure itself, since the entries are already organized according to sports disciplines at the macrostructural level.

Both dictionaries are intended primarily for native speakers of Serbian. Their structure and entry elements indicate that their primary function is communicative, more specifically mediative (or translational, *cf*. Tarp, 2008), insofar as they assist users in identifying translation equivalents. The first dictionary may also fulfil a receptive function when the meaning of the headword can be inferred from the equivalent provided. A similar receptive role is performed by the indexes in the second dictionary. At the same time, the second dictionary may additionally serve a productive translational function from Serbian into French (as well as into the other included languages), particularly when users are already completely familiar with the concepts, meaning and usage of the relevant terms.

It should nevertheless be noted that a more recent dictionary of Olympic and Paralympic sports has also been published (Ujlaki Šubić, 2010). Similar to the second dictionary discussed above, this resource organizes English headwords according to individual sports disciplines and provides parallel equivalents in French and Croatian. The dictionary was compiled primarily through the translation of English and French terminology, while specialists from the respective sports disciplines participated in the revision of the translations. In addition, the main body of the dictionary is supplemented by indexes in all three languages, thereby facilitating the retrieval of terms. Despite its potential usefulness, particularly given that Serbian and Croatian belong to the same pluricentric linguistic system, this dictionary is excluded from the present overview. The reason is that it provides exclusively Croatian standard terms (or proposed terms) as equivalents, which are not always identical to, nor necessarily used within, the Serbian standard variety. Furthermore, the dictionary was published 16 years ago and may therefore not adequately reflect current terminology and usage.

Although the foregoing discussion demonstrates the existence of resources for French–Serbian sports lexicography, the dictionaries in question can no longer be regarded as fully adequate for contemporary use. First, their selection of terms is limited, and none of them provides a systematic coverage of the sports terminology used in the languages concerned. Second, since these dictionaries were compiled more than forty years ago, they are now considerably outdated, both with respect to the selection of headwords and the processing of equivalents, especially given the rapid evolution of sports terminology, as of specialized terminology in general. Furthermore, the meaning of the terms is not explained through either definitions or brief glosses, which prevents users from obtaining precise semantic information. Closely related to this issue is the fact that, when multiple equivalents are provided, no indication is given of the semantic or usage-related differences between them. In addition, the polysemy of certain terms is not clearly represented. Consequently, these dictionaries lack a cognitive function

and do not enable an adequate comparison or contrastive analysis of the terminological systems of the languages involved. Finally, the dictionaries were not compiled on the basis of systematic terminological research, but rather through the decontextualized translation of headwords. For this reason, there is a clear need for a new dictionary that would provide an overview of the current state of French–Serbian sports terminology and that would be grounded in actual language use.

3 Methodology

The previous section demonstrated that the existing French–Serbian dictionaries of sports terminology were compiled through the decontextualized translation of headwords and that, as a result, their usability is limited today. In order to develop a new dictionary in which the lexicographic processing of the meaning and usage of headwords and equivalents would adequately fulfil the principal communicative and cognitive functions (*cf.* Section 1), particularly for users whose native language is Serbian — a user group considerably larger than that of native speakers of French in the context of French–Serbian lexicography — the present study aims to compare and evaluate two approaches to the processing of those microstructural elements of the dictionary entry concerned with the description of term meaning. Namely, the corpus-based method is compared to the AI-assisted entry compilation.

To examine the possibilities and limitations of these two approaches, ten French sports terms were selected according to several criteria. First, the selected terms are not multiword expressions, since this allows for greater collocational span and diversity of usage, as well as a higher likelihood of multiple equivalence. Second, all selected terms are nouns. Restricting the sample to a single word class ensures greater comparability among the terms and avoids differences arising from grammatical category. Third, the terms are used in two or more sports. One exception was nevertheless made in the case of the term *ashi-barai*, which occurs exclusively in karate. As a borrowing from Japanese, this term was included in order to examine how untranslatable loanwords of foreign origin are processed. Fourth, the selected terms are polysemous: either they develop secondary senses within the same sport (e.g. *but* 'goal' meaning both 'area' and 'result') or they are used with different meanings across different sports. Finally, the terms display multiple equivalence relations, either within the same sense, across different senses within the same sport, or across different sports. This final criterion was also intended to make it possible to examine the lexicographic ordering of multiple equivalents.

3.1 Reference entries

In order to enable a comparison of the two approaches, reference entries were constructed, serving as a basis against which the corpus-based and AI-based French–Serbian entries are compared. First, based on the Croatian trilingual dictionary (Ujlaki Šubić, 2010), described in Section 2, we identified the English equivalents of the ten selected French terms. These English terms were then looked up in the English–Serbian dictionary of sports terminology (Milić et al., 2021; *cf.* Milić et al., 2023). As noted in the review by Marjanović (2025), this dictionary can currently be considered an authoritative English–Serbian resource for Serbian sports terminology. It is based on a corpus-driven approach combined with a standardization-oriented perspective. The dictionary provides information not only on the meanings of the English terms, but also on the range of Serbian equivalents in use and their respective normative status. Equivalents are arranged according to their degree of attestation in the corpus used by the authors, as well as in accordance with relevant standardization policy. More specifically, the authors adopt a proscriptive approach (*cf.* Bergenholtz, 2003), the aim of which is not only to describe or prescribe forms already attested in usage, but also to guide users toward

normatively preferred equivalents and away from non-standard ones, thereby contributing to the future development of Serbian sports terminology. All entries in the dictionary were reviewed by specialists in the relevant sports, and the work follows modern lexicographic principles and complies with terminological standards.

The entry in the reference dictionary consists of an English headword, followed by grammatical information indicating the part of speech. It also includes cross-references to synonyms, antonyms, and variants. The Serbian equivalents are presented in numbered order, arranged according to their frequency of use and their status in terminological recommendations. Each sense is accompanied by a gloss, that is, a concise single-sentence definition. The gloss begins with a diatechnical label indicating the sport in which the term is used (Milić et al., 2023, pp. 52–54). For the purpose of the evaluation in this paper, for each identified English equivalent of the ten selected French nouns, the corresponding dictionary entries were consulted, and the sense definitions and Serbian equivalents were extracted for all sports in which both the English equivalent and the original French term (for which the reference entry was being compiled) are used. The Serbian equivalents were reproduced in the same order as in the reference dictionary. Cross-references were omitted, as no relationships of synonymy, antonymy, or variation exist among the ten terms analysed. The resulting reference entries served as a gold standard for the French–Serbian dictionary of sports terminology. A sample of these entries is provided below (*cf.* Table 1).

3.2 Corpus-based entries

The relevance and importance of corpus analysis, along with the numerous contributions made by examining empirical corpus data in the process of dictionary compilation have already been demonstrated in various (meta)lexicographic studies (*cf.* Hanks, 2009; 2012). We implemented those methodologies in conducting corpus-based processing of the selected ten French terms in the French–Serbian dictionary using the French–Serbian sports subcorpus of the multilingual parallel corpus ParCoLab (Miletić et al., 2017; Stosic et al., 2024). The use of this corpus for lexicographic purposes, with a view to compiling a general French–Serbian dictionary, has already been demonstrated through sample entries illustrating the treatment of selected polysemous headwords (e.g., Marjanović et al., 2018; Marjanović et al., 2019; Stosic et al., 2019). At present, ParCoLab contains approximately 105 million tokens distributed across twelve languages: French, Serbian, English, Spanish, Occitan, Italian, German, Portuguese, Russian, Alsatian, Corsican, and Poitevin–Saintongeais. The largest proportion of tokens belongs to Serbian, French, and English.

Originally conceived as a predominantly literary corpus with only a limited proportion of film transcripts (*cf.* Miletić et al., 2017), ParCoLab has since undergone a process of diversification through the inclusion of texts from various specialized domains (Terzić et al., 2020). In 2022, the corpus was further enriched with 42 official sports rulebooks aligned in English, French, and Serbian. This subcorpus was processed and aligned by the authors of this paper; however, its compilation process has not yet been documented. The most recent available versions of these multilingual rulebooks were incorporated for the following sports: athletics, basketball, boules, fencing, field hockey, football, handball, karate, korfball, rugby, swimming, table tennis, volleyball, water polo, and wrestling. As a result, the sports subcorpus currently contains 327,003 tokens in English, 341,390 in French, and 271,855 in Serbian.

Although the size of this sports-related subcorpus remains relatively modest in comparison with other corpora of sports terminology — *cf.* SrFudKo (Lazarević et al., 2023), which contains 10,100,553 tokens, but falls under only the football domain —, it is nevertheless adequate for

the purposes of the present study. Indeed, the terminology contained in the corpus has been validated by the relevant sports federations and is therefore believed free from the inappropriate terminology and textual "noise" commonly encountered in corpora based on, for example, newspaper discourse.

By searching the sports domain within ParCoLab, it is possible to identify Serbian equivalents of French terms and to observe both their usage within particular sports and their frequency of occurrence. For this reason, the present study draws on the ParCoLab sports subcorpus in order to explore the possibilities of applying a corpus-based approach to the compilation of a French–Serbian dictionary of sports terminology.

For each selected term, all relevant corpus occurrences were identified and extracted in.*csv* format, after which the data were subjected to detailed annotation. The annotation process consisted of several stages. First, each occurrence was assigned to a particular sport on the basis of the official sports rulebook in which it appeared. This made it possible to establish the corresponding sport label. The Serbian equivalents attested in the corpus were then identified and recorded for each occurrence. On the basis of both the sport labels and the attested equivalents, the sport-specific senses of the terms were subsequently determined. Semantic distinctions were first identified among occurrences sharing the same equivalent and then among those containing different equivalents. These distinctions were likewise incorporated into the annotation.

Particular attention was paid to occurrences containing either explicit definitions or contextual information that could facilitate the formulation of appropriate definitions. The annotation of sport labels and equivalents was carried out by a single lexicographer, while the annotation of meanings and definitional information was independently verified by three lexicographers in order to ensure consistency and reliability.

The annotated corpus data subsequently served as the basis for the compilation of French–Serbian entries derived from the ParCoLab corpus. In these entries, the Serbian equivalents were ordered according to their frequency of occurrence within a given sense and sport domain. A sample of these entries is presented in the Table 1.

TABLE 1 Reference and corpus-based entry samples

Reference entry sample	Corpus-based entry sample
BARRE, imenica ženskog roda A. *letvica* U atletici, horizontalno postavljena šipka, odnosno prepreka koju skakač uvis ili skakač motkom nastoji da preskoči izvijanjem tela iznad nje. B. *prečka* 1. U fudbalu, vodoravna obla greda kružnog preseka koja spaja stubove gola na visini od 2,44 m, mereno od donjeg kraja prečke. 2. U rukometu, vodoravna greda kvadratnog preseka koja povezuje stative gola na visini od 2 m od tla. 3. U vaterpolu, vodoravna greda pravougaonog preseka koja povezuje stative gola na visini od 90 cm od površine vode.	**BARRE**, imenica ženskog roda A. *šipka koja se preskače* (atletika) **letvica, letva** B. *vodoravna šipka na golu* (fudbal) **prečka** (ragbi) **prečka** (vaterpolo) **prečka**

3.3 AI-based entries

The emergence of AI calls into question traditional techniques and methodologies applied in any kind of human endeavour involving textual data, dictionary compilation not being excluded. The capacity of LLMs to generate and translate texts has been the subject of various evaluations (*cf*. Arias-Arias & Cancela, 2025; Filipović Petrović & Beliga, 2025; Greco et al., 2025). In this context, we evaluated the capacity of three different LLMs to generate dictionary entries and compared it to the results of the abovementioned corpus-based approach. We chose OpenAI's GPT-5.5, Google's Gemini 3.6 Flash set to extended level and Anthropic's Claude Sonnet 4.6 set to high effort as those models were the latest free versions of the respective companies at the time of our evaluation.

In order to provide comparability, the same ten French sports terms selected for the corpus-based approach (*cf*. Section 3.2) were used as a basis for the prompt. The same prompt was used to generate all of the entries in all of the used LLMs. For the entries to be comparable, the prompt was modelled on the structure of the reference entries (*cf*. Section 3.1 and Table 1), drawing on the entry format employed in the authoritative reference dictionary (Milić et al., 2021). The output, thus, includes the headword, grammatical information, numbered standard Serbian equivalent terms as well as definition of meaning starting with a diatechnical label of the respective sport, for example: "U fudbalu, . . . " ('In football . . . '). The following prompt was used to generate dictionary entries:

> "You are a lexicographer specialized in terminology. You are compiling a French-Serbian dictionary of sports terminology particularly from these sports: athletics, basketball, boules, fencing, field hockey, football, handball, karate, korfball, rugby, swimming, table tennis, volleyball, water polo, and wrestling. Your first task is to compile 10 dictionary entries, for the following French terms: "ashi-barai", "barre", "but", "échange", "faute", "hors-jeu", "panier", "passe", "point", "terrain". Each entry should contain the following elements:
>
> 1. headword
>
> 2. grammatical information indicating the part of speech, gender and/or number, and transitivity, where applicable (do not put the abbreviations)
>
> 3. Serbian equivalents, numbered and ordered according to their frequency of use. If different equivalents are used for the same sense in different sports, the relevant equivalent(s) should be preceded by a diatechnical label indicating the sport in which they are used.
>
> 4. definition of meaning, formulated in Serbian. A definition should begin with a diatechnical label indicating the sport in which the term is used (for example: "U fudbalu, . . . ").
>
> If a French term has multiple senses, each sense should be presented in a separate section labelled with a capital letter (A, B, C, etc.)."

After the three models were queried with the same prompt, the output[1] was evaluated by three linguists, which is described in the next section. We decided not to intervene and correct the compiled entries so as to evaluate the output of the LLMs and the level of their accuracy and the adequacy of the obtained result. It was necessary, however, to add a second prompt to the GPT model since its output included metalexicographic terms unsuitable for the target user group of the French–Serbian dictionary of sports terms, as well as grammatical information in English. We intervened, therefore, solely in this case by precising the need for the output to be in the form of the dictionary entry in Serbian and without metalexicographic terms.

[1] Links: https://chatgpt.com/share/6a1b08e8-b98c-83eb-893d-a86533f2fa6c; https://claude.ai/share/bd3e29a1-89c6-4915-8274-56d00a1b0334; https://gemini.google.com/share/99223c75042e

4 Evaluation

For the evaluation of the translation of multiword expressions generated by LLMs, Greco et al. (2025, pp. 709–711) adapted the traditional metrics used in the field of translation. The reason for the adaptation lies in the fact that they evaluated smaller text portions than the traditional translation evaluations, and the subject of their evaluation has more specific implications (Greco et al., 2025, pp. 709–710). Such is the case in our research, which is why we were inspired by the metrics used in their work. We did, however, adapt it further to suit our needs. All of the four microstructural elements demanded from the LLMs have their place in the evaluation as well as the overall syntax and morphological correctness. The final metrics used in our evaluation are:

— **Overall correctness**: is the rendering in the target language clear and correct in terms of spelling, morphology and syntax?

— **Class Adequacy**: is the grammatical information correct?

— **Accuracy and Functional Equivalence**: is the suggested equivalent accurate? Are any relevant equivalents omitted? Does the order of the equivalents match their frequency in use? Does the equivalent belong to the standard register?

— **Context Usage Suitability**: is the sport suggested in the definition correct? Are all of the sports in which the term is used represented?

— **Completeness or Superabundance**: Are all of the proposed senses accurate? Is there a sense that is lacking? Are they correctly separated?

— **Comprehensiveness**: is the definition clear and exhaustive?

Once the metrics were defined, three evaluators proceeded to evaluate the four proposed entries (three AI generated entries and one created manually through the corpus-based approach) for each of the 10 sports terms, using the reference entries as a benchmark (*cf.* Section 3.1). For this purpose, a 5 point Likert-type scale was used with the following point distribution: 1 — 'very poor'; 2 — 'poor'; 3 — 'acceptable'; 4 — 'good'; 5 — 'excellent'. Once the three evaluators finished individual evaluation, the results were discussed and harmonised. We present them in the next section.

5 Results and discussion

As we have already stated (*cf.* Section 3), the gold standard reference entry was based on the entry from the authoritative English–Serbian reference dictionary. The drawbacks of this method emerged during the evaluation. First, some inconsistencies were observed in the reference entries during the evaluation of these ten entries: in some cases, the definitions include specific technical details, such as the radius of the basket in basketball, whereas in others such details are absent. Second, some of the grammatical information was not provided in the reference entry, whereas it was generated by the two LLMs; for example, *nepromenljivo* ('invariable') is a grammatical label missing in the reference entry but produced by GPT and Claude for a Japanese loanword *ashi-barai*. Third, sports covered by the ParCoLab corpus, used for the generation of corpus-based entries, and those covered by the reference dictionary do not overlap entirely. While many of the sports are the same between the two resources, some equivalents and senses were not directly comparable. However, this limitation can also be seen as an advantage, since the comparison of the two resources may help generate entries covering a wider range of sports.

TABLE 2 Results of the evaluation of the AI-based entries by three individual lexicographers. The criteria are distributed as follows: **Overall correctness**: **1** – Rendering in the target language is clear and correct in terms of spelling, morphology and syntax; **Class Adequacy**: **2** – Grammatical information is correct; **Accuracy and Functional Equivalence**: **3** – Suggested equivalents are accurate; **4** – There are no equivalents lacking; **5** – Order of the equivalents matches their frequency in use; **6** – Equivalents belong to the standard register; **Context Usage Suitability**: **7** – Sport suggested in the definition is correct; **8** – All of the sports in which the term is used are represented; **Completeness or Superabundance**: **9** – All of the senses are presented; **10** – There are no senses lacking; **11** – The senses are correctly separated; **Comprehensiveness**: **12** – Definition is clear.; **13** – Definition is exhaustive

Gpt-5.5	ashi-barai	barre	but	échange	faute	hors-jeu	panier	passe	point	terrain	average
1	0.73	0.93	1	1	0.93	1	1	1	1	1	**0.96**
2	1	1	1	1	1	1	1	1	1	1	**1**
3	0.67	1	0.53	0.73	0.87	0.87	1	1	0.80	1	**0.85**
4	0.53	1	0.47	0.53	0.67	0.80	0.73	0.67	0.73	0.80	**0.69**
5	0.53	0.93	0.53	0.73	0.80	0.93	0.93	0.73	0.93	1	**0.81**
6	0.60	1	0.67	0.73	1	0.80	1	1	1	1	**0.88**
7	1	0.93	1	1	0.87	1	1	1	0.87	1	**0.97**
8	1	0.87	1	1	0.73	1	0.80	1	0.93	0.87	**0.92**
9	1	1	0.53	1	0.87	1	0.93	1	0.87	0.93	**0.91**
10	1	1	0.33	1	0.67	1	0.93	1	0.60	0.73	**0.83**
11	1	1	0.60	1	0.60	1	1	1	0.40	0.93	**0.85**
12	0.80	1	0.67	1	0.93	1	1	1	0.80	0.93	**0.91**
13	0.87	0.73	0.47	0.80	0.73	0.93	0.87	0.73	0.67	0.80	**0.76**
avg	**0.83**	**0.95**	**0.68**	**0.89**	**0.82**	**0.95**	**0.94**	**0.93**	**0.82**	**0.92**	
Sonnet 4.6	ashi-barai	barre	but	échange	faute	hors-jeu	panier	passe	point	terrain	average
1	0.87	1	0.87	0.93	1	1	1	1	1	1	**0.97**
2	1	1	1	1	1	0.93	1	1	1	1	**0.99**
3	0.80	0.87	0.80	0.67	0.87	0.80	0.93	1	0.80	1	**0.85**
4	0.60	1	0.93	0.60	0.80	0.80	0.80	1	1	0.80	**0.83**
5	0.53	1	0.93	0.73	0.93	0.87	1	1	0.93	1	**0.89**
6	0.67	1	0.67	0.87	1	1	0.93	1	1	1	**0.91**
7	1	1	1	0.67	0.93	1	1	1	0.93	1	**0.95**
8	1	1	1	0.80	1	1	1	1	0.87	0.87	**0.95**
9	1	1	1	0.67	0.93	1	1	1	0.73	0.93	**0.93**
10	1	1	1	0.80	0.73	1	1	1	0.53	0.73	**0.88**
11	1	1	1	0.53	0.80	1	1	1	0.40	0.80	**0.85**
12	0.80	0.93	0.93	0.67	1	1	1	0.93	0.80	0.93	**0.90**
13	0.87	0.87	0.80	0.67	0.93	0.93	1	0.87	0.67	0.73	**0.83**
avg	**0.86**	**0.97**	**0.92**	**0.74**	**0.92**	**0.95**	**0.97**	**0.98**	**0.82**	**0.91**	

TABLE 2 (*CONTINUED*)

Gemini 3.5	ashi-barai	barre	but	échange	faute	hors-jeu	panier	passe	point	terrain	average
1	1	1	0.93	1	1	1	1	1	1	1	**0.99**
2	0.93	1	1	1	1	1	1	1	1	1	**0.99**
3	1	1	0.80	1	0.87	0.80	0.93	1	0.93	1	**0.93**
4	0.87	0.47	0.87	0.73	0.80	0.80	0.87	1	1	0.80	**0.82**
5	1	0.73	0.87	1	0.93	1	1	1	1	1	**0.95**
6	1	1	0.80	0.80	1	1	0.93	1	1	1	**0.95**
7	1	1	0.93	1	1	1	1	1	1	1	**0.99**
8	1	0.47	1	0.53	1	1	1	1	1	0.87	**0.89**
9	1	0.67	0.73	0.53	0.93	1	1	1	0.87	0.93	**0.87**
10	1	0.47	0.80	0.53	0.80	1	1	1	0.73	0.73	**0.81**
11	1	0.73	0.40	0.60	0.80	1	0.60	1	0.73	0.93	**0.78**
12	1	0.93	0.80	0.93	1	0.93	0.93	1	0.87	0.93	**0.93**
13	1	0.60	0.80	0.73	0.93	1	1	0.93	0.73	0.87	**0.86**
avg	**0.98**	**0.77**	**0.83**	**0.80**	**0.93**	**0.96**	**0.94**	**0.99**	**0.91**	**0.93**	

During the analysis stage, we first compared the three AI models to determine the most suitable one for our lexicographic purposes. The results of the evaluation are presented in Table 2. Even though the evaluators used a 5 point Likert-type scale, the results were rescaled to the [0, 1] range following the methodology used by Greco et al. (2025, p. 711) in order to facilitate possible further evaluations, especially automatic ones.

Even though the AI-generated entries might initially look acceptable, a comparison with the reference entries highlighted major inconsistencies. In particular, sense disambiguation emerged as the primary challenge. Evaluators struggled to identify criteria used by the LLM to distinguish between different senses. For example, there were cases in which the senses were split according to different sports. Conversely, clearly distinct senses were sometimes merged into a single sense within the context of a specific sport, resulting in a single expression being used for two distinct senses.

Quantitative evaluation revealed that Google's Gemini 3.6 Flash achieved the highest overall score. Overall, the entries generated by Gemini showed the highest degree of similarity to the reference entries. Moreover, the shortcomings observed across the AI-generated entries were generally less pronounced in Gemini's output, which further supports the conclusion that this model achieved the best overall performance in the present evaluation. However, certain important discrepancies among evaluators underscored the need for more precise evaluation criteria. For example, ambiguity arose regarding how to score articles that presented some accurate senses but were overall incomplete. This indicated that AI-generated entries alone would not suffice for compiling a high-quality French–Serbian dictionary of sports terms.

We then compared dictionary entries generated by Gemini with those produced through a corpus-based approach. Evaluators rated corpus-based entries higher in quality on almost all of the criteria. The only criteria where the AI model performed better concerned definitions. It is important to emphasize that the corpus-based entries did not contain fully processed definitions, but solely information important to distinguish between senses, which led to poorer

performance in terms of points. Further analyses of the entries reflects that AI-generated definitions tend to be more general, likely due to their training on diverse, non-specialized data sources. As a result, AI models were less capable of producing terminologically precise equivalents and definitions, which is highly relevant since the terms vary from sport to sport and the same term can have a different sense in different sports. In contrast, the ParCoLab corpus used in this study contains authoritative resources, such as the official sports rulebooks, which contribute substantially to terminologically precise and contextually appropriate entries. Consequently, consulting expert-verified sources or domain specialists remains crucial.

TABLE 3 Results of the evaluation of the corpus-based entries by three individual lexicographers. The criteria are the same as in the Table 2

Corpus Metrics	ashi-barai	barre	but	échange	faute	hors-jeu	panier	passe	point	terrain	average
1	0.87	1	1	1	1	1	1	1	1	1	**0.99**
2	0.93	1	1	1	1	1	1	1	1	1	**0.99**
3	0.93	1	0.93	0.93	0.80	1	0.93	1	1	1	**0.95**
4	0.67	1	0.80	0.73	0.80	0.80	0.93	0.87	0.87	0.80	**0.83**
5	0.87	1	0.67	0.87	0.87	0.93	0.93	0.87	1	1	**0.90**
6	0.93	1	0.80	0.67	1	1	1	1	1	1	**0.94**
7	1	1	0.87	1	0.93	1	1	1	1	1	**0.98**
8	1	0.80	1	1	1	1	1	0.93	0.93	0.87	**0.95**
9	1	1	1	0.87	0.87	1	1	1	0.73	0.93	**0.94**
10	1	1	0.87	0.80	0.73	1	1	1	0.60	0.73	**0.87**
11	1	1	0.80	0.60	0.80	1	1	1	0.40	0.80	**0.84**
12	0.73	0.87	0.80	0.80	1	0.87	0.87	0.87	0.80	0.93	**0.85**
13	0.53	0.60	0.67	0.67	0.73	0.73	0.67	0.47	0.47	0.60	**0.61**
average	**0.88**	**0.94**	**0.86**	**0.84**	**0.89**	**0.95**	**0.95**	**0.92**	**0.83**	**0.90**	

Further qualitative examination of definitions produced by Gemini suggested that improvements might be achieved through prompt fine-tuning, particularly to increase alignment with the dictionary's focus on specialized subject and to provide more detailed information on individual sports. Accordingly, we developed a refined prompt explicitly clarifying the dictionary's characteristics and intended functions. These included specifying its target user groups (as outlined previously) and emphasizing its various functions: receptive (to facilitate correct understanding of French terms), productive (to enable accurate and context-appropriate use of French terms and their Serbian equivalents) both in free context and translation.

Despite these refinements, AI-generated entries still contained erroneous information. For instance, after prompt fine-tuning for Japanese term *ashi-barai*, Gemini 3.5 offered two possible entries as illustrated in Figure 1. Option A stated that one of the proposed equivalents (*čišćenje stopalom*) was privileged by the official translation of the World Karate Federation official sports rulebook. However, when we verified it against the ParCoLab corpus, which contains the original source referenced by the AI, this information turned out to be false. On the other hand, option B stated *nožna metlica* as the equivalent, which does not occur in the ParCoLab corpus nor on the simple Internet search. This example underscores that, even with carefully designed prompts, AI-generated dictionary entries are not fully reliable and must undergo verification by a human specialist in sports terminology.

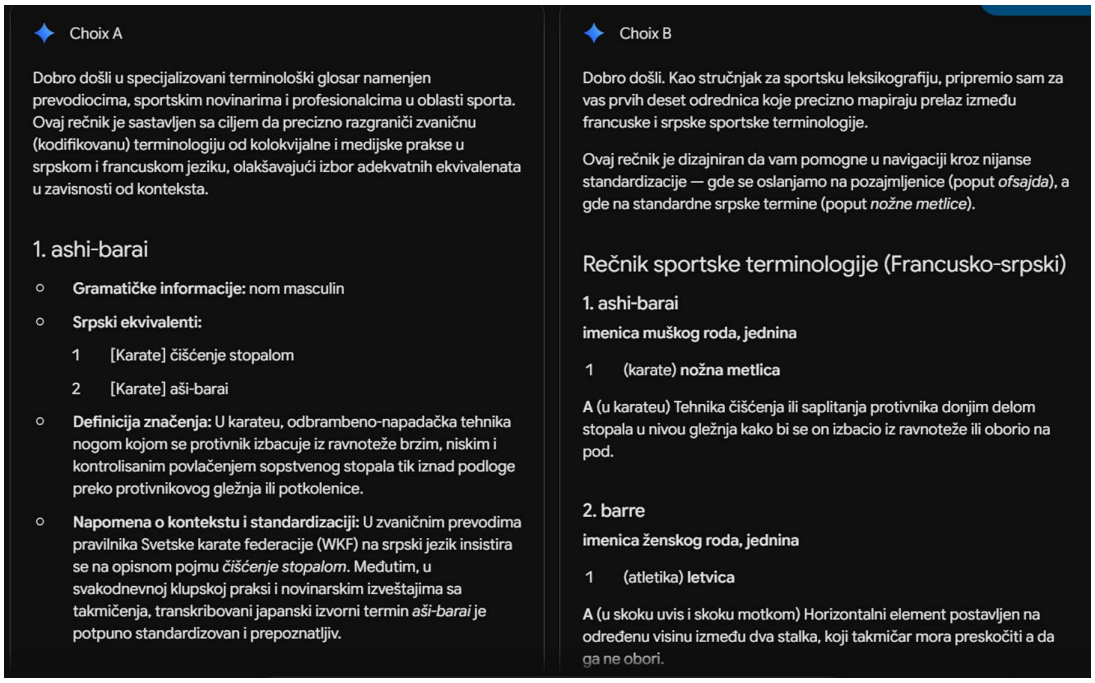


FIGURE 1 Gemini 3.5 output after fine-tuning offering two choices which contain (A) false information (B) non-existing equivalent

These findings confirm our initial hypothesis. Contrary to expectations that AI could significantly reduce editorial effort and time, substantial lexicographic intervention remains indispensable. This process often requires iterative trial and error. Lexicographers must meticulously review and enhance AI outputs while carefully curating corpus-based entries to ensure precision and terminological adequacy. Nonetheless, AI shows potential as an aid for generating definitional templates. Such templates can subsequently be verified and enhanced through corpus-based evidence, and especially by consulting domain experts and terminologists, thus creating a complementary workflow that supports productivity and facilitates the lexicographer's task.

6 Conclusion

In this study we evaluated two approaches — corpus-based and AI-based — for developing a specialized French–Serbian dictionary of sports terminology, motivated by notable gaps in existing lexicographic resources. Our quantitative and qualitative analyses addressed the initial hypotheses and led to several key conclusions. Efficient time management is often cited as a primary advantage of using AI models in lexicographic work. However, our research revealed that while AI offers potential time savings through rapid entry generation, it requires multiple iterations of prompt fine-tuning to generate suitable dictionary entries, resulting in a potentially lengthy trial-and-error process. Moreover, even carefully curated prompts do not guarantee fully reliable outputs. AI outputs could contain inaccuracies and fabricated information, underscoring the essential role of human validation. Conversely, the corpus-based approach consistently produced more precise, contextually relevant equivalents, leveraging authoritative sources that ensure terminological accuracy within the sports domain. It should be noted, however, that these conclusions can only be drawn upon verification with domain experts and terminologists.

Nevertheless, AI-generated entries showed promise as initial templates that lexicographers can further refine using corpus data and by consulting domain experts and terminologists. Such a combined approach could enhance overall time efficiency while ensuring the production of reliable, detailed, and up-to-date dictionary entries. Ultimately, our goal was to identify the most effective methodology for compiling a contrastive French–Serbian sports dictionary tailored to expert users. By demonstrating how AI and corpus-based methods can complement one another, we contribute to lexicographic methodology and specialized dictionary-making. Finding an optimal balance between automated generation and editorial work has the potential to produce the most effective French–Serbian dictionary of sports terminology. Grounded in empirical data and supported by modern computational tools, this lexicographic resource would offer a clear, contrasted presentation of terminological systems, thereby improving comprehension and precise usage within the specialized domain of sports. Such a resource would fill a critical gap in the current French–Serbian specialized lexicographic landscape.

References

Arias-Arias, I. & Cancela, E. (2025). Bridging human and AI perspectives: Semantic annotation of generic nouns in German. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda & A. Salgado (Eds.), (Eds.), *Electronic lexicography in the 21st century (eLex 2025): Intelligent lexicography. Proceedings of the eLex 2025 conference* (pp. 1–18). Lexical Computing CZ s.r.o.

Bergenholtz, H. (2003). User-oriented Understanding of Descriptive, Proscriptive and Prescriptive Lexicography. *Lexikos*, 13. https://doi.org/10.5788/13-0-722

Beusan, A. (Ed.). (1984). *Sportski rječnik*. Svjetlost.

Filipović Petrović, I. & Beliga, S. (2025). Can AI Understand Croatian Idioms? Assessing Large Language Models in Lexicographic Tasks. *Contributions to Contemporary History*, 65(3), 218–242. https://doi.org/10.51663/pnz.65.3.09

Greco, A., Delsanto, M., Di Fabio, A., Mori, L., Onesti, C., Radicioni, D. P. & Scozzaro, C. J. (2025). Compiling Bilingual Dictionaries: AI-Assisted Translation of Italian Multiword Expressions into English and French. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda & A. Salgado (Eds.), *Electronic lexicography in the 21st century (eLex 2025): Intelligent lexicography. Proceedings of the eLex 2025 conference* (pp. 692–718). Lexical Computing CZ s.r.o.

Hanks, P. (2009). The impact of corpora on dictionaries. In P. Baker (Ed.), *Contemporary Corpus Linguistics* (pp. 214–236). Continuum.

Hanks, P. (2012). The corpus revolution in lexicography. *International Journal of Lexicography*, 25(4), 398–436. https://doi.org/10.1093/ijl/ecs026

Janković, M. & Janković, R. (Eds.). (1979). *Sportski rječnik*. Direkcija VIII mediteranskih igara.

Lazarević, J., Stanković, R., Škorić, M. & Rujević, B. (2023). Football terminology: Compilation and transformation into OntoLex-Lemon resource. In S. Carvalho, A. F. Khan, A. Ostroški Anić, B. Spahiu, J. Gracia, J. P. McCrae, D. Gromann, B. Heinisch & A. Salgado (Eds.), *Proceedings of the 4th Conference on Language, Data and Knowledge* (pp. 634–645). NOVA CLUNL.

Marjanović, S., Stosic, D. & Miletic, A. (2018). A Sample French-Serbian Dictionary Entry based on the ParCoLab Parallel Corpus. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (Eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts* (pp. 423–435). Znanstvena založba Filozofske fakultete Univerze v Ljubljani /Ljubljana University Press, Faculty of Arts.

Marjanović, S., Stošić, D. & Miletić, A. (2019). Paralelni korpus ParCoLab u službi srpsko-francuske leksikografije. In J. Novaković & M. Srebro (Eds.), *Srpsko-francuske književne i kulturne veze u evropskom kontekstu 1* (pp. 279–307). Novi Sad: Matica srpska – Bordeaux: Université Bordeaux Montaigne.

Marjanović, S. (2025). Krupan leksikografski (is)korak u dosadašnjoj dvojezičnoj terminografiji: Mira Milić, Olga Panić Kavgić, Aleksandra Kardoš. Novi englesko-srpski rečnik sportskih termina. Novi Sad: Prometej, 2021. - 536 str. (review). *Filološki pregled*, 52(1), 155-167.

Miletic, A., Stosic, D. & Marjanović, S. (2017). ParCoLab: A Parallel Corpus for Serbian, French and English. In K. Ekštein & V. Matoušek (Eds.), *Text, Speech, and Dialogue* (Vol. 10415, pp. 156–164). Springer International Publishing.

Milić, M., Panić Kavgić, O. & Kardoš, A. (Eds.) (2021). *Novi englesko-srpski rečnik sportskih termina*. Prometej.

Milić, M., Panić Kavgić, O. & Kardoš Stojanović, A. (2023). A New English–Serbian Dictionary of Sports Terms in the Light of Contemporary Challenges. *Lexikos*, 33, 47–67. https://doi.org/10.5788/33-1-1796

Stosic, D., Marjanović, S., Bernhard, D., Bach, X., Bras, M., Kevers, L., Retali-Medori, S., Vergez-Couret, M. & Werner, C. (2024). The ParCoLab Parallel Corpus and Its Extension to Four Regional Languages of France. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 16014–16023). ELRA and ICCL.

Šipka, D. (2000). *A Bibliography of Serbo-Croatian Dictionaries: Serbian, Croatian, and Bosnian Muslim*. Dunwoody Press.

Tarp, S. (2008). *Lexicography in the borderland between knowledge and non-knowledge: General lexicographical theory with particular focus on learner's lexicography*. Max Niemeyer Verlag.

Terzić, D., Marjanović, S., Stosic, D. & Miletic, A. (2020). Diversification of Serbian-French-English-Spanish Parallel Corpus ParCoLab with Spoken Language Data. In P. Sojka, I. Kopeček, K. Pala & A. Horák (Eds.), *Text, Speech, and Dialogue* (Vol. 12284, pp. 61–70). Springer International Publishing.

Ujlaki Šubić, Lj. (Ed.) (2010). *Rječnik olimpijskih i paraolimpijskih športova: Engleski, francuski, hrvatski; ljetne igre*. Školska knjiga.

Contact information

Dušica Terzić
University of Belgrade – Faculty of Philology
dusica.terzic@fil.bg.ac.rs

Dejana Mirković-Birtašić
University of Belgrade – Faculty of Philology
dejana.mirkovic.birtasic@fil.bg.ac.rs

Saša Marjanović
University of Belgrade – Faculty of Philology
sasa.marjanovic@fil.bg.ac.rs

Ene Vainik, Heete Sahkai, Ahto Kiil, Jelena Kallas & Geda Paulsen

SEMANTIC CLUSTERING AND CATEGORISATION OF CONSTRUCTIONAL COLLEXEMES USING LLMS

Abstract This paper investigates whether large language models can perform semantic clustering and categorisation of constructional collexemes to support the analysis of constructional meaning and the organisation of collexemes within constructicon entries. As a case study, we examine the collexemes of the Estonian Nominal Quantifier Construction, identified from lexicographic and corpus data. Using OpenAI's ChatGPT-5.4, we conducted an informed free sorting task and a closed sorting task with five input types, consisting either in bare lemmas or lemmas accompanied by different types of context: corpus phrases, corpus sentences, dictionary definitions, and dictionary examples. Model outputs were evaluated against a human-created gold standard using Adjusted Mutual Information, Adjusted Rand Index, and a label quality rating scheme. A scaled closed sorting experiment was also conducted. The results of the free sorting tasks approached human agreement levels, with dictionary definitions yielding the most similar clustering and corpus sentence input the most acceptable labels. The results of the closed sorting task and the scaling experiment demonstrated that a bare list of lemmas provided sufficient input for assigning collexemes to predefined semantic categories.

Keywords Constructicography; lexicography; computational semantics; colloprofile; Estonian

1 Introduction

This paper presents work in progress from the Estonian Constructicon (Vainik et al., 2024), which is being compiled as an extension of the EKI Combined Dictionary within the lexicographic database Ekilex (Tavast et al., 2018) and is intended to be accessible through the EKI Combined Dictionary interface Sõnaveeb (Koppel et al., 2019).

Constructicons are currently being compiled for a wide range of languages (see Zhukova, 2026, p. 22). In addition to information on the form and meaning of constructions, constructicons may also provide information on constructions' CEFR levels (e.g. Janda et al., 2020) and on relationships between constructions (e.g. Ziem et al., 2019). Some constructicons, e.g. CASA – ConstructiCon of the English Language (Herbst & Hoffmann, 2024) and the German Constructicon (Ziem et al., 2019), also include information on constructional collexemes (Stefanowitsch & Gries, 2003), i.e., the words attracted to an open slot of a (semi-) schematic construction. To refer to the set of collexemes associated with a constructional slot, (Herbst, 2018, 2020) introduced the term *constructional collo-profile*. As Ziem & Feldmüller (2023) point out, collo-profiles provide information about which individual words are relevant for particular constructional slots, indicate usage preferences, and serve as a useful indicator of constructional semantics.

The present study addresses the organisation of the collo-profile in a constructicon entry. In existing constructicons, collo-profiles are organised as frequency-based lists of collexemes (Herbst & Hoffmann, 2024, p. 79) or as ranked lists of collexemes together with association scores and the number of sentences in which a collexeme was predicted as a suitable filler (Ziem & Feldmüller, 2023, p. 220). In the Estonian Constructicon, we envisage moving beyond lists and enabling the collo-profile to be sorted and filtered based on multiple criteria: raw co-occurrence frequency of the collexemes and the construction, association strength between

the collexemes and the construction, the CEFR-levels of collexeme-construction pairings (Kallas et al., 2025), and, finally, the semantic categories of the collexemes. The focus of this study is on the semantic categorisation of collexemes. This information is expected to provide a useful way of presenting the collo-profile in a constructicon entry, and, additionally, to contribute to the identification of constructional meaning.

Given the aim of representing a large number of constructions together with their potentially extensive sets of collexemes in a semantically structured way, it is necessary to develop scalable methods for the semantic categorisation of collexemes. This study therefore aims to develop methods to automatise the semantic clustering and categorisation of collexemes in the constructicographic workflow. In particular, we investigate whether and how LLMs can be used for this purpose. In our previous research, we have found that LLMs can successfully identify instances of schematic constructions in L1 corpora (Sahkai et al., 2025) as well as in L2 learner corpora (Kallas et al., 2026). Corpus retrieval of instances of constructions enables identification of the collexemes of (semi-) schematic constructions, along with their co-occurrence frequency and association strength with the construction. In this study, we test whether LLMs can also be used for the semantic clustering and categorisation of collexemes.

The semantic analysis of constructional collexemes predates the emergence of constructicon-building projects. Two broad and complementary approaches can be distinguished. The first approach is interpretive categorisation, where collexemes are semantically grouped based on predefined inventories or conceptual frameworks like event types (Goldberg, 1995, p. 39; Levshina, 2016), verb classes (Barðdal, 2008, pp. 63–63; Dekalo & Hampe, 2017), or FrameNet (Fillmore et al., 2002) frames (Hartmann & Willich, 2024; Sundquist, 2020). Manual categorisation remains widely used in the interpretive categorisation tradition. In this study, we test whether LLMs can be used to assign collexemes to pre-defined categories.

The second approach is data-driven clustering. This approach consists of three steps. First, the pairwise similarity of all the collexemes is calculated, based on their distribution (word embeddings, co-varying collexemes) or semantic relations (WordNet-based similarity metrics). Secondly, a cluster analysis is performed, using a clustering method (Gries & Stefanowitsch, 2010; Dekalo & Hampe, 2017; Perek, 2016, 2018) or network science (Chen, 2022, 2025; Dekalo & Hampe, 2017; Ellis et al., 2014; Perak & Ban Kirigin, 2023). Finally, the clusters must be interpreted and labelled. These approaches offer scalability but require additional interpretation to yield meaningful categories. In the present study, we instruct an LLM to cluster a set of collexemes into categories and to provide labels for the categories, aiming to replace the three-step procedure described above with a single prompt.

In practical constructicography, both the data-driven and interpretive approach may be needed. The interpretive approach can be used when a semantic classification, for instance, an inventory of frames, verb classes or semantic types, has been adopted. The data-driven approach can be used when no such inventory is available. The present study investigates whether LLMs can be used for both approaches, enabling the clustering of collexemes into labelled semantic categories, as well as the automated assignment of collexemes to predefined categories. We also test a hybrid approach in which a human-validated clustering based on a smaller dataset is applied by the LLM to a larger dataset. Our research questions were formulated as follows:

RQ1: Can an LLM produce a semantic clustering of constructional collexemes that approximates human judgement in terms of systematicity, consistency, granularity, and labelling?

RQ2: How well can an LLM assign collexemes to predefined categories?

RQ3: Which type of input data (bare lemmas or lemmas accompanied by corpus contexts or dictionary definitions and usage examples) yields the most reliable results?

RQ4: Can an LLM apply a predefined categorisation system based on a smaller set of data to a larger dataset of a similar kind?

In the following, we first introduce the data and procedures. We then present the results and discuss their implications.

2 Data and procedures

2.1 The example construction: The Estonian Nominal Quantifier Construction

The Estonian Nominal Quantifier Construction (NQC) is a (pseudo-)partitive construction consisting of two nouns (see e.g. (Metslang, 2017; Pilvik et al., 2025). The first noun functions as a quantifier, and the second noun denotes the referent (in partitive use) or the kind of entity (in pseudo-partitive use) quantified over (Koptjevskaja-Tamm, 2001), e.g. *tükk seda torti* [piece DEM.PAR cake.PAR] 'a piece of this cake', *tükk paberit* [piece paper.PAR] 'a piece of paper'. Nouns vary as to their ability to be used as quantifiers: some do so directly (e.g., *liiter* 'litre', *kilogramm* 'kilogram', *hetk* 'moment'), others via productive word formation (e.g. by compounding/affixing with the element *-täis* '-ful' (*lusikas* 'spoon' > *lusikatäis* 'spoonful')) or via meaning change, e.g., via metonymic transfer from the name of a container to the amount of substance/entities contained (e.g. *tass* 'cup', *kott* 'bag').

The collexemes of the NQC (i.e. nouns used as quantifiers) were identified in our earlier study (Sahkai et al., 2025) on the basis of the EKI Combined Dictionary database and the Estonian Balanced Corpus[1]. A total of 412 nominal quantifiers (NQs) were identified from the dictionary database and 435 NQs were identified from the corpus; the intersection of the two sets is 188 and the total number of unique NQs based on both sources is 654. The use of corpus data made it possible to compute measures of association strength, while the use of dictionary data ensured that the lexemes were linked to sense ID numbers in the EKI Combined Dictionary Database, enabling integration with the dictionary database.

The NQC is a suitable test case as several categorisations of nominal quantifiers have been proposed in the literature. The description of the semantics of measure nouns in Estonian grammar is relatively brief, yet it reflects a layered structure that takes into account both the semantics of the measure noun and the type of entity measured (countable vs. uncountable) (see Metslang et al., 2023, pp. 756–758). The classification presented in (1) therefore applies to quantifier phrases, rather than directly to NQs heading them, but it remains informative with respect to the semantic classes distinguished.

(1)

1. Measure nouns complemented by a mass noun:

    a) primary measure nouns (*liiter* 'litre', *meeter* 'metre', *tonn* 'ton', *tund* 'hour', *tükk* 'piece', *ivake* 'a tiny amount', *korvitäis* 'basketful');

    b) secondary measure nouns that primarily denote a container, aggregate, or object (*klaas* 'glass', *korv* 'basket', *kott* 'bag', *ämber* 'bucket', *hunnik* 'pile'), as well as category nouns (*sort* 'sort', *liik* 'type').

[1] https://www.cl.ut.ee/korpused/grammatikakorpus/ (25 May 2026)

2. Nouns with meanings like 'set' or 'group', complemented by plural count nouns (e.g. *komplekt* 'set', *hulk* 'quantity', *rühm* 'group', *kari* 'herd', *paar* 'pair', *kimp* 'bunch', *virn* 'cluster', *trobikond* 'a small group').

3. Measure nouns designating a part of a whole (*osa* 'part', *enamik* 'majority', *lõik* 'section') or a proportion (*viiendik* 'one fifth'), complemented by a mass or count noun.

More fine-grained classifications are found in the cross-linguistic literature. For example, (Keizer, 2007) proposes five subtypes of pseudo-partitives (see 2), while (Koptjevskaja-Tamm, 2001, p. 503) distinguishes seven non-exhaustive subtypes of measure nouns (see 3). These classifications formed the background for the present study.

(2)

a) Quantifier-noun constructions: *a number of people*

b) Measure-noun constructions: *a pint of beer*

c) Container-noun constructions: *a box of chocolates*

d) Part-noun constructions: *a piece of cake*

e) Collection-noun constructions: *a herd of elephants*

(3)

a) Conventionalised measures: *a litre of milk, a kilo of apples*

b) Abstract quantity nouns: *a large amount of apples*

c) Containers: *a cup of tea, a pail of apples*

d) Fractions/parts: *a slice of bread, a quarter of an hour, a large section of students*

e) Quantums (for mass nouns): *a lump of sugar, a drop of milk*

f) Collections (for count nouns): *a group of students, a herd of sheep*

g) Forms (for both mass and count nouns): *a pile of sand/bricks, a bouquet of roses*

2.2 Data

Two datasets were used in the experiments. Dataset A (N = 188) represents the intersection of the sets of NQC collexemes based on lexicographic and corpus data, with the association strength (LLR) between the noun and the construction ranging from 11,334.886 to 0.016, computed using the software Coll.analysis 4.0 (Gries, 2022). Dataset B contains the full set of NQC collexemes from the Estonian Balanced Corpus (N = 435) that fall within the same range of attraction values.

Dataset A was further enriched with different types of contextual input, presented in Table 1. Together with the plain noun lemmas, there were five input conditions (see sections a–e in the Description column of Table 1).

TABLE 1 Sources and types of input in Dataset A

Source of the input	Description of the input	Input reference
Intersection of corpus and dictionary data	(a) List of collexemes (noun lemmas)	"words"
Balanced Corpus of Estonian	(b) Collexemes with quantifier phrases: three randomly selected expressions per noun lemma (fewer where unavailable)	"phrases"
	(c) Collexemes with sample sentences: three randomly selected sentences per noun (fewer where unavailable)	"sentences"
EKI Combined Dictionary	(d) Collexemes with dictionary definitions; multiple definitions included where relevant (polysemy/homonymy)	"definitions"
	(e) Collexemes with dictionary definitions and usage examples (sentence examples only; compounds excluded)	"usage examples"

2.3 Procedures

The most recent OpenAI's model available at the time (ChatGPT-5.4) was used, with the Extended thinking and Web search options enabled. We selected the OpenAI's model based on its previous iteration's best performance in another NLP task – identification of the instances of the NQC in corpus data (Sahkai et al., 2025). All prompts were issued in Estonian, each in a new, empty chat session.

The procedure consisted of five stages: (i) clustering task using the LLM (2.3.1); (ii) parallel human clustering to establish a benchmark (gold standard) (2.3.2); (iii) evaluation of the LLM outputs (2.3.3); (iv) categorisation task using the LLM and evaluation of the outputs (2.3.4); and (v) scaling up the categorisation task to a larger dataset (2.3.5).

2.3.1 Clustering task

The clustering method employed in the experiment can be described as an informed free sorting task (Chollet et al., 2014). It was *free* in that neither the number nor the composition of the clusters was predefined and *informed* in that neither the LLM nor the human annotators were treated as naïve with respect to the task: both could draw on existing knowledge of semantic classifications of measure nouns described in the literature (see Section 2.1). Clustering was applied to dataset A.

We prompted the model with the following instruction, adjusting for the five different input types (see Table 1):

> You are a lexicographer working on Estonian quantifiers. Based on the [input], identify the most optimal, logical, and meaningful categorisation system. If necessary, consult online sources for comparison with other languages. If needed, use www.sõnaveeb.ee to clarify the meanings of the words.

Input drawn from the EKI Combined Dictionary (i.e., definitions and usage examples) required additional pre-processing to filter out irrelevant senses arising from polysemy and homonymy. For example, *viil* 'slice' also has a homonym meaning 'file', and *päts* 'loaf' has a homonym meaning 'bear'. Such senses were excluded from the analysis by prompting the LLM. Prior to disambiguation, dictionary-derived input exceeded the number of collexemes

approximately fivefold (definitions dataset had 965 instances for 188 quantifiers, usage examples raw dataset had 1281 entries); pre-processing reduced it to a tractable and task-relevant subset. All prompts are provided in the GitHub repository[2].

2.3.2 Creating the gold standard

To evaluate the quality of the LLM-based clusterings, a *tertium comparationis* was established in the form of a gold standard. Two human annotators independently performed a free sorting task analogous to that given to the LLM, grouping the nouns in Dataset A into semantically coherent categories. The annotators worked primarily with noun lemmas but were allowed to consult the EKI Combined Dictionary if needed.

Inter-annotator agreement was substantial (Cohen's $\kappa = 0.7$). Differences between the two clusterings were expected, given the open nature of the task, and primarily concerned the level of granularity (15 vs. 10 clusters) and the treatment of polysemous items, as the protocol did not allow multiple labels per word.

The final gold standard was established through discussion, resulting in a compromise solution of 13 clusters with concise labels. Each cluster was accompanied by a short description clarifying its semantic scope. Table 2 provides an overview of the clusters, presented in descending order of their proportion in Dataset A. The last column provides the average association strength (the log-likelihood ratio, LLR) between the construction and the nouns in each cluster.

TABLE 2 A concise description of the gold standard

No	Label	Description	Examples	Proportion	Mean LLR
1	Container-based quantity	Container nouns functioning as quantifiers of substances or objects	*kann* 'jug', *karp* 'box', *klaas* 'glass'	13.76%	73.95
2	Small-amount quantity	Includes minimal amounts, often conceptualised as drop- or particle-like units	*tilk* 'drop', *kübe* 'speck, particle', *näputäis* 'pinch'	12.70%	28.04
3	Collection	Consists of uniform and unorganised entities; both animate and inanimate	*rühm* 'group', *parv* 'flock', *kogum* 'collection'	10.05%	36.82
4	Bounded quantity	Typically, a quantity delimited by a derivation from a container noun	*kotitäis* 'bagful', *ämbritäis* 'bucketful, *klaasitäis* 'glassful'	10.05%	30.88
5	Part and proportion expressions	Includes both discrete parts and proportional shares	*enamus* 'majority', *tükk* 'piece', *osa* 'part', *käär* 'slice'	9.52%	1311.11
6	Structured collection	Consist of internally related or ordered elements	*rida* 'row', *kimp* 'bundle', *kobar* 'bunch'	7.41%	229.18
7	Standard measurement units	Includes spatial, volumetric, and proportional units such as per mille	*gramm* 'gram', *kraad* 'degree', *liiter* 'litre'	7.41%	277.40

[2] https://github.com/keeleinstituut/PRG1978/tree/main/constructions/EURALEX_2026.

TABLE 2 (*CONTINUED*)

No	Label	Description	Examples	Proportion	Mean LLR
8	Consumable quantity	Nouns denoting a typical single unit of consumption	*lonks* 'sip', *ports* 'serving', *sõõm* 'draught'	6.88%	56.27
9	Units of time	Both standardised and vague	*päev* 'day', *minut* 'minute', *viiv* 'moment'	5.29%	289.95
10	Quantity	A general term that encompasses both very abstract and amorphous quantities	*hulk* 'amount', *kogus* 'quantity', *pahvak* 'burst'	4.76%	1386.97
11	Members of a set	Nouns denoting instances or elements within a set of entities or events	*puhk* 'occasion', *seeria* 'series', *valimik* 'selection, sample'	4.76%	27.51
12	Shape-defined quantity	Includes quantities and aggregates of matter or objects structured by shape	*hunnik* 'pile', *kiht* 'lsquo;layer', *plokk* 'block'	4.23%	69.54
13	Type	General-level category nouns	*kategooria* 'category'; *klass* 'class', *sort* 'kind'	3.17%	7.97

In comparison with the classifications reviewed in 2.1, our 13-category gold standard aligns most closely with Metslang et al. (2023) and Koptjevskaja-Tamm (2001), each matching ten categories fully or partially, while Keizer's (2007) five-way scheme covers only eight. The existing frameworks are complementary: Metslang et al. is the only one to isolate category nouns, and Koptjevskaja-Tamm is the only one to distinguish minimal quantities (Quantums) and shape-based aggregates (Forms). One group – Members of a set (*puhk* 'occasion', *seeria* 'lsquo;series', *valimik* 'lsquo;selection, sample') – has no direct counterpart in the other classifications.

2.3.3 Evaluation metrics

The five clustering outputs of the LLM were evaluated at three levels.

(a) Granularity, i.e., the number of clusters compared with the gold standard.

(b) Structural similarity to the gold standard was measured using Adjusted Rand Index (ARI) and Adjusted Mutual Information (AMI), both of which assess agreement between partitions while correcting for chance and are invariant to label permutation (Hubert & Arabie, 1985).

(c) Label quality was evaluated by human annotators using a three-point scale: 1 (correct = the label coincided with a label in the gold standard), 0.5 (partially correct = the label did not coincide with a label in the gold standard, but was an acceptable alternative), and 0 (incorrect = the label did not coincide with a label in the gold standard and was not an acceptable alternative). An analogous prompt was also used to test whether this evaluation step could be automated by the LLM.

2.3.4 Categorisation task and evaluation

Categorisation was performed in the form of a closed sorting task (Divjak et al., 2024), using dataset A, the five input types and the gold standard classification. The LLM was prompted using the following instruction:

> You are a lexicographer working on Estonian quantifiers. Task: distribute [the input, see Table 1] provided in the file in the following clusters [list of clusters in the form of gold standard labels and descriptions, see Table 2]

The results were evaluated semi-manually. In a first step, we used the LLM to compare the category assignments with the gold standard (1 = correct, 0 = incorrect). In a second step, mismatches were reviewed by a human expert. In justified cases (i.e. where alternative affiliations were acceptable), scores were adjusted from 0 to 0.5, following the scoring protocol described in Section 2.3.3.

2.3.5 Scaling up the categorisation

In the final stage, the closed sorting task was applied to a larger dataset (Dataset B; N = 435), using the same prompt as in the initial closed sorting task (see 2.3.4). The categorisation of the Dataset A nouns was evaluated automatically by applying the measures of micro and macro level accuracy. Micro-averaged accuracy measures the proportion of correctly categorised lemmas across the whole dataset, giving each lemma equal weight. Macro-averaged accuracy first calculates accuracy separately for each category and then averages these category-level scores, giving each category equal weight regardless of its size. The categorisation of the novel words not included in the gold standard was evaluated by two human experts, using again the evaluation system described in 2.3.3.

3 Results

In this section, we report the results and address the research questions.

3.1 Clustering task

An overview of the results is presented in Table 3. The number of clusters shows a high degree of convergence across the five input types (mean: 12.4; median: 13), closely matching the gold standard (N = 13). Two input types yielded exactly 13 clusters. However, the number of clusters alone is not informative with respect to how they were assigned to the items in Dataset A.

TABLE 3 Performance of the LLM in the informed free sorting task across five input types, compared to the gold standard

Input reference	No of clusters	AMI	ARI	Labelling (human)	Labelling (LLM)
"words"	**13/13***	0.634	0.522	0.749	0.701
"phrases"	10/13*	0.492	0.373	0.706	0.635
"sentences"	14/13*	0.652	0.514	**0.802**	**0.831**
"definitions"	12/13*	**0.656**	**0.569**	0.738	0.746
"usage examples"	**13/13***	0.613	0.445	0.696	0.783

Note: The asterisk (*) marks the number of clusters in the gold standard; values closest to the gold standard are highlighted in bold.

The AMI and ARI scores indicate that the clustering based on dictionary definitions showed the greatest similarity to the gold standard. The other input conditions also performed well, with the exception of "phrases", which yielded the lowest agreement in terms of pairwise comparison. Interestingly, definition-only input outperformed definition-plus-examples input, suggesting that additional context may introduce noise rather than improve clustering quality.

In contrast, the evaluation of label quality favoured the "sentences" condition. Overall, however, differences across input types were relatively small: all five conditions produced labels with approximately 70–80% overlap with the gold standard. It should be noted that the holistic evaluation of label quality involves a degree of subjectivity, particularly given that the gold standard itself reflects a negotiated compromise. Human evaluators often found the LLM-generated labels semantically plausible and well-formed, even if they did not fully coincide with the gold-standard labels. Notably, human and LLM-based evaluations showed a high degree (~80–83%) of agreement regarding label adequacy.

A further implication concerns the role of polysemy, which introduces systematic variability into clustering behaviour. A substantial proportion of items can be plausibly assigned to more than one category (e.g. *toop* 'jug; pint' denotes both a container and a unit, and *promill* 'per mille' denotes both a unit and a proportion), suggesting that cluster membership should be treated as one-to-many rather than one-to-one.

3.2 Categorisation task

Table 4 reports the accuracy of the closed sorting as agreement with the gold standard. The model achieved an average weighted accuracy of 85.6%.

TABLE 4 Coincidence with the gold standard in the closed sorting task across the five input types

Input reference	Weighted accuracy	Fully correct rate	At-least-partially correct rate	Mismatch rate	Cohen's κ
"words"	**90.2%**	**83.1%**	**97.4%**	**2.7%**	**0.814**
"phrases"	78.0%	70.4%	85.7%	14.3%	0.674
"sentences"	84.2%	76.2%	92.6%	7.4%	0.738
"definitions"	87.3%	78.3%	96.3%	3.7%	0.762
"usage examples"	88.1%	79.9%	96.3%	3.7%	0.779

The bare-lemma input condition "words" outperformed the others, yielding weighted accuracy of up to 90% and the highest Cohen's kappa (0.8). Again, phrase-based input produced weaker results, presumably due to greater noise and/or increased polysemy. These findings indicate that, in a closed sorting setting, additional contextual information in the input does not improve performance. Once the semantic structure of the collexeme set is specified, the LLM is able to assign items with high accuracy.

Systematic comparison further reveals variation across gold-standard clusters in how readily their instances were recognised, which likely reflects differences in semantic distinctiveness. The most accurately classified clusters (>93%) were 'small-amount quantity', 'units of time', and 'bounded quantity'. In contrast, the most challenging cluster was 'members of a set', with only 37.8% weighted accuracy and a 51% zero-score rate; particularly difficult items included

ring 'circle', *seeria* 'series', and *valimik* 'selection'. This cluster was the only one of our gold standard for which no precedent existed in the literature.

3.3 Scaling up the categorisation

The gold standard categorisation system established on the basis of Dataset A (N = 188) was applied to Dataset B (N = 435), using the input type that performed best in the closed sorting task, i.e., "words". The results are shown in Table 5.

TABLE 5 Performance of the LLM on the scaling up experiment

Group	N	Reference	Main measure	Result
Gold-standard collexemes	188	Expert gold standard	Micro accuracy	79.8%
Gold-standard collexemes	188	Expert gold standard	Macro accuracy	76.3%
Novel collexemes	247	Human evaluation	Mean rating of human experts	0.948
Novel collexemes	247	Human evaluation	Fully acceptable assignments	89.5%

The scaling experiment indicated that the 13-category gold standard derived from Dataset A could be applied to the larger Dataset B with substantial success. For the 188 lemmas overlapping with the gold standard, the LLM achieved 79.8% micro-averaged accuracy and 76.3% macro-averaged accuracy, showing that the expert classification was reproduced in approximately four out of five cases. The slightly lower macro accuracy indicates that performance was not fully uniform across categories: the LLM performed well in larger and semantically well-delimited categories, such as container-based quantity, standard measurement units, small-amount quantity, and units of time, but less consistently in smaller or fuzzier ones, especially quantity, member of a set, shape-defined quantity, and structured collection.

For the 247 novel lemmas, human evaluation of the LLM's assignments was even more favourable than in the case of the initial categorisation task reported in 3.2: the mean rating was 0.948, and 89.5% of the assignments were judged fully correct.

3.4 Answers to the research questions

A positive answer can be given to RQ1. The evaluation results suggest that ChatGPT-5.4 can be successfully used for semantic clustering of collexemes, with overlap between human and machine clustering as high as 0.7 (AMI); 0.6 (ARI) and labelling quality reaching 80%.

A positive answer can also be given to RQ2. In the closed sorting task, the LLM assigned collexemes to predefined categories with substantial agreement with the human gold standard, reaching an average weighted accuracy of 85.6%. The best result was obtained with bare lemmas ("words": 90.2%; $\kappa = 0.814$), suggesting that once the category system is specified, additional contextual input does not improve performance. However, accuracy varies across categories: semantically clear categories are assigned reliably, while diffuse categories such as 'members of a set' are more challenging.

Regarding RQ3, the informed free sorting task did not yield a fully decisive answer, as four of the five input types produced broadly comparable outcomes. The highest pairwise agreement with the gold standard was achieved using dictionary definitions (AMI = 0.656, ARI = 0.569),

whereas the highest label-quality scores were obtained with sentence-based input (human-rated 0.802; LLM-rated 0.831). The closed sorting task revealed that the input condition of bare lemmas ("words") was able to provide the highest accuracy and agreement with the gold standard. Taken together, these findings suggest that dictionary definitions provide the most reliable basis for free clustering, while sentence-level corpus context yields the most natural and human-aligned labels, bare lemmas being sufficient for the closed categorisation task.

The scaling experiment that addressed RQ4 showed that the gold-standard categorisation could be applied to a larger dataset with substantial success: for the overlapping gold-standard lemmas, the LLM reached 79.8% accuracy, while 89.5% of novel lemma assignments were judged fully acceptable by human raters. However, performance was weaker for semantically diffuse or borderline categories, such as 'quantity' and 'member of a set'. This suggests that when a pre-defined classification is used, attention should be paid to the clear description of the categories.

4 Discussion and conclusions

The study extends the tradition of interpretive (e.g., Barðdal, 2008; Goldberg, 1995; Levshina, 2016) and data-driven (e.g. Gries & Stefanowitsch, 2010; Ellis et al., 2014; Perak & Ban Kirigin, 2023; Chen, 2022, 2025) categorisation in Construction Grammar, where classes of collexemes are used to infer constructional meaning. This addresses a key challenge in constructicography: how to complement statistical organisation of collo-profiles with semantic structure.

The findings suggest that LLMs can approximate both interpretive and data-driven grouping with minimal manual effort. Compared with previous automated categorisation methods in terms of distributional and similarity-based approaches, LLM-based clustering has the practical advantage of combining grouping and semantic interpretation in one step. This addresses the problem identified by Ziem and Feldmüller (2023): how to organise large collo-profiles in a way that is both empirically grounded and semantically meaningful, and, at the same time, scalable.

However, clustering and categorisation tasks are not fully determinate, as several categorisations may be valid, reflecting the gradient and multidimensional nature of lexical semantics. Some collexemes can plausibly belong to more than one category, suggesting that constructicon entries should allow flexible or multi-cluster representations rather than strictly discrete classifications. This complicated also the evaluation of the results. As a solution, we used a combined evaluation of structural and interpretative quality and allowed for partial acceptability. Partition-similarity measures such as AMI and ARI assessed the organisation of collexemes, while label-quality assessment evaluated the acceptability of the resulting semantic descriptions. This distinction is important because constructicographic clusters must be not only structurally plausible but also interpretable for users.

The comparison of input conditions showed that different inputs support different stages of the workflow. Collexemes supplied with dictionary definitions produced the most reliable cluster structures. This could be explained by the fact that definitions contain structured semantic information, similarly to frame-semantic approaches (Sundquist, 2020; Hartmann & Willich, 2024). Collexemes supplied with corpus sentences supported more satisfying cluster labels, while bare lemmas of collexemes were sufficient for assigning items to an already established category system. Thus, richer context does not necessarily improve performance; in some cases, it introduces noise.

In conclusion, for the purpose of creating a system of labelled categories, collexemes accompanied by dictionary definitions provide the best input to the LLM, also allowing to solve the ambiguity created by lexical polysemy and homonymy. For better controllability, a human-in-the-loop approach could be applied, where LLMs are used to generate an initial category system based on a smaller dataset that can be validated by an expert. This category system can then be applied to a larger dataset. In case of the existence of a predefined category system, bare lemmas provide a sufficient input for the model to assign collexemes into categories. However, to achieve greater reliability, human validation is necessary to identify borderline and ambiguous cases.

In sum, LLM-assisted semantic clustering and categorisation offers a scalable method for structuring constructional collo-profiles. In the Estonian Constructicon, such structuring can be combined with other criteria used to structure the collo-profile – frequency, association strength and CEFR-level information –, allowing to sort and filter the collexemes based on multiple criteria. This study is limited to one model, one language and one construction type, and the gold standard represents one plausible human consensus rather than an objective ground truth. Further work should test the workflow across models, languages and constructions.

References

Barðdal, J. (2008). *Productivity: Evidence from case and argument structure in Icelandic*. John Benjamins. https://doi.org/10.1075/cal.8

Chen, A. C.-H. (2022). Words, constructions and corpora: Network representations of constructional semantics for Mandarin space particles. *Corpus Linguistics and Linguistic Theory*, 18(2), 209–235. https://doi.org/10.1515/cllt-2020-0012

Chen, A. C.-H. (2025). From sequentiality to schematization: Network-based analysis of covarying collexemes in Mandarin degree adverb constructions. *Corpus Linguistics and Linguistic Theory*, *21*(3), 475–515. https://doi.org/10.1515/cllt-2024-0037

Chollet, S., Valentin, D. & Abdi, H. (2014). Free Sorting Task. In P. Varela Tomasco & G. Ares (Eds.), *Novel techniques in sensory characterization and consumer profiling* (pp. 207–223). CRC Press.

Dekalo, V. & Hampe, B. (2017). Networks of meanings: Complementing collostructional analysis by cluster and network analyses. *Yearbook of the German Cognitive Linguistics Association*, *5*(1). https://doi.org/10.1515/gcla-2017-0011

Divjak, D., He, S. & Borowski, M. (2024). *Expertise affects categorisation, in language too: (The) case (of) constructions (preprint)*. https://pure-oai.bham.ac.uk/ws/portalfiles/portal/272333420/PURE_Polish Case_Revised.pdf

Ellis, N. C., O'Donnell, M. B. & Römer, U. (2014). The processing of verb-argument constructions is sensitive to form, function, frequency, contingency and prototypicality. *Cognitive Linguistics*, *25*(1), 55–98. https://doi.org/10.1515/cog-2013-0031

Fillmore, C. J., Baker, C. F. & Sato, H. (2002). The FrameNet Database and Software Tools. *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC'02), Las Palmas, Canary Islands - Spain. European Language Resources Association (ELRA), 1156–1160*.

Goldberg, A. E. (1995). *Constructions: A Construction Grammar Approach to Argument Structure*. University of Chicago Press. https://press.uchicago.edu/ucp/books/book/chicago/C/bo3683810.html

Gries, S. Th. (2022). *Coll.analysis 4.0: A script for R to compute collostructional analyses*. https://www.stgries.info/teaching/groningen/index.html

Gries, S. Th. & Stefanowitsch, A. (2010). Cluster analysis and the identification of collexeme classes. In S. Rice & J. Newman (Eds.), *Empirical and experimental methods in cognitive/functional research* (pp. 73–89). CSLI Publications/Center for the Study of Language and Information. Conceptual

Structure, Discourse and Language Conference, Stanford, Calif. https://www.stgries.info/research/2010_STG-AS_ClusteringCollexemes_EmpExpMeth.pdf

Hartmann, S. & Willich, A. (2024). Collostructional analysis meets construction semantics: Revisiting the English way-construction and its German equivalents. *Zeitschrift Für Anglistik Und Amerikanistik*, 72(3), 319–349. https://doi.org/10.1515/zaa-2024-2029

Herbst, T. (2018). Is Language a Collostructicon? A Proposal for Looking at Collocations, Valency, Argument Structure and Other Constructions. In P. Cantos-Gómez & M. Almela-Sánchez (Eds.), *Lexical Collocation Analysis* (pp. 1–22). Springer International Publishing. https://doi.org/10.1007/978-3-319-92582-0_1

Herbst, T. (2020). Constructions, generalizations, and the unpredictability of language: Moving towards colloconstruction grammar. *Constructions and Frames*, 12(1), 56–95. https://doi.org/10.1075/cf.00035.her

Herbst, T. & Hoffmann, T. (2024). *A Construction Grammar of the English Language: CASA – a Constructionist Approach to Syntactic Analysis* (Vol. 5). John Benjamins Publishing Company. https://doi.org/10.1075/clip.5

Hubert, L. & Arabie, P. (1985). Comparing partitions. *Journal of Classification*, 2(1), 193–218. https://doi.org/10.1007/BF01908075

Janda, L. A., Endresen, A., Zhukova, V., Mordashova, D. & Rakhilina, E. (2020). How to build a constructicon in five years. The Russian example. *Belgian Journal of Linguistics*, 34(1), 161–173. https://doi.org/10.1075/bjl.00043.jan

Kallas, J., Kiil, A., Sahkai, H., Paulsen, G. & Saul, K. (2026). Using LLMs to extract instances of schematic constructions from unannotated L2 learner corpora. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference* (LREC 2026) (pp. 10517–10524). ELRA Language Resources Association (ELRA).

Kallas, J., Paulsen, G. & Saul, K. (2025). Language proficiency levels for constructions. *Pedagogical Application of Constructicography Workshop 2025*. https://static1.squarespace.com/static/5bc797b1523958aa328c4f66/t/682b5e0c56c18d2b941f4254/1747672591962/Estonian+constructicon_Language+proficiency+levels+for+constructions_G%C3%B6teborg+2025.pdf

Keizer, E. (2007). *The English noun phrase: The nature of linguistic categorization*. Cambridge University Press.

Koppel, K., Tavast, A., Langemets, M. & Kallas, J. (2019). Aggregating dictionaries into the language portal Sõnaveeb: Issues with and without a solution. In I. Kosem, T. Zingano Kuhn, M. Correia, J. P. Ferreira, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (Eds.), *Proceedings of the eLex 2019 conference* (pp. 434–452). Lexical Computing CZ, s.r.o. https://elex.link/elex2019/wp-content/uploads/2019/09/eLex_2019_24.pdf

Koptjevskaja-Tamm, M. (2001). Partitive and pseudo-partitive nominal constructions in the Circum-Baltic languages: "A piece of the cake" and "a cup of tea". In Ö. Dahl & M. Koptjevskaja-Tamm (Eds.), *Circum-Baltic Languages: Volume 2: Grammar and Typology* (pp. 523–568). John Benjamins Publishing Company. https://doi.org/10.1075/slcs.55.11kop

Levshina, N. (2016). A geometric exemplar-based model of semantic structure: The Dutch causative construction with *laten*. In J. Yoon & S. Th. Gries (Eds.), *Constructional Approaches to Language* (Vol. 19, pp. 241–262). John Benjamins Publishing Company. https://doi.org/10.1075/cal.19.09lev

Metslang, H. (2017). Kvantorifraas. In M. Erelt & H. Metslang (Eds.), *Eesti keele süntaks* (pp. 463–478). Tartu Ülikooli Kirjastus.

Metslang, H., Erelt, M., Habicht, K., Hennoste, T., Kasik, R., Teras, P., Viht, A., Asu, E. L., Lindström, L., Lippus, P., Pajusalu, R., Plado, H., Rääbis, A. & Veismann, A. (2023). *Eesti grammatika*. Tartu Ülikooli Kirjastus. https://doi.org/10.12697/EG

Perak, B. & Ban Kirigin, T. (2023). Construction Grammar Conceptual Network: Coordination-based graph method for semantic association analysis. *Natural Language Engineering*, 29(3), 584–614. https://doi.org/10.1017/S1351324922000274

Perek, F. (2016). Using distributional semantics to study syntactic productivity in diachrony: A case study. *Linguistics*, 54(1), 149–188. https://doi.org/10.1515/ling-2015-0043

Perek, F. (2018). Recent change in the productivity and schematicity of the way-construction: A distributional semantic analysis. *Corpus Linguistics and Linguistic Theory*, 14(1), 65–97. https://doi.org/10.1515/cllt-2016-0014

Pilvik, M.-L., Lindström, L., Plado, H. & Simmul, C. E. (2025). Nimisõnafraasi ja hulgafraasi piirimail: 'osa', 'enamik' ja 'enamus' hulgasõnadena. *Eesti Rakenduslingvistika Ühingu Aastaraamat*, 21(0), Article 0. https://doi.org/10.5128/ERYa21.13

Sahkai, H., Paulsen, G., Vainik, E., Kallas, J., Kiil, A., Tsepelina, K., Saul, K. & Tavast, A. (2025). Exploring the constructicographic potential of lexicographic data and language models: The case of the Estonian Nominal Quantifier Construction. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda & A. Salgado (Eds.), *Electronic lexicography in the 21st century (eLex 2025): Intelligent lexicography. Proceedings of the eLex 2025 conference* (pp. 673–691). Lexical Computing CZ s.r.o.

Sundquist, J. D. (2020). Productivity, richness, and diversity of light verb constructions in the history of American English. *Journal of Historical Linguistics*, 10(3), 349–388. https://doi.org/10.1075/jhl.19009.sun

Stefanowitsch, A. & Gries, S. Th. (2003). Collostructions: Investigating the interaction of words and constructions. *International Journal of Corpus Linguistics*, 8(2), 209–243

Tavast, A., Langemets, M., Kallas, J. & Koppel, K. (2018). Unified Data Modelling for Presenting Lexical Data: The Case of EKILEX. In *Proceedings of the XVIII EURALEX International Congress: EURALEX: Lexicography in Global Contexts* (pp. 749–761). Ljubljana University Press, Faculty of Arts.

Vainik, E., Paulsen, G., Sahkai, H., Kallas, J., Tavast, A., & Koppel, K. (2024). From a dictionary to a constructicon – Putting the basics on the map. In K. Š. Despot, A. Ostroški Anić & I. Brač (Eds.), Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress (pp. 209–217). Institute for the Croatian Language.

Zhukova, V. Z. (2026). *A Constructivist Approach to Russian Grammar: Lessons from the Russian Constructicon* [Doctoral thesis, UiT The Arctic University of Norway, Universitetet i Tromsø]. https://nva.sikt.no/registration/019c8ea8772a-098f7547-a914-4323-b0b0-b908239dd408

Ziem, A. & Feldmüller, T. (2023). Dimensions of constructional meanings in the German Constructicon: Why collo-profiles matter. *Yearbook of the German Cognitive Linguistics Association*, 11(1), 203–226. https://doi.org/10.1515/gcla-2023-0010

Ziem, A., Flick, J. & Sandkühler, P. (2019). The German Constructicon Project: Framework, methodology, resources. *Lexicographica*, 35(2019), 15–40. https://doi.org/10.1515/lex-2019-0003

Contact information

Ene Vainik
The Institute of the Estonian Language
ene.vainik@eki.ee

Heete Sahkai
The Institute of the Estonian Language
heete.sahkai@eki.ee

Ahto Kiil
The Institute of the Estonian Language
ahto.kiil@eki.ee

Jelena Kallas
The Institute of the Estonian Language
jelena.kallas@eki.ee

Geda Paulsen
The Institute of the Estonian Language
Uppsala University
geda.paulsen@eki.ee, geda.paulsen@moderna.uu.se

Acknowledgements

This work was supported by the Estonian Research Council grant (PRG 1978).

Zongyuan Wang, Valeria Caruso & Anna De Meo

COLLABORATIVE LEXICOGRAPHY FOR HERITAGE LANGUAGE LEARNING

Abstract This study explores the potential of collaborative dictionary compilation to promote vocabulary retention and contributes to research in Chinese heritage language education. A pilot intervention was conducted with 22 second-generation Chinese students attending a complementary school in Italy. The participants studied literary vocabulary from Lu Xun's *Village Opera* and collaboratively compiled dictionary entries using Lexonomy. To assess the broader effectiveness of the protocol and compensate for the absence of a control group of heritage learners in Italy, the framework was replicated with two age-matched groups in China: an experimental group of 50 students engaged in dictionary-compilation activities and a control group of 53 students receiving traditional teacher-led instruction. Vocabulary knowledge was assessed through a pre-test and a delayed post-test administered three months after the intervention. A mixed-effects logistic regression analysis showed that the experimental group was approximately twice as likely as the control group to improve its retention of the literary vocabulary studied. The findings provide preliminary evidence that collaborative lexicography can support medium-term literary vocabulary acquisition and culturally oriented language learning, although larger-scale studies involving students from complementary schools are needed.

Keywords Heritage language learning; collaborative lexicography; Chinese as a heritage language; vocabulary acquisition; dictionary compilation; lexonomy; literacy development

1 Preliminary Remarks on Dictionaries and Societies

This paper explores the role that lexicography can play in societies and among specific groups of people whose lives and social engagement may be improved through the use of dictionaries. If lexicography is to become a discipline more closely concerned with people and real-world contexts, it may have to sacrifice some degree of precision, run the risk of relying on unbalanced data, and even produce partial or negative evidence. However provisional or even disappointing a real-world context may appear because of its imbalanced components, it should not be disregarded but rather used as a starting point for further improvement, accumulating evidence, replication studies, and clearer demonstrations of the role dictionaries play in the lives of their users.

The current study addresses a complex global context, populated by migrant communities with strong cultural identity that are rapidly evolving in relation both to their host societies and to their cultures of origin. Despite the common features that newly established Chinese communities abroad may share, and the Chinese state's "resilient capacity" to facilitate "diasporic participation in China's socio-economic and political processes" (Liu, 2011, p. 813), the Chinese diaspora has been described as "one of the most hyperdiverse groups, with substantial subgroup differences that distinguish it from most other diasporas" (Guo, 2022, p. 847).

Against this social background, Chinese complementary schools[1], which are found across the world, play a particularly important role. Their social mission varies from country to country,

[1] A complementary school is a community-based organisation, often established by parents, that provides language and cultural education outside regular school hours to maintain a community's linguistic and cultural heritage. Such schools are a widespread and distinctive feature of the Chinese diaspora (Curdt-Christiansen & Hancock, 2014). These schools are also referred to as *heritage language schools, community language schools*, and *supplementary schools*. It is worth noting that a UK government report commissioned by the Department for Children, Schools and Families uses the term *supplementary schools* (Maylor at al., 2010).

depending largely on the characteristics of the host society and of the migrant community itself. In Italy, these schools began to appear in the second half of the 1990s and, because they are self-governed and are often established through the initiatives of private members of the immigrant community –such as entrepreneurs, they have developed without coherent planning, display highly diverse characteristics, and still lack a comprehensive census (Pellin, 2025), let alone a clear overview of the teaching practices adopted in each institution.

With respect to the Italian context, Brigadoi Cologna (2023) notes the tendency to promote a standardised, supranational culture centred on Pǔtōnghuà (普通话), with the support of the PRC government, and the difficulties that this may create within migrant communities whose members predominantly speak Wu varieties of Sinitic, which belong to a branch distinct from Mandarin.

Setting aside the complex issue of which language should be taught in complementary schools, it should be emphasised that conducting coherent and structured research within these institutions is as complex as it is desirable. Their geographical dispersion and the absence of coordinating bodies[2] make it difficult to establish collaborations and carry out research involving their students. Such research is nevertheless highly desirable, since support for the development of young learners' L1 is recognised in the relevant literature as a key factor in the educational success of students from immigrant backgrounds (see, at least, Cummins, 1979; Thomas & Collier, 2002).

However, given the current organisation of Chinese complementary schools in Italy and the objective difficulty of recruiting large numbers of learners, the present study should be regarded as a starting point for investigating the benefits of integrating innovative dictionary-based programmes into this highly specific educational context. It provides preliminary evidence of both the effectiveness of the learning protocol and its feasibility for second-generation Chinese students attending these schools across Italy. Nevertheless, larger-scale studies involving more students and a greater number of complementary schools are still needed.

2 Introduction

In the teaching of Chinese as a heritage language, particular importance is attached to the systematic integration of cultural and community dimensions into language instruction (He, 2010; Carreira & Kagan, 2011). Building on this premise, the present study proposes a pedagogical approach that combines digital lexicographic technologies with the study of Chinese literature. Its aim is to strengthen heritage language learners' connection with the culture of their country of origin while improving their command of different language registers, given that their linguistic repertoires are often unevenly distributed across domains and varieties (Carreira, 2004; Valdés, 2001). To implement this dictionary-compilation framework, which transforms the dictionary from a source of information into a repository of structured lexical data compiled by the learners themselves, the study addresses two interrelated research questions: whether collaborative dictionary compilation is an effective strategy for learning new vocabulary, and whether this approach can also be successfully implemented in more complex social settings, such as heritage language education in a diasporic context.

[2] Two recent initiatives are seeking to provide a clearer picture of Chinese complementary education in Italy, thereby addressing the limited attention paid by Italian institutions to migrant-community multilingualism and to the educational success of second-generation students. These are the PRIN 2022 PNRR project ECCO-Italy – Education of Children of Chinese Origin in Italy (Pellin, 2025) and the Observatory on Italian Heritage Schools currently being developed by the Centre for Research on Minorities (CERM) at the University of Insubria (Brigadoi Cologna, 2023).

A pilot study was conducted with a group of second-generation students of Chinese origin (Heritage Group, HG) attending a Chinese complementary school in Italy. The experiment aimed to assess the overall design of the teaching protocol and the types of activities to be carried out alongside dictionary compilation. Regrettably, the limited number of participants made it impossible to establish a control group, while the lack of computers for compiling the learners' dictionary prevented the study from meeting all the conditions required for a fully robust experimental design.

The participants attended a Chinese complementary school in Italy that offers extracurricular classes alongside mainstream public education, also preparing learners for the HSK language certification (Hànyǔ Shuǐpíng Kǎoshì, 汉语水平考试). The participants, who were enrolled in Grades 7 and 8 of the Chinese school, had only recently begun studying literary texts, thereby minimising the risk of bias arising from prior knowledge of the vocabulary selected for testing.

Within the Chinese education system, literary education—including the reading and appreciation of classical and modern poetry, prose, fiction, and other culturally significant texts—is regarded as part of the broader development of linguistic competence, cultural awareness, and students' core competencies. This conception is also reflected in the terminology used within the nine-year compulsory education system. Literature is not taught as a separate subject but is integrated into Yuwen (语文), the national curriculum area conventionally rendered in English as "Chinese Language" or "Chinese Language and Literature"[3].

Within this educational setting, a module on Chinese language and culture based on literary reading was piloted, adopting a lexical approach and a lexicographic methodology that frames dictionary compilation as an experiential and cooperative learning practice rather than as pure consultative activity (Campoy Cubillo, 2002; Caruso, 2024; Caruso & De Meo, 2025; Chaves-Yuste & Márquez-Cruz, 2022, 2024).

To alleviate the practical difficulty of establishing a control group for the Chinese heritage learners in Italy, and to assess whether the Heritage Group's vocabulary knowledge was comparable to that of age-matched students in China, the same framework was replicated with two groups of learners in China. One group followed the same activities implemented in Italy (Experimental Group, EG), while the other received more teacher-centred instruction based on textbook activities (Control Group, CG).

Results from a delayed post-test administered three months after the classes point to the efficacy of the protocol among Chinese learners, with the experimental group showing significantly greater gains over time. These findings encourage further investigation into the efficacy of the protocol also among heritage learners in Italy.

3 The learning framework

3.1 Collaborative lexicography for learning

Building on and adapting the protocol proposed in Caruso (2024), an experimental group of 22 Chinese heritage language learners (HG, mean age 12.7) read and analysed the short story *Village Opera* (社戏; *Shèxì*) by Lu Xun (1922), a canonical text in school curricula for the

[3] The relevant national framework is the *Compulsory Education Curriculum Plan and Curriculum Standards (2022 Edition)*, issued by the Ministry of Education of the People's Republic of China and implemented from the autumn semester of 2022. See http://en.moe.gov.cn/news/press_releases/202205/t20220507_625532.html.

transmission of modern Chinese literary language and cultural identity (He, 2006; Curdt-Christiansen & Hancock, 2014). The instructional intervention pursued a dual objective: on the one hand, to improve learners' linguistic and cultural competences through the acquisition of literary vocabulary; on the other hand, to foster the development of effective learning strategies for vocabulary acquisition and Chinese character literacy (He, 2020).

The experimental intervention focused on the learning of 15 target words over a ten-hour instructional period. The course combined teacher-led activities devoted to guided reading and textual analysis with laboratory-based work designed to familiarize learners with the use of Lexonomy (Mechura, 2017; Jakubíček et al., 2018, 2022). Through lexicographic training and collaborative dictionary compilation, students were expected to develop greater awareness of structured lexical description as a means of facilitating vocabulary retention, while also benefiting from peer-supported use of the digital tool (Elia, 2009; Caruso & De Meo, 2025). Target words were selected based on their frequency in contemporary Chinese, as determined through the zhTenTen21 corpus (Kilgarriff et al., 2014). Only the highest-frequency words were included among the target items for assessment, although learners were free to choose any lemma when compiling entries for their dictionary.

In line with Papert's constructivist pedagogy (Papert, 1980; Papert & Harel, 1991), widely adopted in educational technology, students were encouraged to explore Lu Xun's short story, *Village Opera* (社戏; *Shèxì*), autonomously and to select, working in small groups, the lemmas to be included in their dictionary. Papert's learning model, grounded in Piagetian learning theory, emphasizes the active involvement of learners in constructing knowledge through experiential activities and task-based learning. Within this framework, technology is effective when it supports learners' active participation and helps them create meaningful products. The collaborative compilation of digital dictionary entries was therefore expected to promote vocabulary growth and encourage learners to develop new ways of practicing unfamiliar words, such as searching for example sentences online and providing Italian translations.

3.2 Participants and Group Characteristics

Due to infrastructural constraints, the 22 heritage students could only use tablets to carry out the lexicographic activities, and some insisted on using their own smartphones instead of the devices provided by the Chinese school in Italy. This represented a significant limitation for the learning framework, since Lexonomy has been designed for larger PC screens, particularly with respect to dictionary compilation tasks. Only recently has the tool become more suitable for smartphone-based consultation and not for its compilation. A relevant finding of the pilot study involving the HG concerned the duration of the instructional intervention: the initially estimated ten hours proved to be excessive.

To address the difficulty of establishing a control group of Chinese heritage learners in Italy, and to assess the knowledge of the literary vocabulary targeted in the study among Chinese students of the same age and school grade, the same instructional framework was replicated with two groups of learners in China. Their results showed that students in this age group were also largely unfamiliar with the language of Lu Xun's short story.

To evaluate the effectiveness of the dictionary compilation protocol, a control group of 53 Chinese students (Control Group, CG) received traditional instruction based on teacher-led explanations and textbook activities. A second group of 50 students from the same school (Experimental Group, EG) followed the same protocol as the heritage learners in Italy but used computers to compile dictionary entries with Lexonomy. The instructional module

implemented with the Chinese students, in both the experimental and control groups, lasted half as long as the one conducted in Italy.

The control group did not engage in any lexicographic activities. Instead, students completed textbook exercises and received more detailed teacher explanations of the text. The textbook activities also had a lexical focus: students were asked to summarize episodes from the story using only four characters, and to reflect on the different narrative techniques employed by the author, including descriptions, the expressions of feelings, and his reflections.

To evaluate the efficacy of the experimental framework, all groups completed an initial cloze test on the 15 target words, which required them to insert the words into sentences drawn from Lu Xun's literary text (literary context) and from other usage examples extracted from the zhTenTen21 corpus (non-literary context). All tests were administered online via Google Forms. Except for the HG, the number of students completing the cloze post-test was lower than the original group size: 37 out of 50 in the experimental group and 50 out of 53 in the control group. Students in the HG and EG also completed an acceptance questionnaire designed to assess their satisfaction with the learning framework, the group work, and the Lexonomy dictionary editor.

TABLE 1 Participant group composition and activities undertaken

Student Type	School Year	Participants in the pre-test	Participants in the delayed post-test	Gender Distribution	Mean Age	Activities
Heritage students	Seventh and eighth school year	22	22	15 females, 7 males	12.7	10 meetings (500 minutes); collaborative dictionary compilation
Experimental Group	Seventh school year	50	37	24 females, 26 males	13.3	5 meetings (235 minutes); collaborative dictionary compilation
Control Group	Seventh school year	53	50	27 females, 26 males	13	5 meetings (235 minutes); predominantly teacher-led activities

3.3 *Learning objectives and Assessment of Students' Vocabulary Knowledge*

The teaching protocol implemented in the experimental and heritage group was designed as an interdisciplinary learning experience combining cultural and literary knowledge within a lexical approach supported by state-of-the-art lexicographic technologies. The project aimed to enhance students' lexical competence through guided exposure to Chinese literary vocabulary and the compilation of lexicographic entries including definitions, usage examples, collocations, and translations. Within this framework, vocabulary was conceived not merely as a set of linguistic units to be learned, but also as a means of accessing cultural and metalinguistic knowledge through reflection on the types of lexical information included in dictionary entries.

By way of example, the entries compiled by the **Heritage** and **Chinese** students (EG) for the same lemma—朦胧 (*ménglóng*, 'hazy', 'dim', or 'blurred')—can be compared, while the total number of entries compiled by each group is summarised in Table 2.

The number of students who contributed to the dictionary compilation differs from the number of students who participated in the tests. In the Heritage Group, 27 students took part in the instructional intervention and contributed to the collaborative dictionary, but only 22 completed both the pre-test and the delayed post-test. In the Experimental Group, all 50 students participated in the instructional intervention; however, only 47 entered their names when compiling the dictionary entries, and only 37 completed both the pre-test and the delayed post-test. Consequently, only the performance of these 37 students was included in the statistical analysis. The number of compiled dictionary entries is therefore reported on the basis of the students who actually contributed to the dictionary compilation.

TABLE 2 Composition of the groups and number of dictionary entries compiled

Group	Number of groups compiling entries	Average group size	Number of entries compiled	Average number of entries per group
Heritage	10	2.7	33	3.3
Experimental	14	3.6	29	2.1

From a cultural perspective, the protocol sought to introduce students to the culture of *Shèxì* (or 'village opera', 社戏)[4] and, more broadly, to the heritage of traditional Chinese opera, with particular attention to Peking Opera (*Jīngjù* 京剧)[5]. Lu Xun's short story *Village Opera* (社戏; *Shèxì*) describes the young protagonist's experience of attending a traditional opera performance in rural southern China, where stages were set up along the riverbanks and spectators watched the performances from boats on the water. During classroom activities, students were introduced to this theatrical tradition through audiovisual materials that enabled them to explore its main character types, the symbolic significance of masks, and the relationship between theatre, collective ritual practices, and cultural memory.

In this context, the literary text and its lexicon served as valuable entry points into cultural traditions from which even younger generations in China are gradually becoming detached. The experimental framework, which combined vocabulary learning with literary study, therefore proved relevant not only for heritage learners but also for students in China. While the control group received the same culture-oriented explanations and was exposed to the same audiovisual materials, only the experimental group participated in the collaborative dictionary-compilation activities, following the same protocol as the heritage group in Italy.

Moreover, the integration of literature, culture, and vocabulary learning was supported by collaborative group work and the critical use of digital tools. These activities were designed to promote students' active participation in the creation of a cultural product, such as a dictionary. Role rotation and peer review further encouraged cooperative, reflective, and metalinguistic skills, helping learners take an active role in the co-construction of linguistic and cultural knowledge.

[4] Shèxì (社戏, literally "community performance" or "community opera," from shè 'community' and xì 'theatrical performance') refers to traditional operatic performances held during village festivals and communal celebrations in rural China.

[5] Peking Opera is included in UNESCO's Representative List of the Intangible Cultural Heritage of Humanity: https://ich.unesco.org/en/RL/peking-opera-00418.

Results from the pre-test questionnaire showed the unfamiliarity with the target literary vocabulary in all groups (19% correct responses for EG; 18% for CG; 13% for HG), not only for heritage learners. Familiarity scores were higher when the same words were provided in a non-literary context (or *general*, see Figure 1). Baseline differences among groups on the pretest were assessed using a mixed-effects logistic regression model (Table 3). The initial model included register, group, and their interaction, with random intercepts for question and word:

$$Pre \sim 1 + \text{Register} + \text{Group} + \text{Register} \times \text{Group} + (1|\text{Question}) + (1|\text{Word})$$

A random intercept for participant ID was deliberately excluded, as the focus was on inter-group differences. Model selection identified the following reduced model as the best-fitting model:

$$Pre \sim 1 + \text{Group} + (1|\text{Question}) + (1|\text{Word})$$

TABLE 3 Predicted probabilities of success estimated using a mixed-effects logistic regression model. Although Group significantly predicted pretest performance, the 95% confidence intervals around the predicted success probabilities overlapped across all groups, revealing no clear pairwise differences between specific groups

Group	Predicted Probability of success	95% CI
Control	0.11	[0.06, 0.21]
Experimental	0.11	[0.06, 0.21]
Heritage	0.07	[0.03, 0.14]

Group was retained as a predictor in the final model. However, confidence intervals for the predicted scores all overlap with predicted success rates, suggesting no differences between particular groups. The predicted probabilities of success were 0.11 for both the Control and Experimental groups and 0.07 for the Heritage group.

To assess medium-term lexical retention, groups were administered the same sentence-completion test used in the pre-test three months after the instructional intervention. On the delayed post-test, students obtained better scores, suggesting improvement relative to baseline performance (Figure 1).

4 Results

To assess the validity of the proposed learning framework, statistical analyses were conducted using mixed-effects logistic regression models. These models are used to analyse binary outcomes (e.g., success vs. failure) while simultaneously accounting for both fixed effects (experimental predictors of theoretical interest) and random effects (variability associated with participants, items, or tasks). This approach is particularly appropriate for repeated-measures linguistic and psycholinguistic data, where observations are not fully independent. Compared to classical inferential statistics (e.g., ANOVA or standard logistic regression), mixed-effects models offer the advantage of simultaneously modeling experimental effects and the variability associated with participants and linguistic items. This is particularly important in psycholinguistic and lexical research, where observations are not fully independent, since the same participants respond to multiple items and the same items are seen by multiple participants. By accounting for these sources of random variation, mixed-effects models provide more reliable and generalizable estimates of the effects of interest.To analyze the data, we began with a maximal model including four fixed predictors — Group, Register of context, Compiled, and Frequency — and all their interactions, alongside random intercepts for

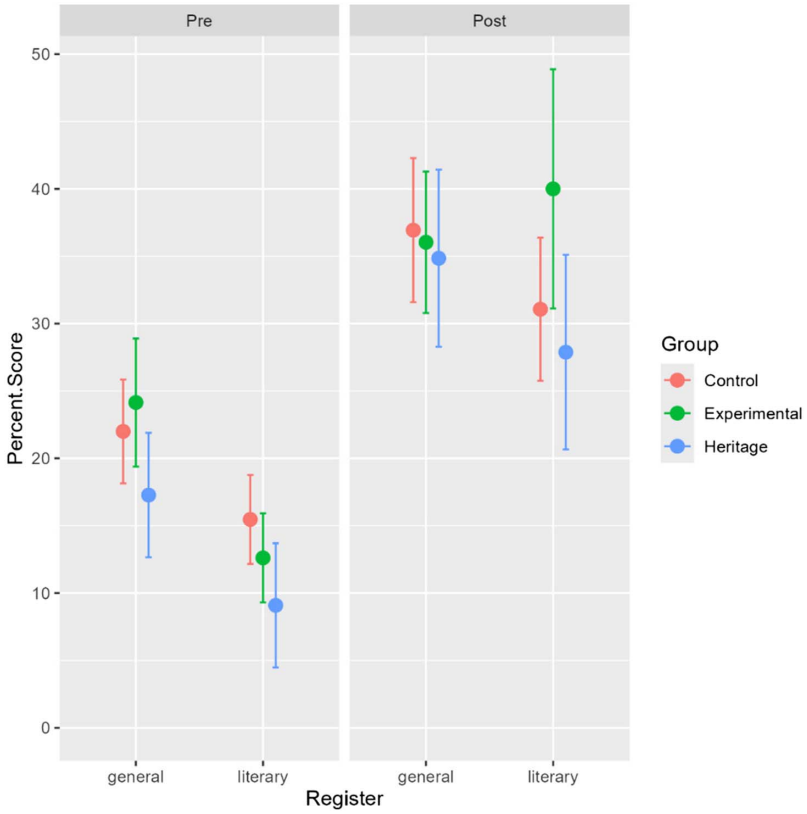


FIGURE 1 Percentage of correct item completion in the pre-test and delayed post-test in the heritage, experimental and control groups for both contexts: literary and non-literary

participants, questions, and words. The predictor *Compiled* indicated whether a given target word had been selected by students for dictionary compilation. The predictor *Frequency* captured differences in word frequency in the zhTenTen21 corpus. The 15 target words were classified as high-, medium-, or low-frequency based on their corpus frequency values. We then applied a principled model reduction procedure which progressively removes terms that do not contribute to model fit, arriving at the most parsimonious model supported by the data. The final model retained Group, Register, and their interaction as fixed effects, with random intercepts for participants ($N = 109$), questions ($N = 30$), and words ($N = 15$).

Importantly, most of the variability in our data came from these random effects — that is, from individual differences between participants, differences in how questions were phrased, and differences between the target words themselves — rather than from the experimental manipulation. This is a normal and expected finding in lexical research, and our model (as *mixed* or *hierarchical* models in general) accounts for all of this background noise before estimating the effects of interest.

Looking at the fixed effects, neither register nor group membership alone made a difference. What did matter was their combination: participants in the Experimental Group were roughly twice as likely to show improvement when the context was literary (OR = 2.08, $p < .001$), compared to the Control group in the same condition (see Figure 2). No such effect was found for the Heritage group, whose sensitivity to context register did not differ reliably from Controls. Predicted improvement was broadly comparable across groups in the general register (approximately 13–17%), whereas in the literary register the model predicted the highest improvement (about 27%) for the Experimental group, compared with the Control (17%) and Heritage (19%) groups (see Figure 3).

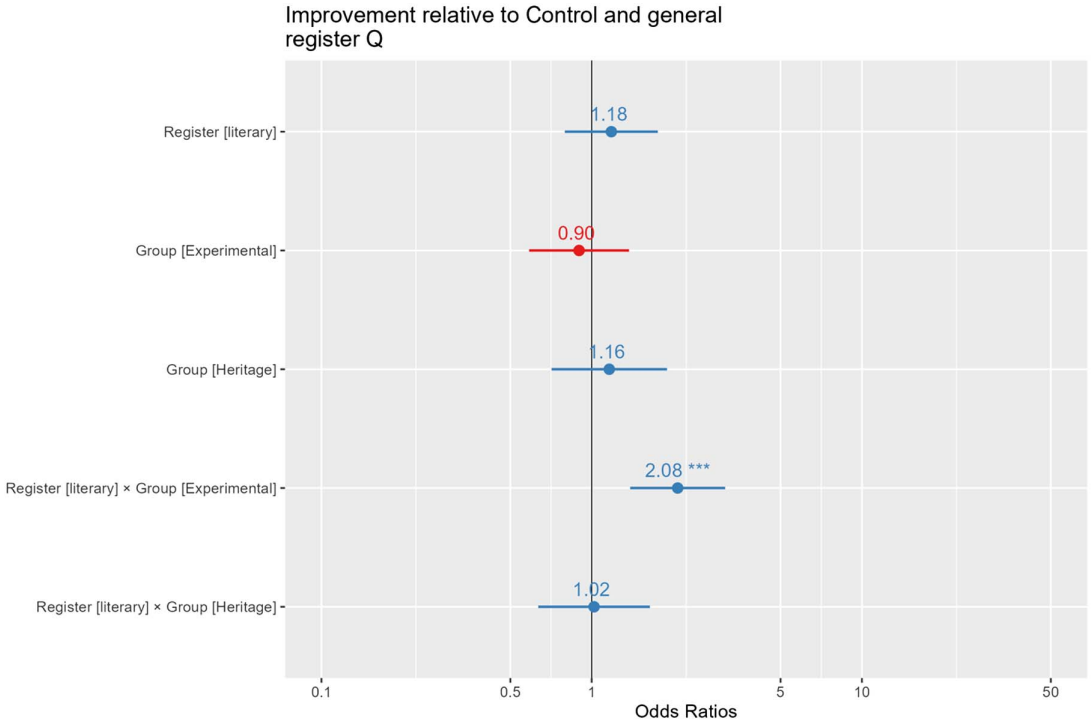


FIGURE 2 Forest plot of odds ratios from the mixed-effects logistic regression model predicting improvement relative to the Control group and the general register condition. Only the interaction between literary register and the Experimental group was statistically significant (OR = 2.08, $p < .001$), suggesting that participants in the Experimental condition were approximately twice as likely to improve when working with literary-register items

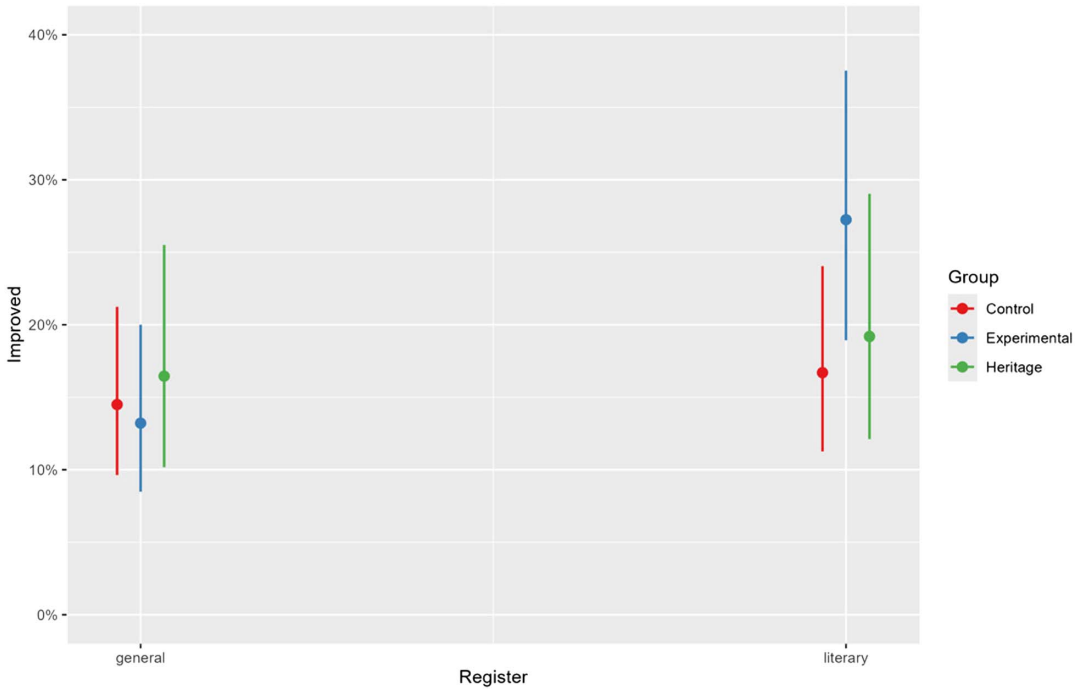


FIGURE 3 Improvement from pre-test to post-test across registers (general vs. literary) and participant groups (Control, Experimental, and Heritage). Error bars indicate 95% confidence intervals

5 Concluding remarks

Given the numerous limitations involved in assessing the effectiveness of the learning framework with heritage students in Italy, the initial research question concerning its potential to enhance L1 vocabulary development and strengthen cultural identity of Heritage Learners could not be conclusively answered on the basis of the present data. Two major factors have limited the effectiveness of the intervention.

The first concerns the lack of access to computers, which likely reduced students' engagement with the dictionary-compilation task. Participants could only use tablets, and some even preferred completing the activity on their smartphones, despite Lexonomy having been designed primarily for larger screens. The second limitation was the absence of a control group of Chinese heritage learners in Italy.

More conclusive evidence regarding the broader question of whether dictionary compilation is an effective strategy for learning new vocabulary was obtained by replicating the framework in China. The study compared the performance of an experimental group following the dictionary-compilation framework with that of a control group receiving traditional, culturally oriented instruction. Notably, students who participated in the dictionary-compilation activities (EG) showed substantial gains in vocabulary acquisition. Over the three months following the instructional intervention, their scores nearly doubled relative to the initial test, and they significantly outperformed the students who received traditional teacher-led instruction.

Another relevant finding relates to the vocabulary selected for testing, which proved to be largely unfamiliar to both groups in China (EG and CG) as well as to the heritage learners (HG), thereby supporting the appropriateness of the lexical selection.

References

Brigadoi Cologna, D. (2023). Sviluppo, prospettive e problemi delle scuole ereditarie cinesi in Italia. *OrizzonteCina*, 14(1), 71–75. https://doi.org/10.13135/2280-8035/8528

Campoy-Cubillo, M. C. (2002). Dictionary use and dictionary needs of ESP students: An experiential approach. *International Journal of Lexicography*, 15, 206–228.

Carreira, M. (2004). Seeking explanatory adequacy: A dual approach to understanding the term "heritage language learner." *Heritage Language Journal*, 2(1), 1–25.

Carreira, M. & Kagan, O. (2011). The results of the National Heritage Language Survey: Implications for teaching, curriculum design, and professional development. *Foreign Language Annals*, 44(1), 40–64.

Caruso, V. (2024). Between collocation and colligation: An experiment in collaborative lexicography. *International Journal of Lexicography*, 37(1), 75–94.

Caruso, V. & De Meo, A. (2025). Tecnologie e apprendimento esperienziale-cooperativo: Dizionari fraseologici digitali per le lingue seconde. *Rassegna Italiana di Linguistica Applicata*, 57(1–2–3), 145–168.

Ceccagno, A. (2003). New Chinese migrants in Italy. *International Migration*, 41(3), 187–213. https://doi.org/10.1111/1468-2435.00246

Chaves-Yuste, B. & Márquez Cruz, M. (2022). Lexicografía digital y adquisición de léxico en inglés: Una experiencia b-learning en los tiempos de Covid-19. In B. Puebla Martínez, R. Vinader Segura & N. Navarro Sierra (Eds.), *Covid-19, multiverso de disciplinas. Una mirada desde la salud, la educación y la comunicación* (pp. 350–367). Dykinson.

Chaves-Yuste, B. & Márquez Cruz, M. (2024). A cooperative digital dictionary to improve EGP reading comprehension in secondary education. In J. Bobkina & E. Domínguez Romero (Eds.), *Transformative trends in language education: Pioneering autonomous learning technologies* (pp. 171–188). Peter Lang.

Confucius Institute (孔子学院). (2014). *HSK zhen ti ji = Official examination papers of HSK*. Gao.

Cummins, J. (1979). Linguistic interdependence and the educational development of bilingual children. *Review of Educational Research*, 49(2), 222–251.

Curdt-Christiansen, X. L. & Hancock, A. (Eds.). (2014). *Learning Chinese in diasporic communities: Many pathways to being Chinese*. John Benjamins.

Elia, A. (2009). Can a collaborative wiki Weblish dictionary project help academic writing of ICT language learners. In I. González-Pueyo., C. Foz Gil., M. Jaime Siso & M. J. Luzón Marco (Eds.), *Teaching academic and professional English online* (pp. 153–180). Peter Lang.

Guo, S. (2022). Reimagining Chinese diasporas in a transnational world: Toward a new research agenda. *Journal of Ethnic and Migration Studies*, 48(4), 847–872.

Harel, I. & Papert, S. (Eds.). (1991). *Constructionism: Research reports and essays, 1985–1990*. Ablex Publishing Corporation.

He, A. W. (2006). Toward an identity theory of the development of Chinese as a heritage language. *Heritage Language Journal*, 4(1), 1–28.

He, A. W. (2010). The heart of heritage: Sociocultural dimensions of heritage language learning. *Annual Review of Applied Linguistics*, 30, 66–82.

He, A. W. (2020). *Heritage language education: A research-based approach*. Routledge.

Jakubíček, M., Měchura M., Kovář, V. & Rychly P. (2018). Practical post-editing lexicography with Lexonomy and Sketch Engine. In *Proceedings of the XVIII EURALEX International Congress* (pp. 431–440). Ljubljana University Press.

Jakubíček, M., Kovář, V., Měchura, M. & Rambousek, A. (2022). Using NVH as a Backbone Format in the Lexonomy Dictionary Editor, in A. Horák, P. Rychlý & A. Rambousek (Eds.), *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN 2022* (pp. 55–61). Tribun EU.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, 1, 7–36.

Leeman, J. (2015). Heritage language education and identity in the United States. *Annual Review of Applied Linguistics*, 35, 100–119.

Liu, H. (2011). An emerging China and diasporic Chinese: Historicity, state, and international relations. *Journal of Contemporary China*, 20(72), 813–832.

Maylor, U., Glass, K., Issa, T., Kuyok, K. A., Minty, S., Rose, A., Ross, A., Tanner, E., Finch, S., Low, N., Taylor, E., Tipping, S. & Purdon, S. (2010). *Impact of supplementary schools on pupils' attainment: An investigation into what factors contribute to educational improvements* (Research Report No. DCSF-RR210). Department for Children, Schools and Families. https://dera.ioe.ac.uk/id/eprint/810/

Mechura, M. (2017). Introducing Lexonomy: An open-source dictionary writing and publishing system. Online. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek & V. Baisa (Eds.), *Electronic lexicography in the 21st century: Proceedings of eLex 2017 conference* (pp. 662–679). Lexical Computing

Papert, S. (1980). *Mindstorms: Children computers and powerful ideas*. Basic Books.

Pellin, T. (2025). Le scuole complementari cinesi in Italia dopo la pandemia, fra grandi città, provincia e l'Europa. *OrizzonteCina*, 16(1), 4–15. https://doi.org/10.82002/117661

Valdés, G. (2001). Heritage language students: Profiles and possibilities. In J. K. Peyton, D. A. Ranard & S. McGinnis (Eds.), *Heritage languages in America: Preserving a national resource* (pp. 37–77). Delta Systems.

Thomas, W. P. & Collier, V. P. (2002). *A national study of school effectiveness for language minority students' long-term academic achievement*. Center for Research on Education, Diversity & Excellence, University of California, Santa Cruz.

Contact information

Zongyuan Wang
Università di Napoli L'Orientale
z.wang1@unior.it

Valeria Caruso
Università di Napoli L'Orientale
vcaruso@unior.it

Anna De Meo
Università di Napoli L'Orientale
ademeo@unior.it

Acknowledgements

This paper is the result of ongoing collaboration among the authors, who jointly endorse its methodology, findings, and claims. Responsibility for drafting the individual sections is distributed as follows: Sections 1, 3.2, 3.3, and 4 were written by Valeria Caruso; Sections 2 and 3.1 by Zongyuan Wang; and Section 5 by Anna De Meo.

Yelena Yerznkyan & Diana Movsisyan

ENCODING EMOTIVITY IN DICTIONARY DEFINITIONS

A Corpus-Based Study of Verbs of Understanding

Abstract The study identifies the emotional profiles of verbs of understanding and demonstrates their relevance for corpus-driven lexicography, contributing to corpus-based research on emotivity and offering implications for lexicographic description. The paper examines the interaction between understanding and emotion as reflected in the contextual behaviour of English verbs of understanding and provides empirical evidence for models of emotionally grounded meaning representation, supported by the emotive clusters most frequently associated with understanding. From a lexicographic perspective, the findings substantiate the inclusion of the emotive component in dictionary descriptions of the verbs in question. Thus, the research argues for a more refined approach to emotivity in lexicographic description of verbs of understanding considering that the emotional meaning makes the axes of the semantics of verbs of understanding.

Keywords Verbs of understanding; collocational pattern; emotivity; corpus linguistics; lexicography; definitions of words

1 Introduction

Understanding as a cognitive and emotive process, has been studied in cognitive linguistics, psycholinguistics, corpus research with growing interest in the interaction between conceptual meaning and emotional dimensions. Understanding is often expressed through a variety of verbs in the language. While all of these verbs denote cognitive processes, some carry implicit or explicit emotional components, such as empathy, sympathy, or emotional awareness. This paper focuses on identifying how emotional dimensions are reflected and realized in contextual uses, thereby contributing to lexicography, where emotive nuances present a challenge for the formulation of dictionary definitions and representation of usage patterns.

Although standard dictionary definitions often prioritize the denotative meanings of the verbs of understanding, the emotional components of these verbs are clearly observed through collocational patterns. In other words, the verbs acquire understanding-related meanings largely through the emotional markers collocated with them.

In this study emotion is treated as an inherent component of meaning rather than an external layer. Proceeding from the assumption that meaning emerges through collocation we argue for a usage-based, corpus-driven approach to meaning. In this regard, the research is in line with the claim that lexical meaning is shaped by co-occurrence patterns (Sinclair, 1991; Stubbs, 2001). It aligns with the discussion that a consistent aura of meaning is imbued by its collocates, explaining how words acquire evaluative meaning through typical contexts (Louw, 1993).

The role of corpus methodologies in revealing collocational patterns (Hunston, 2002), is also evident in recent publications in lexicography and corpus linguistics that tend to integrate collocational and usage-based patterns into dictionaries. Contemporary lexicography relies on corpus-derived collocational patterns and frequency data to refine and enhance dictionary definitions, while user-based perceptual survey data further contribute to shaping

lexicographic entries (Frankenberg-Garcia et al., 2020; Nelson, 2020). This approach comes to argue that dictionary definitions should be empirically grounded and usage-based.

Corpus-based studies further confirm that collocational patterns are central to meaning construction and language acquisition. Collocational profiles have been shown to systematically vary depending on learners' proficiency. It was found that corpus use contributed to a greater collocation output than dictionary use. The reason is that corpus users focused on word combinations and looked up high-frequency collocations, while the dictionary users focused on meaning, then looked up and memorized the collocations (Du et al., 2022; Satake, 2022). These findings substantiate the idea that lexical meaning is inherently usage-based and that collocation is a significant mechanism through which meaning is encoded.

The present study examines how English verbs of understanding collocate with words conveying emotional meaning. By analyzing the frequency with which these verbs co-occur with emotional lexemes, the research aims to uncover the cognitive-emotive mechanisms embedded in linguistic usage and to demonstrate their relevance for lexicography, particularly in structuring dictionary definitions. As many verbs of understanding simultaneously encode conceptual processing and interpersonal evaluation, their emotional profiles reveal not only semantic tendencies but also deeper cognitive-emotional structures used in language and require careful lexicographic representations.

Thus, the paper views verbs of understanding as cognitive-emotive units. Their meanings are dynamically constructed through the interaction of cognitive meanings and emotional collocational patterns. From a lexicographic perspective, this interaction is particularly significant, as it contributes to a more accurate representation of semantic nuances in dictionary definitions.

2 Methodology

Understanding constitutes a complex conceptual domain rather than a single cognitive operation, encompassing a wide range of processes. Therefore, the present study adopts a broad spectrum of verbs of understanding as lexical units that profile one or more stages of the process of making sense of information, experience, or another person's mental and emotional state. A corpus-based and cognitively-oriented approach is applied to a total of 46 verbs denoting various aspects of understanding. The list comprises the following set of verbs presented in descending order of their raw frequency: *know, see, feel, read, understand, follow, reach, realize, note, draw, accept, imagine, recognize, respect, relate to, appreciate, be aware, view, gather, acknowledge, dig, perceive, detect, interpret, take in, sense, distinguish, seize, diagnose, grasp, conceive, make out, envision, penetrate, identify with, validate, comprehend, discern, pinpoint, infer, sympathize, decipher, apprehend, fathom, bottom, empathize*.

The selection of the 46 verbs was developed through a multi-stage procedure. First, through componential analysis a preliminary set of the verbs was identified to reveal their association with cognition and comprehension. Dictionary definitions were consulted to determine whether each verb encoded a meaning component related to understanding. The candidate verbs were examined to determine whether their definitions contain such markers/determiners as *understand, comprehend, perceive*. Second, the verbs were classified according to their dominant cognitive profiles, such as acquisition, perception, recognition, interpretation, conceptualization, inference, socio-emotional comprehension, and interpersonal or empathetic connection. This approach ensured that the selected verbs represented the multidimensional structure of understanding rather than focusing exclusively on intellectual comprehension. Third, the verbs were examined in their immediate linguistic contexts to verify

their actual usage patterns and frequency of occurrence. The final selection was also determined by the availability of adequate contextual data required for the co-occurrence analysis with emotion-related markers. Since the study employs a corpus-based approach, verbs with insufficient corpus representation were excluded. This ensured that the final inventory represented both theoretically relevant conceptual categories and empirically attested language use.

The analysis was conducted using the *Corpus of Contemporary American English* (COCA), a large, balanced corpus comprising both spoken and written registers of modern English. This study employs the COCA as the primary empirical basis for the corpus-based analysis of emotivity in verbs of understanding. The choice of COCA is determined by its substantially larger size and greater temporal coverage, also it is more balanced across registers, all of which are essential for investigating emotive meaning as a context-dependent semantic-pragmatic phenomenon. With over one billion words spanning from the 1990s to the present, COCA provides access to contemporary usage patterns and helps in identifying current emotive tendencies. Consequently, COCA offers methodologically appropriate corpus for capturing the dynamic and usage-based nature of emotivity in verbs of understanding. A co-occurrence window of ±5 words was applied to identify instances where these verbs occurred in proximity to emotional words.

In order to achieve fairly balanced and comprehensive coverage of emotional categories, we rely on a combined framework of emotions and integrate both basic and fine-grained emotions to ensure theoretical reliability and broader emotional representation. Accordingly, our classification is based on well-established categorizations of basic emotions, proposed by Paul Ekman (1992), Robert Plutchik (1980), and Carroll Izard (1993), as well as on the results of recent research which demonstrate that emotional taxonomy extends far beyond the core schemes and provide a wider coverage of emotions. In this regard, *GoEmotions* goes beyond basic emotion categories and covers such emotions as *pride, embarrassment, remorse, and curiosity; some other papers identify the positive emotions of love, pride, interest*, and *joy* (Demszky et al., 2020; Keltner & Cowen, 2021). Furthermore, a recent publication in the field compares the existing systems of emotions and studies their distribution based on different schemes, including Ekman's, Plutchik's, GoEmotions and other classification frameworks, and points out that *sadness, joy (happiness), fear*, and *anger* have a relatively large share of the corpus, with their proportions being very close to each other in each scheme (Zhang et al., 2024). This convergence between classical and contemporary models reinforces our integrative classification approach and highlights the most salient and expressive emotional categories. Accordingly, our research encompasses the following 15 categories: *happiness/joy, sadness, fear, anger, surprise, anticipation/hope, trust, disgust, love, interest, contempt, pride, shame, guilt, and embarrassment, as well as two general evaluative polarities – positive* and *negative*.

In total, 69 evaluative-emotive markers were included in the analysis, representing both evaluative polarities and discrete emotional connotations. These markers comprised the following lexemes: *positive, negative, good, bad, right, wrong, happiness, happy, joy, joyful, cheerful, delighted, glad, sadness, sad, devastated, upset, lonely, despair, pain, boring, bored, fear, afraid, anxious, anger, angry, furious, annoyed, surprise, surprised, astonished, amazed, shocked, anticipation, eager, excited, worried, hope, hopeful, optimistic, trust, trusting, comfortable, confident, safe, secure, devoted, disgust, disgusted, disgusting, unpleasant, love, loving, interest, interested, interesting, contempt, disdain, disdainful, pride, proud, shame, ashamed, shy, guilt, guilty, embarrassment, embarrassed*.

In this paper the emotivity of verbs of understanding is examined through quantitative-qualitative analysis, relying not on frequency alone but also integrating emotive co-occurrences and pragmatic contextual markers. To measure the frequency of collocability of verbs of understanding with emotional markers, a Co-occurrence Frequency Score (CFS)

was calculated for each verb according to the following formula: *CFS=Frequency of verb co-occurring with emotional words/Total frequency of the verb in the corpus×100* (Rychlý, 2008; Pecina, 2010). This score provides a normalized measure of how strongly each verb of understanding tends to occur in emotionally charged contexts. Higher CFS values indicate a stronger association between the verb and emotional language, thus reflecting the relative emotivity of the verb within natural language use.

3 Results and Discussion

3.1 Emotivity Rate of the Verbs of Understanding

The verbs of understanding are categorized by CFS, which reveals the emotivity level of the verbs (see Table 1). The thresholds used are 7.01–18 for very high emotivity; 5.01–7.00 for high emotivity; 3.01–5.00 for moderate emotivity; below 3.00 for low emotivity, as shown in the table below.

TABLE 1 Distribution of CFS Across Emotivity Levels

Very high emotivity	High emotivity	Moderate emotivity	Low emotivity
Feel (17.595)	Sense (6.954) See (6.315) Respect (5.883) Bottom (5.882) Know (5.735) Appreciate (5.607) Realize (5.567) Empathize (5.064)	Read (4.971) Discern (4.836) Imagine (4.474) Understand (4.406) Perceive (4.146) Recognize (4.074) Acknowledge (3.986) Distinguish (3.977) Read into (3.783) Accept (3.720) Note (3.357) Dig (3.270) Apprehend (3.239) Follow (3.088) Sympathize (3.083) Draw (3.082) Take in (3.080) Grasp (3.057)	View (2.972) Seize (2.911) Validate (2.860) Pinpoint (2.802) Make out (2.762) Fathom (2.736) Comprehend (2.710) Identify with (2.653) Interpret (2.653) Relate to (2.487) Be aware (2.473) Reach (2.455) Conceive (2.336) Penetrate (2.316) Infer (2.314) Gather (2.239) Detect (2.121) Decipher (2.079) Envision (1.931) Diagnose (1.830)

The resulting data were subsequently analyzed to identify patterns and tendencies across the verbs under study, offering insights into the interaction between cognitive and emotional components in linguistic representation. The findings (see Table 2) indicate that the most frequent verbs primarily related to cognition and perception (*know, see, feel*) also exhibit a high degree of emotional loading.

TABLE 2 Comparative Frequency and Emotivity Data for Verbs of Understanding

Verbs of understanding	Raw frequency of the verb	CFS of the verb	Emotivity rate
Know See Feel	2762921 1961110 743031	5.735 6.315 17.595	Very high and high emotivity
Read Understand Follow	391918 349203 300266	4.971 4.474 3.088	Moderate emotivity

Although the raw frequency was calculated for all the meanings of the verbs under study (including highly polysemous verbs such as *know, see, read,* etc.), the semantic analysis concentrated on occurrences expressing the meaning of understanding. Thus, the study integrates the application of quantitative analysis of lexical frequency with qualitative semantic interpretation. Notably, this approach revealed that emotional co-occurrence contributes to the activation of understanding-related interpretations of sensory, perceptual, and action-based physical verbs. The presence of emotional markers in the immediate context served as one of the indicators that the verb was functioning within the conceptual domain of understanding rather than merely expressing a literal physical or perceptual action. The corpus evidence indicates that emotional co-occurrence constitutes one of the contextual mechanisms through which embodied verbs are construed as expressing understanding. The data also suggest that fundamental, high-usage verbs tend to serve (although not always) as crucial linguistic anchors for expressing or perceiving emotion. Similarly, verbs with lower raw frequency tend to have moderate emotivity rate. This indicates that the frequency of usage of the verb is one of the determining factors (although not the only one) directly related to the emotivity rate.

The analysis of the 69 markers associated with the verbs revealed a diverse distribution. The total counts for positive (204319) and negative (51734) evaluative polarities indicate that verbs of understanding tend to have a positive profile nearly 3.94 times more often than a negative one. The observation of the primary emotion categories across the verbs under study suggests that they are predominantly positively connotated. The same tendency works in the distribution of positive, negative and ambiguous emotions: verbs of understanding are 2.14 times more frequently collocated with positive emotions rather than with negative ones. On the whole, understanding in English tends to be conceptualized as a positive, socially oriented process. The more detailed counts are introduced in Table 3.

TABLE 3 Classification of Emotions into Positive, Negative and Ambiguous Categories

Positive emotions	Negative emotions	Ambiguous emotions
love (44318), trust (38302), happiness/joy (30060), anticipation/hope (28667), interest (26972)	sadness (30916), fear (14444), guilt (11858), anger (9856), shame (5821), embarrassment (2709), disgust (2002), contempt (803)	surprise (13901), pride (6073)
Total counts: 168319	Total counts: 78409	Total counts: 19974

Further data are collected and correlated in Figure 1, a dual-axis figure showing both raw frequency and prominence of each emotion among verbs. The X-axis indicates the range of emotions under study – *love, trust, sadness,* etc. The Left Y-axis (bars) displays the frequency of verb usage associated with each emotion. The Right Y-axis (red line) shows the number of verbs where that emotion ranks in the top 5.

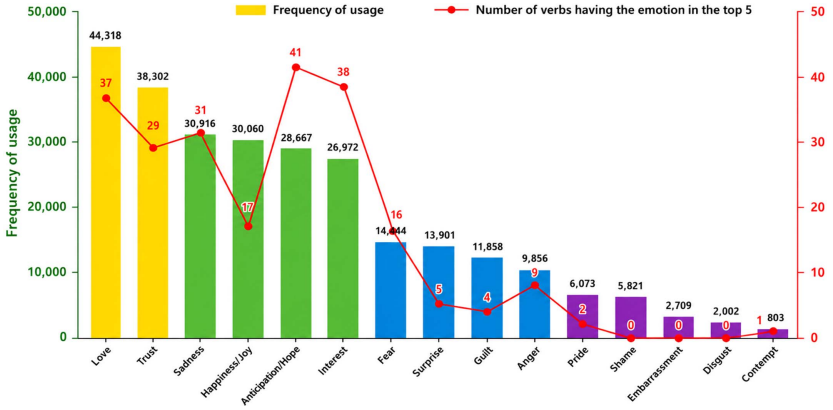


FIGURE 1 Verbs-Emotions Correlation Profile: Frequency and Top-5 Occurrence

Yellow and green bars (positive emotions: *love, trust, happiness/joy, anticipation/hope, interest*; the negative emotion: *sadness*) dominate the frequency spectrum. The core tendency observed is the association of verbs of understanding with *love*, which constitutes the highest percentage marker for every verb examined. This suggests that much of the emotional loading in verbal communication is fundamentally relational and affiliative. The categories of *love* and *trust* are dominant in total co-occurrence counts, significantly surpassing all other emotions, suggesting that the verbs in the dataset are most frequently used in contexts associated with *love* and *trust*. *Sadness*, *happiness/joy*, *anticipation/hope*, and *interest* follow as the next most dominant categories. These findings highlight the particular centrality of affection and bonding, combined with alertness and caution, in the discourse represented by this verb set.

The top-5 counts (red line) do not strictly follow frequency. Some emotions like *anticipation/hope* and interest peak in prominence. Even though *love* and *trust* have the highest raw frequencies, *anticipation/hope* and *interest* are reflected in the top-5 counts more often. Despite their slightly lower frequency, these emotions most consistently characterize the emotional profile of verbs of understanding. Similarly, although *sadness* and *happiness/joy* have nearly equal raw frequencies, sadness more consistently ranks among the top-5 emotional markers of the verbs. Notably, *sadness*, occurring more than twice as frequently as *happiness/joy* in the emotional profiles of the verbs, is the only negative emotional marker consistently appearing in the top-5 range. Likewise, the top-5 prominence of *happiness/joy* and *fear* is nearly equal, although the raw frequencies significantly differ.

Blue and purple bars (negative and ambiguous emotions: *fear, surprise, guilt, anger, pride, shame, embarrassment, disgust*, and *contempt*) have lower frequencies resulting in low linguistic prominence. Top-5 counts drop sharply with *fear, surprise, guilt, anger*, and *pride*. The representation of these emotions in the top-5 emotive markers within verbs of understanding is minimal. These emotions appear frequently but are less often a primary emotion for verbs, indicating they are secondary associations rather than dominant. The findings of the study also indicate that the emotions *shame, embarrassment, disgust*, and *contempt* almost never appear among the top-5 emotions associated with any of the examined verbs, that is to say, they do not emerge as dominant emotions for any verb.

Emotions occur not in isolation but in interconnected clusters that reflect the interplay between cognitive and emotive processes in language use. In this regard, *love* most frequently co-occurs with *anticipation/hope, interest*, and *trust*. *Anticipation/hope*, in turn, aligns most strongly with *interest* and *trust*, while *interest* also pairs with *trust*. These emotions consistently

cluster together and are commonly represented among the top-5 emotions. This cluster persists even for verbs with lower frequency of usage and fewer instances of emotional representation. These patterns demonstrate a stable, interrelated network of positive emotions that consistently underpins the semantic and emotional profiles of the verbs of understanding.

Notably, the emotional cluster *anticipation/hope-trust-love-interest* frequently co-occurs with a broad range of verbs of understanding. The cluster collocates with *understand, know, discern, distinguish, apprehend, interpret, be aware*, which denote cognitive processing, internal mechanisms of comprehension and the acquisition of conceptual clarity. The cluster collocates with *recognize, appreciate, respect, validate* as well, the verbs denote recognition and evaluation, encode acts of acknowledgment and positive evaluative stance. The verbs *envision, reach, dig, follow,* denoting conceptual expansion and mental projection and highlighting forward-directed cognitive activity oriented toward discovery, insight, also collocate with the cluster under discussion. The emotional pattern of *anticipation/hope-trust-love-interest* emerges because verbs of understanding inherently require positive cognitive engagement, evoke confidence and certainty, and frequently involve evaluative or socially affiliative meanings. Moreover, verbs implying mental projection or conceptual advancement align with future-oriented emotional states such as *hope* and sustained *interest*; the verbs naturally attract constructive, *trust*-based, and affectively positive emotional associations.

While studying the emotion of surprise in the top-5 range, two clusters are identified: *anticipation/hope-love-surprise* (*see, imagine, make out, diagnose*) and *anticipation/hope-interest-surprise* (*see, make out, detect, diagnose*). The data further suggest that *anticipation/hope* serves as a central and stable emotional element in the verbs' profiles, whereas *surprise*, though constant, aligns with different variables – either with *love* or with *interest*. As can be inferred, the cluster is linked to visual perception-based verbs. Overall, these recurrent pairings indicate that *surprise* functions as a key marker within perception-driven cognitive processes, reinforcing its role during acts of visual understanding. Consequently, it is argued that perceptual processing is a primary trigger for cognitively charged emotional responses.

One of the most frequent emotional clusters observed is *sadness-fear-anger*. These emotions enter the top-5 range for certain verbs. Specifically, the full *sadness-fear-anger* negative emotional cluster is attested in *sense, empathize, sympathize, seize, fathom, diagnose*. A reduced *sadness-fear* cluster emerges in *perceive, acknowledge, grasp, pinpoint, comprehend, identify with* as well, while the *sadness-anger* pairing is characteristic of *sense, empathize, sympathize, seize, fathom, penetrate, diagnose*. Some of the verbs highlight analytic or inferential cognition (*grasp, comprehend, diagnose*), others emphasize intuitive or affective attunement (*sense, empathize, sympathize, identify with*), while still others highlight socially grounded recognition and acknowledgment (*acknowledge, respect*). As seen, across these patterns, *sadness* consistently functions as the dominant emotive marker, followed by *fear* and *anger*, with *anger* most often occurring only in combination with *fear*. These tendencies suggest that verbs of understanding frequently encode a layered negative affective structure in which *sadness* anchors the cluster and conditions the co-occurrence of other negative emotions.

The analysis of the data shows there are a couple of negative clusters which match rather frequently: *sadness-guilt, fear-guilt, anger-guilt*. *Guilt* is paired with *sadness* (*feel, relate to*), it is also paired with *fear* (*apprehend*), as well as with *anger* (*infer*). These patterns indicate that whenever *guilt* is in the top-5 emotions in the emotional profile of a verb, it systematically aligns with other negative emotions. These recurrent combinations reveal that *guilt* is not an isolated emotion but functions as regularly co-occurring with other negative states. Verbs of empathetic and relational understanding (*feel, empathize*), as well as those of sensory and perceptual

understanding (*sense, perceive*) show the highest concentrations of specific negative emotions matching with the cluster *fear-anger-sadness-guilt* suggesting their primary role in communication is related to processing or sharing adverse emotional experiences, that is they are verbs of emotional exchange.

Taken together these findings, it is claimed that

1. High-frequency emotions do not always correlate with top-5 occurrence.
2. Some emotions (*anticipation/hope, interest, sadness*) are highly significant across verbs despite comparatively lower raw frequency.
3. Verbs of understanding are not typically linked to negativity. Negative and ambiguous emotions are rare in verb usage and seldom dominant. Namely, negative emotions such as *disgust, contempt, shame,* and *guilt* rarely align with this group of verbs.
4. Understanding is emotional, and mostly in a positive direction. The verbs of understanding under study are typically associated with positive emotional profiles: they co-occur with *love, anticipation/hope, interest,* and *trust*. Positive emotions dominate both usage and top-5 prominence. The verbs of understanding usually carry a strong positive or prosocial emotional load.
5. The verbs of understanding constitute a subclass that conceptualizes understanding as an active, often emotive process involving depth, insight, and a dynamic interplay between cognition and emotion.
6. Verbs gain their meaning of understanding largely through the emotional markers that accompany them. In other words, it is the emotional load embedded in their usage – whether conveyed through connotation, context, or collocational patterns – that enables these verbs to develop and express semantic nuances associated with understanding. This emotional component shapes and reinforces the meaning of understanding, transforming otherwise neutral, merely cognitive or action-based verbs into carriers of interpretive, perceptive or empathetic understanding.

3.2 From Emotions to Definitions

Emotivity, as a context-sensitive semantic phenomenon, requires qualitative contextual/pragmatic study in order to reveal how emotional associations interact with verb meaning and how they can contribute to the definitions of the verbs.

While the CFS provides a systematic mapping of emotivity across the verb set, frequency-based evidence alone cannot fully explain how emotional meaning is linguistically encoded, interpreted and pragmatically realized. To address this limitation, the second part of the study undertakes a qualitative, corpus-driven semantic analysis of the selected verbs with emotivity levels ranging from 4.0 to 17.595, identified in the quantitative phase. These case studies demonstrate how statistically observable emotional profiles of cognitive verbs acquire affective depth. By integrating quantitative distributional patterns with qualitative semantic interpretation, the study offers a comprehensive account of the interaction between understanding and emotion in contemporary English through suggested definitions that incorporate information on recurrent collocational behaviour, emotive load and evaluative polarity. The corpus-driven indicators of emotional involvement are treated as integral components of meaning representation, thus reflecting both the semantic structuring and usage-based functioning of this verb class.

To encode emotivity in dictionary definitions of the verbs of understanding, the study proposes a systematic algorithm. The procedure includes corpus-based collocational analysis and lexicographic modelling and is applied across the verbs under study. The following steps have been undertaken to work out a more elaborate definition.

1. As a **first step**, the **emotional profile** of each verb is extracted based on corpus data. Emotions co-occurring with the verb of understanding are identified and quantified, then the dominant and secondary emotional associations are determined.

2. As a **second step**, the **CFS** is used as an indicator measuring the overall emotivity level: very high, high, moderate, low.

3. The **third step** defines the **evaluative polarity** of the verbs under study: positive, negative, or dual-balanced. Evaluative polarity reflects generalized positive or negative judgment encoded through evaluative markers.

4. The **fourth step** determines the **emotional connotation**, verbs are classified based on the distribution of emotional markers: positive, negative, or mixed. Distribution of emotional markers reflects the collocability of verbs of understanding with specific emotional categories.

5. The **fifth step** places each verb on a **cognitive-emotive continuum**; distinguishing between primarily emotive verbs, emotive-cognitive verbs, cognitive-emotive verbs, and predominantly cognitive verbs.

6. In parallel, we defined the **functional role of emotion**: emotions functioning as mechanisms of interpersonal connection (*trust, love*), indicators of vulnerability (*sadness, fear, guilt*), or triggers of cognitive transformation (*surprise, interest, anger, embarrassment*). This step ensures that emotivity is not treated as incidental but as structurally meaningful.

7. Based on these parameters, an **emotive profile statement** is worked out for each verb which summarizes its emotional configuration in a standardized format. This profile serves as the foundation for definition construction.

8. Two types of definitions are then proposed. The **first definition** is more formal and integrates cognitive and emotional components into a lexicographic description. The **second definition** is more general, concise and user-friendly and covers the core emotive content. The suggested definitions are illustrated by corpus-driven examples.

The proposed algorithm ensures coherence within all the stages; it enables consistent incorporation of information about emotivity, dominant emotional associations, collocational preferences and evaluative polarity into the semantic description of the verbs.

The definitions proposed in this study are intended as alternatives to traditional dictionary descriptions rather than as replacements for existing lexicographic entries. Their primary aim is to make the emotional dimension of verbs of understanding explicitly signalled. The proposed model demonstrates how corpus-based emotional data can enrich dictionary entries and provide users with a more comprehensive account of word meaning. The definitions proposed may serve as entries for a specialized cognitive-semantic lexical resource, particularly an online dictionary or database of English verbs denoting cognitive processes. Such expanded corpus-informed definitions would complement traditional lexicographic descriptions by integrating semantic, cognitive, and corpus-based information, including conceptual profiles, emotional associations, and contextual patterns of usage, constituting as an integral aspect of lexical meaning. The proposed format is therefore designed primarily for advanced users, such as researchers, lexicographers, and language specialists.

To make the whole picture more organized and vivid the verbs are classified into 3 main categories, according to the type of understanding they denote: empathetic and relational understanding, intellectual and analytical understanding, sensory and perceptual understanding. The classification of the verbs into three categories was conducted through manual qualitative analysis guided by cognitive-semantic principles. The classification procedure consisted of a combination of lexicographic analysis and corpus-based examination. First, dictionary definitions were analyzed to identify the core semantic components associated with each verb. Second, corpus examples were examined to determine how these meanings are realized in authentic contexts. The assignment criteria focused on the primary cognitive operation represented by the verb: understanding another person's emotional or experiential state (empathetic and relational – *feel, empathize, sympathize, respect, appreciate*, and *relate to*), processing information and constructing knowledge (intellectual and analytical – *know, understand, realize, read,* and *infer*), or acquiring understanding through perception and recognition (sensory and perceptual – *sense, see, perceive, recognize, discern*, and *detect*). In cases where a verb exhibited characteristics of more than one category, the classification was based on its most prominent semantic function within the conceptual domain of understanding. This approach allows the classification to capture the cognitive organization of the lexical field while acknowledging the inherent flexibility and overlap of semantic categories. The study selected the most prominent verbs with the highest CFS values from each category to illustrate the definition-generation process following the proposed algorithm.

3.2.1 Empathetic and Relational Understanding

To demonstrate the applicability of the proposed algorithm, the verb *feel* is analyzed step-by-step according to the model proposed for encoding emotivity in dictionary definitions.

1. **Emotional profile extraction**. Corpus analysis reveals that the verb *feel* frequently co-occurs with a wide range of emotions. *Trust* (21112), *sadness* (12567), *guilt* (9011), *love* (4906), *happiness/joy* (4752) are found among the top-5 emotions collocated with the verb. *Trust, sadness,* and *guilt* have a more dominant role, while *love, happiness/joy*, along with *fear, anger, shame* are secondary emotions. As can be seen the emotional range is rather broad.

2. **CFS**. *Feel* has the highest CFS (17.595) among all verbs analyzed, positioning it as the most 'emotional' verb in the dataset.

3. **Evaluative polarity**. The positive polarity of the verb under study counts at 43425, while the negative polarity counts at 14918. Thus, *feel* has a primarily positive evaluative polarity.

4. **Emotional connotation**. Corpus data show that *feel* is collocated with a wide range of emotions, both positive and negative. The distribution of positive (35556) and negative (33645) emotional markers is quite broad and balanced, covering both high-frequency positive and high-intensity negative emotions, as well as some elements of ambiguous connotation (3190). Thus, *feel* incorporates both positive and negative components, which enables us to consider it as a mixed type of emotional connotation.

5. **Cognitive-emotive continuum**. Unlike other verbs of understanding, *feel* does not present a pure cognitive process, it also denotes immediate emotional experience. Due to its purely experiential nature the verb is tagged as a primarily emotive verb.

6. **Functional role of emotion**. The verb *feel* functions as an immediate and direct lexical link to internal emotional states, affective cognition and relational attunement. In *feel*, emotion not only functions as primary access to understanding and cognitive awareness of one's internal state, but also as connection to one's own self and to others (*trust, love*), and as vulnerability (*sadness, guilt*).

7. **Emotive profile statement**. *Feel* displays a maximally broad, balanced emotive profile characterized by dominant associations with *trust, sadness*, and *guilt*, alongside a full spectrum of secondary emotions including *happiness/joy, love, fear, anger, shame*, and *anticipation/hope*, indicating its role as the primary linguistic marker of direct emotional experience.

8. **Definition generation (Academic/Formal and General/Concise)**. (1) **Feel** (v.) – a highly emotive experiential verb distributing a positive evaluative polarity and denoting direct awareness of one's own or another's emotional state. Its emotional profile, characterized by both positive and negative connotations with dominant associations *to trust, sadness, guilt, love*, and *happiness/joy,* reflects a broad spectrum of human affect. The verb encodes a pre-reflective, embodied mode of understanding in which emotional experience precedes cognitive interpretation or rational understanding. (2) **Feel** (v.) – to experience emotions or physical sensations directly and personally as an immediate way of being aware of oneself or others. **Sample sentences**: *It's public and that makes it **feel safe**; we can trust each other through our shared experiences. You **feel blue**, you **feel sad**? Magnus does **feel** a small stab of **guilt**.*

The consistent application of the proposed algorithm demonstrates that *feel* functions as an emotive verb of understanding, integrating the widest emotional range, the highest emotivity score, and the most direct form of experiential understanding. This makes *feel* a prototypical verb for examining how emotional meaning is directly encoded in verbs of understanding. This confirms its role as the central reference point against which other verbs of understanding can be comparatively evaluated.

Referring to its strong relational component, maximum emotivity and direct emotional experiences it is argued that *feel* expresses empathetic and relational type of understanding. Here also belong the verbs *appreciate* and *empathize*, the definitions of these verbs are presented below. As mentioned earlier in the paper, the study employs an alternative definitional template that integrates cognitive meaning with emotionally relevant semantic information and illustrative corpus examples.

Appreciate (v.) – to recognize and feel grateful for the value or importance of someone or something, combining understanding with a range of positive emotional associations – *love, anticipation/hope*, as well as *interest, happiness/joy,* and *trust*; an emotive-cognitive verb collocating mainly with positive emotions and characterized by a positive evaluative polarity*: I grew to **love** the music and to **appreciate** all of its subtleties. . . . which you might **appreciate** as an **interesting** and perhaps somewhat unique view of western cultural history*.

Empathize (v.) – to understand and share another person's emotions, especially their *sadness, love, fear, anger*, and *anticipation/hope*, establishing emotional connection with their experience; an emotive-cognitive verb with a positive evaluative polarity while more frequently collocating with negative emotions: *People could also **empathize** with her **sadness** that she had . . . So, the Obama administration is taking the opportunity to **empathize** with the **anger** that many Americans feel about that*.

3.2.2 Intellectual and Analytical Understanding

1. **Emotional profile extraction**. Corpus data indicate that the verb *know* tends to co-occur with positive emotions: *love* (18974), *trust* (8339), *happiness/joy* (7436), *anticipation/hope* (7115), *interest* (7011) which align with the top-5 emotions. Negative emotions, such as *sadness* (6939), *fear* (3915), *anger* (2131), and *guilt* (1305), rank after positive emotions in

frequency of occurrence. This means the distribution is broad and complex, spanning both positive and negative domains.

2. **CFS**. The CFS of the verb is 5.735, the verb demonstrates a high level of emotivity.

3. **Evaluative polarity**. The comparison of the positive polarity (69890) with the negative polarity (18269) shows that the verb has predominantly positive evaluative polarity.

4. **Emotional connotation**. The emotional configuration includes strong positive (48695) and a solid share of negative (16606) elements, as well as some traces of neutral elements (5009). Based on the distribution of emotional markers the verb reveals a prevailing positive emotional connotation.

5. **Cognitive-emotive continuum**. The verb *know* is fundamentally cognitive, but carries substantial emotional weight. That is to say, cognition leads this process, while emotion enriches it. Thus, the verb is considered to be cognitive-emotive.

6. **Functional role of emotion**. Emotions function as deepening knowledge, creating moral, and interpersonal awareness, intensifying the impact of understanding. The verb under study covers integration of interpersonal connection with some degree of vulnerability and cognitive transformation.

7. **Emotive profile statement**. *Know* displays a broad emotive profile with dominant emotions of *love, trust, happiness/joy, anticipation/hope,* and *interest* and a wide range of secondary emotions, such as *sadness, fear, anger,* and *guilt*. Due to its strong co-occurrence with *love* and *trust, know* is positioned as an emotionally connected form of understanding. On the other hand, the strong presence of *sadness* and *fear* signals that knowing can be associated with vulnerability. This indicates that *know* functions as a marker of cognitive mastery and emotional attunement, reflecting its function as an integrative form of understanding characterized with cognitive awareness inseparable from emotional engagement. Thus, *know* embodies the total fusion of intellect and emotion.

8. **Definition generation (Academic/Formal and General/Concise)**. (1) **Know** (v.) – a cognitive-emotive verb indicating a positive polarity and denoting possession of understanding through cognitive awareness, reflection, experience or familiarity. Its emotional profile, characterized by dominant associations with *love, trust, happiness/joy, anticipation/hope,* and *interest,* along with some negative emotions, reflects the integration of intellectual certainty with emotional component. The verb encodes a reflective mode of understanding in which cognition is enriched by affective resonance. (2) **Know** (v.) – to be in possession of information or understanding about something or someone, often involving both clear cognitive awareness, intellectual certainty, and emotional connection. **Sample sentences**: *I **know** he would be a **loving** and caring father. We need to **know** why we have to **trust** the current regime. I **know** I've been very **happy** to be here. I wanted to **know** more about her **sadness** and her **aloneness**.*

Know represents a cognitively anchored form of understanding in which emotivity functions as an integrative and enriching component rather than the primary mechanism. The verb *know* is classified as denoting intellectual and analytical type of understanding with strong emotive integration. The verbs *realize* and *understand* also belong to this type of understanding. The suggested refinements of definitions of the verbs are presented below.

Realize (v.) – to become fully aware of something, often unexpectedly, in a way that changes how you think or feel, evoking emotions such as *trust, anticipation/hope, love, sadness, happiness/joy*, along with such secondary emotions as *surprise* and *interest, fear* and *anger*; a cognitive-

emotive verb encoding positive emotional connotations and possessing a positive evaluative polarity: *I **realized** I was **comfortable** with my past. That's when I **realized** how **happy** I was to just have the option available for me. The **fear** only came when I **realized** later how close I came to the end.*

Understand (v.) – to know or realize the meaning of something or someone in a way that combines clear thinking with emotional awareness, often linked *to anticipation/hope, love, sadness, trust*, and *interest*; a cognitive-emotive verb characterized predominantly by positive emotions combined with quite strong associations with negative emotions and comprising a positive evaluative polarity: *I **hope** to make the public **understand** how dangerous smoking can be. If you **love** me, you'll **understand** that I needed time. I **understand** you're **upset** that I have perfectionist tendencies.*

3.2.3 Sensory and Perceptual Understanding

1. **Emotional profile extraction**. The verb *sense* is more frequently collocated with the following top-5 emotions: *sadness* (175), *fear* (165), *anger* (86), *anticipation/hope* (65), and *interest* (61). As can be inferred, the dominant top-3 emotions associated with the verb under study are negative, while the positive emotions of *anticipation/hope, interest*, along with other positive emotions of *happiness/joy* (58), *love* (51), *trust* (45) tend to be secondary ones. The emotional range is balanced and moderately broad.

2. **CFS**. The CFS of the verb *sense* counts at 6.954, the verb shows a high level of emotivity.

3. **Evaluative polarity**. The verb shows nearly equal presence of positive (260) and negative (257) polarity. Thus, the evaluative polarity of the verb is balanced, *sense* is dual-balanced.

4. **Emotional connotation**. Given the asymmetrical distribution of emotions, it is claimed that *sense* has primarily negative (494) emotional connotation, meanwhile reflecting some sensitivity to both positive (280) and neutral (49) emotional states.

5. **Cognitive-emotive continuum**. The verb *sense* implies perceptual awareness with emotional sensitivity, but does not involve deep emotional immersion. It is classified as a cognitive-emotive verb with perceptual and intuitive grounding.

6. **Functional role of emotion**. Negative emotions (*sadness, fear*) associated with the verb *sense* function as indicators of tension, discomfort, or vulnerability; *anger* and *interest* add a layer of cognitive transformation; while *anticipation/hope, interest*, and *love* introduce the component of empathy and to some extent function as attunement to emotional atmospheres and connection. On the whole, we observe here emotional sensitivity guiding perception without full emotional involvement.

7. **Emotive profile statement**. *Sense* exhibits a balanced emotive profile characterized by near symmetry between positive and negative polarities, with a prevailing negative connotation caused by dominant associations to *sadness* and *fear*, and secondary links to *interest, anticipation/hope*, and *love*, reflecting its function as an emotionally receptive but controlled mechanism of perceptual and intuitive awareness. It reflects a neutral yet emotionally aware state – *sensitivity* both to external phenomena and to emotional undercurrents.

8. **Definition generation (Academic/Formal and General/Concise)**. (1) **Sense** (v.) – an emotively balanced cognitive-emotive verb denoting awareness achieved through sensory perception or intuition. While exhibiting a symmetrical distribution of positive and negative evaluative polarities, *sense* is characterized primarily by negative emotional connotation, marked by dominant associations with *sadness, fear, anger*. The verb reflects attentiveness to subtle emotional and situational cues. It encodes a perceptual mode of

understanding in which emotional sensitivity functions as an auxiliary rather than a determining factor in cognitive awareness. (2) **Sense** (v.) – to become aware of something through the senses or intuition, often noticing subtle feelings or emotional signals in people or situations. **Sample sentences**: *They smiled back at him, **sensing** the **sadness** coming from his eyes. Don't let him **sense** your **fear**.*

The analysis shows that *sense* occupies an intermediate position in the system discussed above. It is less cognitively explicit than *know*, yet more controlled than purely emotive verbs. Emotion functions as a background perceptual signal, enabling detection of subtle meanings rather than constituting the core of experience. Thus, *sense* represents a form of emotionally attuned perception, where understanding emerges through sensitivity to implicit emotional and situational cues without full emotional immersion.

The verb *sense* is classified as denoting sensory and perceptual type of understanding, along with *see* and *recognize*. The proposed enhancements of dictionary descriptions are illustrated below.

See (v.) – to understand something clearly, often recognizing emotions like *happiness/joy, love, anticipation/hope, interest,* and *surprise*, as well as *sadness, fear*, and *anger*; a highly emotive perceptive-cognitive verb more frequently aligned with positive emotions and a positive evaluative polarity*: But I **see** it as the best **hope** for making our planet cleaner. I can **see** you're **upset** about that.*

Recognize (v.) – to know or identify someone or something, or to acknowledge their importance or value, frequently associated with *trust, anticipation/hope, love, interest*, as well as *happiness/joy, sadness*, and *fear*; a cognitive-emotive verb mainly aligned with positive emotions, although strongly linked to negative emotions as well and exhibiting a positive evaluative polarity: *I have learnt to **trust** by **recognizing** the truth about myself. It's why the FDA insists that sugar is "generally **recognized** as **safe**" despite considerable evidence suggesting otherwise.*

4 Conclusion

The comparison between lexicographic descriptions and corpus evidence reveals several points of systematic variation in the treatment of verbs of understanding.

First, emotional meaning is significantly underrepresented in dictionaries, which tend to prioritize cognitive interpretations and treat affective uses as secondary or figurative, whereas corpus data demonstrate that emotional meanings are both central and frequent in actual usage.

Second, dictionaries largely fail to capture stable collocational patterns, such as *feel + emotion*, *realize + emotional state*, or *see/read + emotion*, which emerge in the corpus as regular and predictable rather than incidental combinations.

Third, while lexicographic tradition maintains a clear division between cognition (*know, understand*) and emotion (*feel*), corpus findings instead point to a continuum in which cognitive and affective processes are deeply intertwined.

Fourth, the intersubjective dimension of the verbs under study is notably underdescribed in dictionaries: in authentic discourse, they are extensively used to interpret others' emotional

states, validate feelings, and construct empathy. This interactional and relational function, however, remains largely absent from traditional dictionary definitions, further highlighting the gap between lexicographic abstraction and real language use.

Fifth, the paper provides potential solutions for constructing definitions of emotionally loaded terms. In particular, it argues that emotional implications are not peripheral but constitutive of lexical semantics and should therefore be systematically and explicitly included in dictionary entries following the model proposed in this study to ensure a more comprehensive and accurate description of word meaning.

To summarize, the usage-based patterns and cognitively oriented approach enabled us to propose a better lexicographic description of the verbs under study through the new format of definitions. The study provides empirical evidence for models of emotionally grounded meaning representation and argues for a more refined approach to emotivity in dictionary definitions.

References

Demszky, D., Movshovitz-Attias, D., Ko, J., Cowen, A., Nemade, G. & Ravi, S. (2020). GoEmotions: A Dataset of Fine-Grained Emotions. In D. Jurafsky, J. Chai, N. Schluter & J. R. Tetreault (Eds.), *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 4040–4054). Association for Computational Linguistics.

Du, X., Afzaal, M. & Al Fadda, H. (2022). Collocation Use in EFL Learners' Writing Across Multiple Language Proficiencies: A Corpus-Driven Study. *Frontiers in Psychology,* 13, 752134. https://doi.org/10.3389/fpsyg.2022.752134

Ekman, P. (1992). An argument for Basic Emotions. *Cognition and Emotion*, 6(3–4), 169–200. https://doi.org/10.1080/02699939208411068

Frankenberg-Garcia, A., Rees, G. P., Lew, R. (2021). Slipping through the Cracks in e-Lexicography. *International Journal of Lexicography*, 34(2), 206–234. https://doi.org/10.1093/ijl/ecaa022

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge University Press.

Izard, C. E. (1993). Four Systems for Emotion Activation: Cognitive and Noncognitive Processes. *Psychological Review*, 100(1), 68–90. https://doi.org/10.1037/0033-295X.100.1.68

Keltner, D. & Cowen, A. (2021). A Taxonomy of Positive Emotions. *Current Opinion in Behavioral Sciences*, 39, 216–221. https://doi.org/10.1016/j.cobeha.2021.04.013

Louw, B. (1993). Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds.), *Text and Technology: In Honour of John Sinclair* (pp. 157–176). John Benjamins.

Nelson, K. (2020). Informing Lexicographic Choices Through Corpus and Perceptual Data. *International Journal of Lexicography*, 33(3), 251–268. https://doi.org/10.1093/ijl/ecz030

Pecina, P. (2010). Lexical Association Measures and Collocation Extraction. *Language Resources and Evaluation*, 44(1–2), 137–158. https://doi.org/10.1007/s10579-009-9101-4

Plutchik, R. (1980). *Emotion: A Psychoevolutionary Synthesis*. Harper & Row.

Rychlý, P. (2008). A Lexicographer-Friendly Association Score. In P. Sojka, A. Horák (Eds.), *Proceedings of Recent Advances in Slavonic Natural Language Processing 2008* (pp. 6–9). Masaryk University.

Satake, Y. (2022). The Effects of Corpus Use on L2 Collocation Learning. *The JALT CALL Journal*, 18(1), 34–53. https://doi.org/10.29140/jaltcall.v18n1.520

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.

Stubbs, M. (2001). *Words and Phrases: Corpus Studies of Lexical Semantics*. Blackwell. The Corpus of Contemporary American English (COCA). https://www.english-corpora.org/coca/

Zhang, F., Chen, J., Tang, Q., Tian, Y. (2024). Evaluation of Emotion Classification Schemes in Social Media Text: An Annotation-Based Approach, *BMC Psychology*, 12(1), 503. https://doi.org/10.1186/s40359-024-02008-w

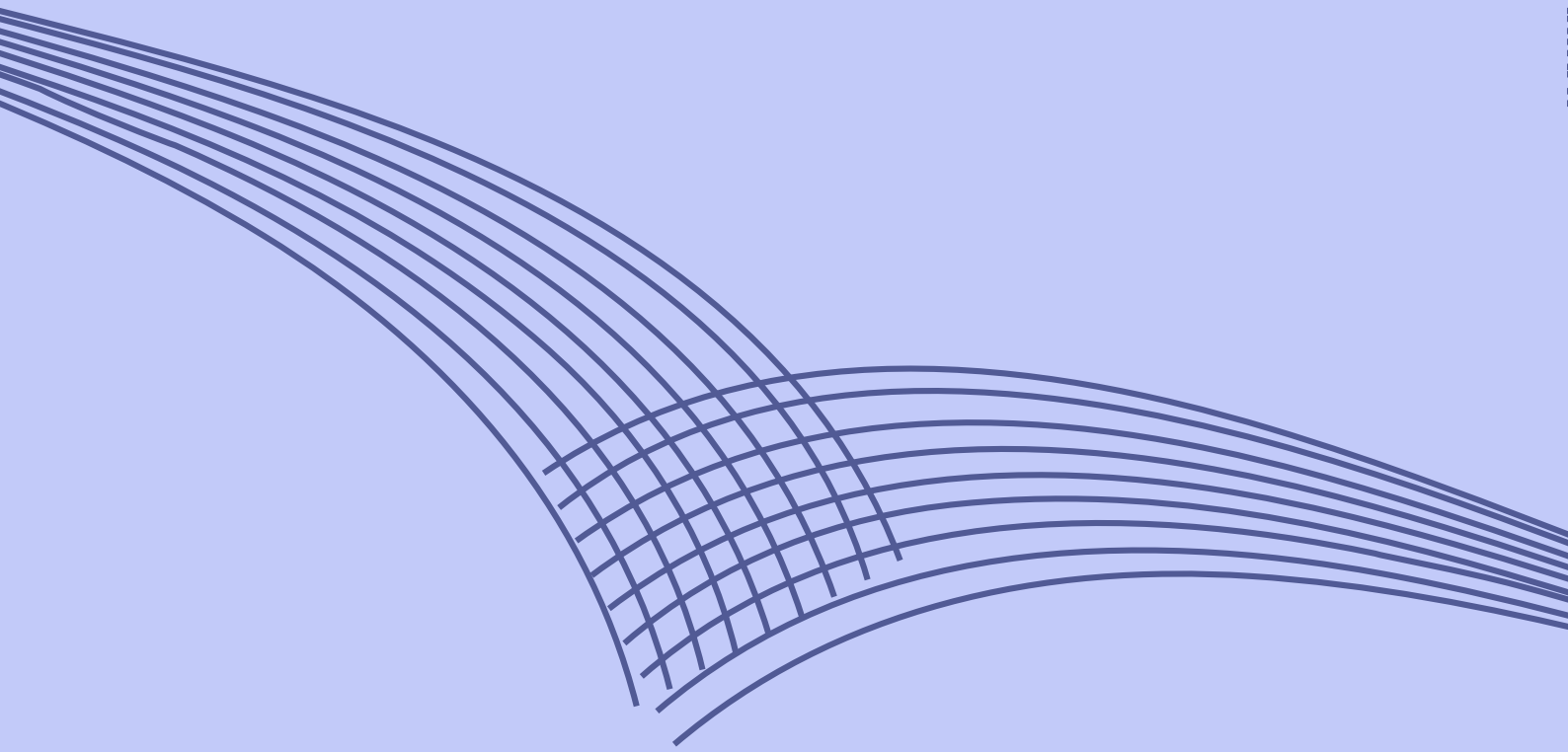
Contact information

Yelena Yerznkyan
Yerevan State University
yerznkyan@ysu.am

Diana Movsisyan
Armenian State University of Economics
movsisyan.diana@gmail.com

DIGITAL LEXICOGRAPHY: DESIGN, DATA MODELLING & THEORY

EURALEX 2026



Rukiye Smilla Burkart & Susanne Kabatnik

WIE DEFINIERT MAN DAS UNBEKANNTE?

Prompt Engineering als Methode der KI-gestützten Bedeutungsdefinition im Pandemictionary

Abstract The formulation of definitions is a core task of lexicography, traditionally carried out manually. Large language models now offer new opportunities to automate this workflow. While AI-assisted definition generation has been well studied for well-known vocabulary, this is not yet the case for lesser-known historical vocabulary, which is scarcely represented in training data. This paper therefore investigates how far targeted prompt engineering can improve the quality of AI-generated definitions for historical pandemic vocabulary, using data from the cholera corpus developed within the *Pandemictionary* project. Eight headwords of varying frequency and semantic complexity are analysed across three prompt configurations, focusing on the role of corpus evidence in generating lexicographical definitions. The results show that AI systems can generate formally correct, lexicographically structured definitions. Prompts without evidence draw on general model knowledge and often produce anachronistic or hallucinated content, whereas evidence-based prompts yield definitions that are more historically accurate and firmly grounded in the corpus. Corpus evidence proves a crucial quality factor, especially for rare or poorly documented lexemes, though challenges remain regarding hallucinations, polysemy, and reducing definitions to their essential elements. The study highlights the potential of prompt engineering for digital lexicography while underscoring the continuing central role of human expertise in quality assurance.

Keywords Prompt engineering; digital lexicography; definition of meaning; historical pandemic vocabulary; AI

1 Einleitung

Die Formulierung von Bedeutungsdefinitionen stellt eine zentrale Aufgabe einsprachiger Wörterbücher dar (Viehweger, 1982, S. 147). Traditionell erfolgt dieser Prozess manuell, ist jedoch zeitaufwändig und unterliegt einer potenziell subjektiven Prägung. Vor diesem Hintergrund wurden in den letzten zwei Jahrzehnten zunehmend (semi-)automatisierte Verfahren entwickelt (Rundell & Kilgarriff, 2011). Mit der breiten Verfügbarkeit KI-gestützter Anwendungen wie ChatGPT eröffnen sich neue Möglichkeiten zur Automatisierung lexikografischer Arbeitsprozesse, insbesondere im Bereich der Bedeutungsdefinition (Pham et al., 2025). Der vorliegende Beitrag untersucht, wie sich durch gezieltes Prompt Engineering qualitativ hochwertige Bedeutungsdefinitionen für historischen Wortschatz generieren lassen – also für Lexeme, die in den Trainingsdaten großer Sprachmodelle kaum repräsentiert sind. Der Beitrag gliedert sich wie folgt: Nach einem Überblick über Forschungsstand und Qualitätskriterien (Abschnitt 2) werden Daten und Methode (Abschnitt 3) vorgestellt, gefolgt von der Analyse (Abschnitt 4) und einem Fazit (Abschnitt 5). Die Analyse verfolgt drei zentrale Ziele: Erstens wird der Einfluss verschiedener Prompt-Strategien auf die Qualität, Präzision und Kontextangemessenheit der generierten Bedeutungsdefinitionen bestimmt. Zweitens werden typische Fehlertypen sowie grundlegende Grenzen der KI im Umgang mit historischem Pandemiewortschatz identifiziert. Drittens werden auf dieser Grundlage Empfehlungen für den gezielten, regelbasierten Einsatz von Prompt Engineering in der digitalen Lexikografie entwickelt.

2 Forschungsstand

Aktuelle Studien zeigen, dass KI-generierte Bedeutungsdefinitionen bereits ein hohes Qualitätsniveau erreichen können. Sie sind häufig flüssig formuliert, entsprechen lexikografischen Konventionen (Lew, 2023) und sind teilweise mit menschlich erstellten Definitionen vergleichbar (Pham et al., 2025, S. 4; Lew, 2023, 2024). Herausforderungen bestehen jedoch weiterhin bei seltenen, archaischen oder historisch gewandelten Lexemen (De Schryver, 2023; Lew, 2024) sowie bei Polysemie, Metaphern und Idiomen (Geyken, 2024; Kabatnik & Burkart, 2025, S. 174), wodurch Halluzinationen und Falschinformationen entstehen können. Als Bewertungsrahmen dienen die im Pandemictionary-Projekt entwickelten Kriterien für Bedeutungsdefinitionen (Kabatnik & Klee, 2025, S. 136–138). Sie umfassen sechs Kategorien und bilden die Grundlage der vorliegenden Untersuchung:

Orthografische Kriterien (1) betreffen eine normgerechte und konsistente Schreibweise des Lemmas sowie die einheitliche Angabe lexikografischer Informationen (z. B. Genus oder zeitliche Einordnung).

Syntaktische Kriterien (2) verlangen grammatisch korrekte, vollständige Definitionen in etablierten Definiensmustern (z. B. “Ein X ist . . . ”), die dem Substituierbarkeits- bzw. Äquivalenzprinzip folgen, sodass Definiendum und Definiens semantisch gleichwertig sind (Wiegand, 1994; Mel’čuk & Polguère, 2018). Stilistische und textlinguistische Kriterien (3) fordern verständliche, prägnante und neutrale Formulierungen. Definitionen sollen nur die notwendigen Bedeutungsmerkmale enthalten (*Minimal Decomposition Principle*; s. Mel’čuk & Polguère, 2018), Fachsprache möglichst vermeiden und Synonyme lediglich ergänzend verwenden (Mel’čuk & Polguère, 2018). Semantische Kriterien (4) beziehen sich auf eine eindeutige Erfassung der denotativen Bedeutung. Definitionen folgen idealerweise der Struktur *genus proximum* und *differentia specifica*, berücksichtigen Polysemie und semantische Relationen und vermeiden Zirkelschlüsse oder rein synonymische Umschreibungen (Moerdijk, 2003; Mel’čuk & Polguère, 2018). (Korpus-)pragmatische Kriterien (5) verlangen eine empirische Fundierung auf Basis von Korpusdaten. Dazu gehören insbesondere Kollokationen, Kookkurrenzen, authentische Belege sowie Angaben zu Register, Stil oder fachsprachlichen Verwendungen (Kilgarriff, 2006; Geyken & Lemnitzer, 2016). Lexikografische Kriterien und Funktionen (6) umfassen übergeordnete Gestaltungsprinzipien lexikografischer Definitionen. Nach Mel’čuk und Polguère (2018) sind dies die minimale semantische Dekomposition, die hierarchische Strukturierung von Bedeutungsmerkmalen sowie deren Einbindung in semantische Netzwerke.

Bedeutungen sind systematisch anzuordnen (z.B. logisch-semantisch, nach Häufigkeit oder historisch; Moerdijk, 2003, S. 285; Kipfer, 1984, S. 101; Werner, 1989) und klar zu gliedern bzw. zu nummerieren bei mehreren Bedeutungen (Geeraerts, 2003). Synonyme erfüllen eine unterstützende Funktion innerhalb der Mikrostruktur, ersetzen jedoch nicht die eigentliche Definition. Insgesamt müssen Definitionen hinreichend häufig und unzureichend vorhersehbar sein (SFIP-Prinzip: ”sufficiently frequent, insufficiently predictable“; s. Kilgarriff, 2006).

Prompt Engineering – die gezielte Gestaltung von Eingabeaufforderungen zur Steuerung von LLM-Outputs (Sahoo et al., 2024; Marvin et al., 2024) – hat sich als eigenständiges Forschungsfeld etabliert. Techniken wie Few-Shot- oder Chain-of-Thought-Prompting werden in verschiedenen Anwendungsbereichen eingesetzt (Sahoo et al., 2024). Auch in den Geisteswissenschaften zeigen Studien, dass gezielt gestaltete Prompts die Erschließung und Analyse historischer Textkorpora unterstützen (Liu et al., 2025). Für die (historische) Lexikografie ist ihr Einsatz bislang jedoch kaum untersucht. Erste Arbeiten belegen, dass strukturierte Prompts die Qualität KI-generierter Definitionen, insbesondere bei seltenen oder

historischen Lemmata, deutlich verbessern können (De Schryver, 2023; Jürviste & Jakobson, 2025). Daran anknüpfend untersucht die vorliegende Studie systematisch, welche Prompt-Konfigurationen für die Definition historischen Pandemiewortschatzes besonders geeignet sind.

3 Daten und Methode

Die vorliegende Untersuchung ist im Rahmen des Wörterbuchprojekts zum historischen Pandemiewortschatz ("Pandemictionary") am Trier Center for Digital Humanities (TCDH) angesiedelt, gefördert durch die Forschungsinitiative Rheinland-Pfalz und Teil des Forschungsverbunds *Linked Open Data in den Geisteswissenschaften* (LODinG). Als Datengrundlage dienen ein historisches Cholera-Korpus sowie das darauf aufbauende Cholera-Glossar (Kabatnik & Klee, 2024). Das Korpus umfasst Daten überwiegend aus dem 18. bis 20. Jahrhundert und basiert auf historischen Wörterbüchern, Sekundärliteratur und zwei digitalen Referenzkorpora (DWDS, ca. 1,53 Mrd. Tokens; HIST-gesamt des IDS, ca. 70 Mio. Tokens). Mithilfe der Textanalyseplattform Sketch Engine wurden Frequenzanalysen, Keyword-Erkennungen, Wortbildungsmuster und Kollokationen ausgewertet, woraus rund 1.300 Stichwörter gewonnen wurden (Kabatnik & Klee, 2026, S. 128). Für die vorliegende Analyse wurden acht Lemmata ausgewählt, die hinsichtlich Korpusfrequenz und semantischer Komplexität variieren (Kabatnik & Burkart, 2025, S. 179; Kabatnik & Klee, 2025):

TABELLE 1 Übersicht Lemma-Auswahl

Begriff	Treffer	Auswahlgrund
Cholera	8.093	hohe Frequenz, Kern des Wortschatzes
Cholerakranke	312	monosem, hohe Frequenz
Choleraklette	1	sehr geringe Frequenz, Grenzfall, Bedeutung erst im Kontext erschließbar
Ferkelcholera	4	geringe Frequenz, Fachvokabular, Bedeutung erst im Kontext erschließbar
Cholerablut	17	geringe Frequenz, Kompositum, Bedeutung erst im Kontext erschließbar
Choleratropfen	12	geringe Frequenz, Pharmazievokabular, Bedeutung erst im Kontext erschließbar
Wasser	599	allgemeinsprachlich, ambig, hohe Frequenz
Cholerazeit	173	temporal, mittlere Frequenz

Mittels gezielten Prompt Engineerings (Marvin et al., 2024) wird eine Methode zur Generierung von Bedeutungsdefinitionen für Lexeme entwickelt, die in den Trainingsdaten großer Sprachmodelle nicht oder unzureichend repräsentiert sind (De Schryver, 2023). Durch klar formulierte Rollen, Zielvorgaben und Bewertungsmaßstäbe lässt sich die Qualität KI-generierter Definitionen gezielt verbessern (Lew, 2023; De Schryver, 2023). Die methodische Grundlage bilden experimentelle Studien mit ChatGPT-5, in denen drei Prompt-Konfigurationen miteinander verglichen werden (Sahoo et al., 2024): In einem ersten Schritt wird die KI mit den zuvor festgelegten Kriterien (s. Abschnitt 2) angereichert: *Ich gebe dir folgende Kriterien für eine gute Bedeutungsdefinition, die du dir bitte merken sollst: [. . .]*

Prompt a (ohne Rolle, ohne Belege, ”Zero-Shot“-Prompts):[1] *Du sollst eine Bedeutungsdefinition nach den dir bekannten Kriterien für eine gute Bedeutungsdefinition zum Lexem [xxx] formulieren. Ziel ist es, ein digitales Wörterbuch zu historischen Pandemien zu erstellen.*

Prompt b (mit Rolle, ohne Belege; ”Zero-Shot“-Prompts): *Stelle dir vor, du bist Lexikograf und sollst eine Bedeutungsdefinition nach den dir bekannten Kriterien für eine gute Bedeutungsdefinition zum Lexem [xxx] formulieren. Ziel ist es, ein digitales Wörterbuch zu historischen Pandemien zu erstellen.*

Prompt c (mit Rolle und Belegen, ”Few-Shot“-Prompts): *Stelle dir vor, du bist Lexikograf und sollst aus den folgenden Belegen die Bedeutungsdefinition nach den dir bekannten Kriterien für eine gute Bedeutungsdefinition zum Lexem [xxx] formulieren. Beziehe dabei auch die angeführten Belege mit ein, die wir dir aus unserem Cholera-Korpus zur Verfügung stellen. Ziel ist es, ein digitales Wörterbuch zu historischen Pandemien zu erstellen. [1. . . . 2. . . . 3. . . . 4. . . . 5. . . .]*

Lediglich die Ergebnisse der Prompts b und c werden im Folgenden verglichen, da Prompt a als Baseline (ohne Rolle) dient und sich in den Ergebnissen nicht wesentlich von Prompt b unterscheidet. Die in Prompt c (”Few-Shot“-Prompts) verwendeten Belege wurden zuvor von einer studentischen Hilfskraft – nach den im Rahmen des Wörterbuchprojekts Pandemictionary definierten Kriterien zur Belegauswahl – zusammengestellt (Kabatnik & Klee, 2025; Kabatnik & Burkart, 2025).

4 Analyse

Im Mittelpunkt der Analyse steht der Vergleich der durch unterschiedliche Prompt-Konfigurationen erzeugten Definitionen anhand der in Abschnitt 2 entwickelten Kriterien. Die ausgewählten Lemmata unterscheiden sich hinsichtlich Korpusfrequenz, semantischer Komplexität und Spezialisierungsgrad und ermöglichen so eine differenzierte Bewertung der Leistungsfähigkeit und Grenzen KI-generierter Definitionen. Da alle acht Lemmata in beiden Varianten die orthografischen und syntaktischen Kriterien erfüllen, werden im Folgenden ausschließlich inhaltliche Unterschiede diskutiert.

4.1 Cholera

Das Lemma *Cholera* weist mit 8.093 Treffern die höchste Korpusfrequenz im Datensatz auf und stellt damit einen günstigen Ausgangspunkt für die KI dar.

Zwischen beiden Prompt-Varianten zeigen sich deutliche Unterschiede. Die Definition ohne Belege orientiert sich am heutigen medizinischen Wissen, indem sie Erreger, Übertragungsweg und Symptome beschreibt. Für ein historisches Wörterbuch ist dies jedoch nur eingeschränkt geeignet, da die zeitgenössische Krankheitsperspektive unberücksichtigt bleibt und einzelne Angaben historisch überprüft werden müssten. Die Definition mit Belegen fokussiert dagegen die historische Wahrnehmung der Cholera. Sie betont den raschen, oft tödlichen Verlauf, gesellschaftliche Ängste sowie zeitgenössische Bezeichnungen wie *Cholera morbus, Gallenruhr* oder *asiatische Cholera*. Zudem übernimmt sie sprachliche Elemente und Kollokationen (z. B. *furchtbare Cholera*) direkt aus den Belegen, während die Kollokationen

[1] Wird in der Analyse (Abschnitt 4) nicht weiter aufgenommen, da sich die Ergebnisse von Prompt a nicht wesentlich von jenen von Prompt b unterscheiden. Der Fokus liegt daher auf dem Vergleich der Ergebnisse von Prompt b und c.

der beleglosen Version offenbar auf allgemeinem Modellwissen beruhen. Beide Definitionen sind für eine klassische Bedeutungsdefinition zu umfangreich, da sie periphere Bedeutungsmerkmale und Gebrauchshinweise ausführlich integrieren. Polysemie wird in keiner der beiden Varianten berücksichtigt.

4.2 Cholerakranke

Beim Lemma *Cholerakranke* stimmen beide Varianten in der Kerndefinition überein: Beide bezeichnen Personen, die an Cholera erkrankt sind. Unterschiede ergeben sich aus der jeweiligen Informationsgrundlage. Die beleglose Definition ergänzt allgemeines medizinisches Wissen und nennt Symptome wie Durchfall, Erbrechen und Austrocknung. Die belegbasierte Definition orientiert sich dagegen eng an den Belegen und beschreibt *Cholerakranke* vor allem als Gegenstand medizinischer Behandlung und behördlicher Maßnahmen (z. B. Lazarett, Isolierung, Desinfektion, zeitgenössische Therapien), während Symptome weitgehend fehlen. Dies verdeutlicht den zentralen Unterschied beider Prompt-Varianten: Die belegbasierte Version paraphrasiert primär die Beleginhalte und greift deutlich seltener auf allgemeines Modellwissen zurück. Dadurch entstehen historisch besser fundierte, zugleich aber potenziell unvollständige Definitionen, wenn die Belege nur Teilaspekte eines Lemmas abdecken.

4.3 Choleraklette

Das Lemma *Choleraklette* bildet mit nur einem Beleg den methodisch anspruchsvollsten Fall des Datensatzes. Nach dem SFIP-Prinzip (Kilgarriff, 2006) wäre ein solches *Hapax legomenon* zwar grundsätzlich nicht wörterbuchwürdig, bei historischen Korpora lässt sich die tatsächliche Gebrauchshäufigkeit jedoch aufgrund der lückenhaften Überlieferung kaum bestimmen. Ohne Beleg erzeugt die KI eine Halluzination und definiert *Choleraklette* fälschlich als amtliches Kennzeichen während einer Choleraepidemie. Mit Beleg beschreibt sie das Lemma hingegen korrekt als Pflanze, der heilende Eigenschaften gegen Cholera und Tollwut zugeschrieben wurden. Die KI übernimmt dabei die im Beleg enthaltene implizite Definition und paraphrasiert sie. Der Befund bestätigt, dass beleglose KI-Definitionen bei seltenen historischen Lexemen besonders fehleranfällig sind (De Schryver, 2023; Lew, 2024). Belege sind hier eine wesentliche Voraussetzung für inhaltlich belastbare Definitionen, ersetzen jedoch die lexikografische Prüfung nicht.

4.4 Cholerablut

Cholerablut bezeichnet in der Medizinsprache des 19. Jahrhunderts das Blut von Cholerakranken, dem eine krankhaft veränderte Zusammensetzung zugeschrieben wurde. Beide Prompt-Varianten definieren das Lemma übereinstimmend als pathologisch verändertes Blut von Cholerakranken. Unterschiede zeigen sich jedoch im Detail: Ohne Belege bezieht die KI auch Verstorbene ein und beschreibt allgemein eine gestörte Zusammensetzung oder Funktion des Blutes. Mit Beleg beschränkt sie sich auf Erkrankte, da der Beleg ausschließlich Blutuntersuchungen an Kranken behandelt. Dies verdeutlicht, dass belegbasierte Definitionen bei einseitiger Beleglage inhaltlich verengt sein können. Gleichzeitig liefert die belegbasierte Variante deutlich präzisere historische Informationen, indem sie den Mangel an Faserstoff (Fibrin) sowie einen erhöhten Färbestoffgehalt nennt. Diese unmittelbar aus dem Beleg übernommenen Merkmale spiegeln die zeitgenössische medizinische Wissensordnung authentischer wider und sind daher für historische Wörterbücher besonders wertvoll.

TABELLE 2 *Cholera*

Lexem: Cholera Treffer: 8.093	
Bedeutungsdefinition Prompt b	Bedeutungsdefinition Prompt c (5 Belege)
Cholera, *f.* (Medizin; historische und aktuelle Pandemien) **1.** *Schwere bakterielle Infektionskrankheit des Menschen*, die durch das Bakterium Vibrio cholerae verursacht wird und sich meist über mit Erregern verunreinigtes Trinkwasser oder Nahrungsmittel überträgt, **gekennzeichnet durch** plötzlich einsetzenden, massiven wässrigen Durchfall und Erbrechen, die unbehandelt rasch zu starkem Flüssigkeits- und Elektrolytverlust, Austrocknung und zum Tod führen können. *Periphere semantische Komponenten:* – Erreger: gramnegatives Bakterium (*Vibrio cholerae*) – Übertragungsweg: fäkal–oral, v. a. über Wasser – Verlauf: akut, potenziell lebensbedrohlich – Historischer Kontext: Ursache mehrerer weltweiter Pandemien seit dem 19. Jahrhundert, besonders in Regionen mit mangelhafter Wasserversorgung und Hygiene *Typische Kollokationen:* **an Cholera erkranken, Cholera-Epidemie, Cholera-Pandemie, Cholera-Ausbruch, Cholera-Tote** *Gebrauchshinweis:* In historischen Texten häufig als Prototyp einer verheerenden Seuche mit hoher Sterblichkeit verwendet; in modernen Kontexten medizinisch klar definiert und gut behandelbar, sofern frühzeitig therapiert.	**Cholera**, *f.* (Medizin; Seuchengeschichte; historisch) **1.** *Schwere, hochgefährliche Infektionskrankheit des Menschen*, die sich epidemisch oder pandemisch ausbreitet und **durch einen raschen, oft tödlichen Krankheitsverlauf** gekennzeichnet ist, **wobei** die Erkrankung plötzlich einsetzt und innerhalb kurzer Zeit – teils innerhalb weniger Stunden – zum Tod führen kann, **sodass** sie in den betroffenen Gesellschaften große Angst, Verzweiflung und tiefgreifende soziale Reaktionen hervorruft. *Zentrale semantische Komponente (genus proximum):* – Krankheit/Seuche des Menschen *Periphere semantische Komponenten (differentiae specificae, historisch motiviert):* – Verlauf: akut, heftig, häufig tödlich – Wirkung: rasche körperliche Schwächung, Zusammenbruch lebenswichtiger Funktionen – Wahrnehmung: lange Zeit als unbekannte, rätselhafte Krankheit verstanden – Historischer Ursprung (zeitgenössische Zuschreibung): aus Asien stammend (“asiatische Cholera“) – Gesellschaftliche Folgen: Angst, moralische Verunsicherung, Unterstützung oder Akzeptanz staatlicher Zwangsmaßnahmen – Medizinische Deutung (historisch): Störung innerer Körperfunktionen, insbesondere von Atmung, Blut und Herz, als angenommene Todesursache *Historische Bezeichnungen/Synonyme:* **Cholera morbus, Gallenruhr, Brechruhr, asiatische Cholera** (historisch) *Typische Kollokationen (Korpusbezug):* **furchtbare Cholera, Auftreten der Cholera, an Cholera sterben, Cholera wütet, Cholera-Seuche, Angst vor der Cholera** *Gebrauchshinweis:* In historischen Texten erscheint *Cholera* nicht nur als medizinische Bezeichnung, sondern auch als Chiffre für eine unkontrollierbare, bedrohliche Seuche mit hoher Sterblichkeit, deren Ursachen lange unklar blieben und die tief in das soziale und moralische Gefüge der Gesellschaft eingriff.

TABELLE 3 *Cholerakranke*

Lexem: Cholerakranke Treffer: 312	
Bedeutungsdefinition Prompt b	Bedeutungsdefinition Prompt c (5 Belege)
Cholerakranke, *Pl.* (historisch; Medizin; Seuchengeschichte) **1.** *Personen, die an Cholera erkrankt sind,* **bei denen** sich die Krankheit durch schwere, akute Symptome äußert und **die** im Kontext historischer Epidemien als Träger einer hochgefährlichen, oft tödlich verlaufenden Seuche galten, **weshalb** sie Gegenstand medizinischer Behandlung, hygienischer Maßnahmen und behördlicher Kontrolle waren. *Zentrale semantische Komponente (genus proximum):* – erkrankte Personen/Kranke *Periphere semantische Komponenten (differentiae specificae):* – Krankheitsbezug: Cholera – Symptomatik (zeitgenössisch): starker Durchfall, Erbrechen, rascher Kräfteverfall, Austrocknung – Verlauf: häufig plötzlich einsetzend, oft lebensbedrohlich – sozialer Status: isoliert, überwacht oder besonders behandelt – hygienisch-administrativer Kontext: Quarantäne, Desinfektion von Kleidung und Umgebung, spezielle Pflege – gesellschaftliche Wahrnehmung: Quelle von Angst, Mitgefühl oder Stigmatisierung *Typische Kollokationen:* **Cholerakranke pflegen, Kleidung der Cholerakranken, Cholerakranke isolieren, Behandlung der Cholerakranken, Ausleerungen der Cholerakranken** *Gebrauchshinweis:* Das Lexem *Cholerakranke* wird in historischen Quellen meist kollektiv gebraucht und verweist weniger auf individuelle Krankengeschichten als auf eine epidemiologisch relevante Gruppe. Es spiegelt die enge Verknüpfung von Krankheit, öffentlicher Hygiene und staatlicher Seuchenpolitik im 19. Jahrhundert wider.	**Cholerakranke**, *Pl.* (historisch; Medizin; Seuchengeschichte) **1.** *Personen, die an Cholera erkrankt sind,* **bei denen** die Krankheit akut und häufig lebensbedrohlich verläuft und **die** im Kontext historischer Epidemien als medizinisch zu behandelnde und zugleich hygienisch zu kontrollierende Gruppe galten, **weshalb** sie transportiert, isoliert, untersucht und besonderen sanitätspolizeilichen Maßnahmen unterworfen wurden. *Zentrale semantische Komponente (genus proximum):* – erkrankte Personen/Kranke *Periphere semantische Komponenten (differentiae specificae, korpusbasiert):* – Krankheitsbezug: Cholera – Verlauf: heftig, oft tödlich, teils Tod während des Transports – medizinische Behandlung (zeitgenössisch): verdünnende Getränke, wässerige Einspritzungen, Salzzusätze; zuvor auch Blutentziehungen – wissenschaftlicher Status: Objekt chemischer und physiologischer Untersuchungen (z. B. Blutanalysen) – institutioneller Kontext: Unterbringung in Hospitälern und eigens vorbereiteten Lazaretten – hygienisch-administrative Maßnahmen: sofortige Isolierung, Desinfektion von benutzten Anlagen, Vermeidung von Krankheitsherden – gesellschaftliche Funktion: potenzielle Quelle weiterer Ansteckung im öffentlichen Raum *Typische Kollokationen (korpusnah):* **Cholerakranke transportieren, Cholerakranke behandeln, Blut der Cholerakranken, Cholerakranke isolieren, Benutzung durch Cholerakranke, Lazarette für Cholerakranke** *Gebrauchshinweis:* Das Lexem *Cholerakranke* wird in historischen Texten überwiegend kollektiv verwendet und bezeichnet weniger individuelle Patienten als eine epidemiologisch relevante Gruppe. Die Belege zeigen Cholerakranke zugleich als Leidtragende der Krankheit, als Gegenstand medizinisch-experimenteller Therapieansätze und als Fokus staatlicher Hygiene-, Isolations- und Kontrollmaßnahmen.

TABELLE 4 *Choleraklette*

Lexem: Choleraklette Treffer: 1	
Bedeutungsdefinition Prompt b	Bedeutungsdefinition Prompt c (5 Belege)
Choleraklette, *f.* (historisch; Seuchengeschichte) **1.** *Amtlich ausgegebenes Kennzeichen*, das während Choleraepidemien an Personen, Gebäuden oder Gegenständen angebracht oder mitgeführt wurde, **um** eine tatsächliche oder vermutete Verbindung zur Cholera kenntlich zu machen, **insbesondere** zur Markierung von Erkrankten, Genesenen, Quarantänefällen oder infizierten Haushalten und zur Durchsetzung behördlicher Kontroll- und Schutzmaßnahmen. *Zentrale semantische Komponente (genus proximum):* – Kennzeichen/Abzeichen mit behördlicher Funktion *Periphere semantische Komponenten (differentiae specificae):* – Funktion: Sichtbarmachung von Seuchengefahr, Warnung und Kontrolle – Gebrauchskontext: Choleraepidemien, vor allem im 19. Jahrhundert – Träger/Objekt: Personen, Häuser oder Hausstände – Sozial-pragmatische Wirkung: Stigmatisierung, soziale Ausgrenzung, erleichterte Überwachung durch Behörden – Material/Form (variabel): kleines Zeichen, Abzeichen oder sichtbare Markierung (z. B. aus Stoff, Metall oder Papier) *Typische Kollokationen (historisch belegt):* **eine Choleraklette tragen, mit einer Choleraklette versehen, Häuser mit Choleraketten kennzeichnen, Anordnung der Choleraklette** *Gebrauchshinweis:* Das Lexem *Choleraklette* gehört zum historischen Verwaltungsvokabular der Seuchenbekämpfung und bezeichnet keine medizinische Maßnahme, sondern ein symbolisches und administratives Mittel zur Regelung des öffentlichen Lebens während Choleraausbrüchen.	**Choleraklette**, *f.* (historisch; Volksmedizin; Seuchengeschichte) **1.** *Pflanzliches Gewächs*, das nach zeitgenössischen Berichten erstmals im Zusammenhang mit einer Choleraepidemie erwähnt wird und **dem** in der Volksvorstellung eine heilende Wirkung gegen Cholera sowie gegen andere schwere Krankheiten, insbesondere Tollwut (*Hundswuth*), zugeschrieben wurde, **ohne** dass seine medizinische Wirksamkeit eindeutig belegt oder wissenschaftlich erklärt war. *Zentrale semantische Komponente (genus proximum):* – Pflanze/pflanzliches Gewächs *Periphere semantische Komponenten (differentiae specificae):* – historischer Kontext: Auftreten bzw. Bekanntwerden während einer Choleraepidemie – zugeschriebene Funktion: Heilmittel gegen Cholera und weitere lebensbedrohliche Krankheiten – Wissensstatus: Wirkung unklar, spekulativ oder mystifiziert – kulturelle Einordnung: Bestandteil volkstümlicher oder außerwissenschaftlicher Krankheitsdeutungen *Typische Kollokationen (korpusnah rekonstruierbar):* **die Choleraklette soll helfen, eine Choleraklette als Heilmittel, Auftreten der Choleraklette, Wirkung der Choleraklette** *Gebrauchshinweis:* Das Lexem *Choleraklette* bezeichnet in historischen Texten kein eindeutig identifizierbares botanisches Taxon, sondern ein im Kontext von Seuchenangst entstandenes Benennungsobjekt, das für den Versuch steht, neuartige oder unerklärliche Krankheiten durch vermeintlich natürliche Heilmittel verständlich und beherrschbar zu machen.

TABELLE 5 *Cholerablut*

Lexem: Cholerablut Treffer: 17	
Bedeutungsdefinition Prompt b	Bedeutungsdefinition Prompt c (5 Belege)
Cholerablut, *n.* (historisch; Medizin; Seuchengeschichte) **1.** *Blut von an Cholera erkrankten oder verstorbenen Menschen*, das in historischen medizinischen und naturwissenschaftlichen Kontexten **als Träger krankhafter Veränderungen** betrachtet und **zur Erklärung des Krankheitsverlaufs, der Todesursachen oder der Ansteckung** herangezogen wurde. *Zentrale semantische Komponente (genus proximum):* – Blut (Körperflüssigkeit des Menschen) *Periphere semantische Komponenten (differentiae specificae):* – Bezugskrankheit: Cholera – Zustand: als pathologisch verändert angesehen – medizinische Zuschreibung (historisch): gestörte Zusammensetzung oder Funktion des Blutes – Erkenntnisinteresse: Erklärung von Symptomen, Organversagen oder plötzlichem Tod – Wissensrahmen: vorkeimtheoretische bzw. frühe naturwissenschaftliche Krankheitsmodelle *Typische Kollokationen (historisch):* **Veränderungen des Cholerablutes, Untersuchung des Cholerablutes, Beschaffenheit des Cholerablutes, Zersetzung des Cholerablutes** *Gebrauchshinweis:* Das Lexem *Cholerablut* ist Teil der historischen Medizinsprache und spiegelt frühere Vorstellungen wider, nach denen Krankheiten als qualitative Veränderung von Körperflüssigkeiten verstanden wurden. Es bezeichnet keinen eigenständigen Stoff im modernen medizinischen Sinn, sondern eine konzeptuelle Einheit innerhalb älterer Krankheits- und Seuchenmodelle.	**Cholerablut**, *n.* (historisch; Medizin; Seuchengeschichte) **1.** *Blut von an Cholera erkrankten Menschen*, das in der zeitgenössischen medizinischen Forschung **als pathologisch veränderte Körperflüssigkeit** beschrieben wird und **sich durch eine auffällige Abweichung seiner Zusammensetzung vom gesunden Blut** auszeichnet, insbesondere durch einen Mangel an gerinnungsbildenden Bestandteilen und einen als stark erhöht beschriebenen Anteil an färbenden Substanzen. *Zentrale semantische Komponente (genus proximum):* – Blut (Körperflüssigkeit des Menschen) *Periphere semantische Komponenten (differentiae specificae, korpusbasiert):* – Krankheitsbezug: Cholera – chemisch-physiologische Zuschreibung (historisch): Mangel an Faserstoff/Fibrin – weitere angenommene Eigenschaft: ungewöhnlich hoher Gehalt an Färbestoff im Vergleich zu gesundem Blut – epistemischer Status: Ergebnis beobachtender, vorkeimtheoretischer Medizin – funktionale Deutung: Erklärung für gestörte Lebensfunktionen, Kreislaufversagen und Tod *Typische Kollokationen (korpusnah):* **Mangel an Fibrin im Cholerablut, Überschuss an Färbestoff im Cholerablut, Beschaffenheit des Cholerablutes, Analyse des Cholerablutes** *Gebrauchshinweis:* Das Lexem *Cholerablut* gehört zur historischen medizinischen Fachsprache und verweist auf frühere Krankheitsmodelle, in denen Cholera als tiefgreifende Veränderung der Blutbeschaffenheit interpretiert wurde. Die beschriebenen Eigenschaften spiegeln zeitgenössische Beobachtungen und Deutungen wider und entsprechen nicht der modernen pathophysiologischen Erklärung der Krankheit.

TABELLE 6 *Wasser*

Lexem: Wasser Treffer: 599	
Bedeutungsdefinition Prompt b	Bedeutungsdefinition Prompt c (5 Belege)
Wasser, *n.* (Naturstoff; Umwelt; Seuchengeschichte) **1.** *Farb- und geruchlose, flüssige chemische Verbindung*, die aus Wasserstoff und Sauerstoff besteht und **für Menschen, Tiere und Pflanzen lebensnotwendig ist,** in historischen Pandemien jedoch zugleich als zentrales Übertragungsmedium von Krankheiten, insbesondere der Cholera, wirkte, **wenn** es mit krankheitserregenden Stoffen verunreinigt war. *Zentrale semantische Komponente (genus proximum):* – natürliche Flüssigkeit/Grundstoff des Lebens *Periphere semantische Komponenten (differentiae specificae, pandemiebezogen):* – hygienischer Status: sauber/verunreinigt – epidemiologische Funktion (historisch): Träger und Verteiler von Krankheitserregern – Gebrauchskontext: Trinkwasser, Brunnenwasser, Flusswasser – gesellschaftliche Relevanz: Schlüsselrolle bei Ausbruch, Eindämmung oder Verschärfung von Seuchen – Erkenntnisgeschichte: lange Zeit unterschätzte Ursache von Epidemien; zentrale Rolle in der Entwicklung moderner Hygiene- und Seuchentheorien *Typische Kollokationen (pandemiehistorisch):* **verunreinigtes Wasser, Trinkwasser, Wasser als Infektionsquelle, Wasser übertragen, Wasser und Cholera, sauberes Wasser** *Gebrauchshinweis:* Im Kontext historischer Pandemien bezeichnet *Wasser* nicht nur einen natürlichen Stoff, sondern eine entscheidende epidemiologische Größe. Seine Qualität und Zugänglichkeit bestimmten maßgeblich Ausbreitung, Intensität und soziale Wahrnehmung von Seuchen und machten Wasser zu einem zentralen Ansatzpunkt staatlicher Gesundheitsmaßnahmen und moderner Prävention.	**Wasser**, *n.* (historisch; Hygiene; Medizin; Seuchengeschichte) **1.** *Natürliche, flüssige Substanz*, **die** für den menschlichen Gebrauch als Trink-, Haushalts- und medizinisches Hilfsmittel dient, **die** im Kontext historischer Pandemien eine **zentrale hygienische und epidemiologische Rolle** spielt, **indem** sie je nach Beschaffenheit Krankheiten übertragen, deren Ausbreitung begünstigen oder zu ihrer Eindämmung genutzt werden kann. *Zentrale semantische Komponente (genus proximum):* – Flüssigkeit/Naturstoff *Periphere semantische Komponenten (differentiae specificae, korpusbasiert):* – hygienischer Status: sauber, desinfiziert oder verunreinigt – epidemiologische Funktion: Übertragungsmedium für Cholera, Typhus und andere Krankheiten bei schlechter Qualität – Nutzungskontexte: – Trinkwasser und Haushaltswasser (Brunnen, Wasserleitungen) – Reinigungs- und Desinfektionsmittel (z. B. phenylsäurehaltiges Wasser zum Waschen von Kleidung Cholerakranker) – medizinische Anwendung (z. B. Einspritzung großer Mengen Wasser in die Blutadern bei Kranken) – physiologische Rolle: Hauptbestandteil der bei Cholera verlorenen Körperflüssigkeiten; Träger weiterer Blutbestandteile – präventive Bedeutung: Ausschluss unsicherer Brunnen, Sicherung zentraler Wasserleitungen *Typische Kollokationen (korpusnah):* **schlechtes Wasser, verunreinigtes Wasser, Trinkwasser, Haushaltungswasser, Wasserleitungen, Wasser einspritzen, Krankheiten durch Wasser verbreiten** *Gebrauchshinweis:* In historischen Texten zur Cholera bezeichnet Wasser nicht nur einen elementaren Naturstoff, sondern eine entscheidende Variable im Spannungsfeld von Krankheit, Therapie und Prävention. Die Belege zeigen Wasser zugleich als Gefahrenträger, als medizinisches Mittel und als Ansatzpunkt staatlicher Hygienemaßnahmen, lange bevor seine mikrobiologische Bedeutung vollständig verstanden war.

4.5 Wasser

Das Lemma *Wasser* ist das einzige allgemeinsprachliche Lemma im Datensatz und bildet damit einen wichtigen Kontrastfall. Beide Prompt-Varianten unterscheiden sich bereits in der Grunddefinition. Ohne Belege beschreibt die KI Wasser naturwissenschaftlich als chemische Verbindung aus Wasserstoff und Sauerstoff und erwähnt lediglich den Übertragungsweg der Cholera über verunreinigtes Wasser. Die Definition bleibt damit auf allgemeines Enzyklopädiewissen beschränkt. Mit Belegen wird Wasser dagegen als hygienisch und medizinisch relevante Substanz im Kontext historischer Pandemien definiert. Genannt werden u.a. Trink- und Haushaltswasser, phenylsäurehaltiges Wasser zur Desinfektion sowie therapeutische Wassereinspritzungen – Informationen, die direkt aus den Belegen stammen. Der Fall *Wasser* zeigt exemplarisch, dass Belege insbesondere bei allgemeinsprachlichen Lemmata notwendig sind, um deren historische und fachspezifische Bedeutungsdimension angemessen abzubilden.

4.6 Ferkelcholera, Choleratropfen, Cholerazeit

Die übrigen drei Lemmata bestätigen in kompakterer Form die bisherigen Tendenzen. Bei Ferkelcholera (4 Treffer, 1 Beleg) beschreibt die beleglose Definition die Krankheit als auf Ferkel beschränkt und ergänzt Symptome wie Fieber und Durchfall aus dem Modellwissen der KI. Die belegbasierte Definition übernimmt dagegen die historische terminologische Unschärfe: Ferkelcholera erscheint als synonyme Bezeichnung zu Schweinecholera bzw. Hog-Cholera, ohne eindeutige Eingrenzung auf Ferkel, dafür mit dem im Beleg genannten wirtschaftlichen Kontext der Tierverwertung – ein Mehrwert gerade für ein historisches Spezialwörterbuch.

Bei Choleratropfen (12 Treffer, 4 Belege) ergänzt die beleglose Version plausible, aber nicht korpusgestützte Angaben zur stofflichen Zusammensetzung (Alkohol, Opium, Kampfer) und zur therapeutischen Funktion. Mit Belegen entfallen diese Angaben; stattdessen rekonstruiert die KI historische Nutzungskontexte wie Dosierung, Handverkauf, Hausapotheken sowie das Missbrauchspotenzial als Träger von Giftstoffen. Die belegbasierte Definition ist damit nicht vollständiger, eröffnet aber eine historisch und pragmatisch reichhaltigere Perspektive.

Bei Cholerazeit (173 Treffer, 5 Belege) stimmen beide Varianten in der Kerndefinition weitgehend überein; die belegbasierte Version ergänzt lediglich die semantische Klasse ”öffentliche Hygiene” sowie konkrete historische Aspekte aus den Belegen, etwa Hausbesuche zur Trinkwasserkontrolle und den Hinweis, dass die Cholerazeit Giftmorde begünstigte, weil Cholerasymptome denen einer Arsenvergiftung ähnelten. Gleichzeitig verwendet die KI den Begriff epidemisch, obwohl er in keinem Beleg vorkommt – ein Beleg dafür, dass auch belegbasierte Definitionen Halluzinationen enthalten können und einer lexikografischen Prüfung bedürfen.

5 Fazit und Limitationen

Die Analysen zeigen, dass ChatGPT-5 in beiden Prompt-Varianten orthografisch und syntaktisch korrekte Definitionen mit *genus-proximum*-Struktur erzeugt (Kriterien 1–2). Der zentrale Unterschied liegt in der Informationsgrundlage: Beleglose Definitionen greifen auf allgemeines Modellwissen zurück und führen bei historischen Speziallemmata häufig zu anachronistischen oder halluzinierten Aussagen. Belegbasierte Definitionen sind historisch besser fundiert, können bei einseitiger Beleglage jedoch unvollständig oder semantisch verengt ausfallen (z.B. *Cholerablut*). Beide Varianten überschreiten den Umfang klassischer

Wörterbuchdefinitionen und berücksichtigen Polysemie nicht systematisch. Halluzinationen treten insbesondere bei geringer Belegdichte auf, sind aber auch in belegbasierten Definitionen nachweisbar.

Insgesamt besitzen KI-Systeme ein erhebliches Potenzial zur Unterstützung lexikografischer Arbeitsprozesse. Die Ergebnisse bestätigen, dass die Qualität der Definitionen mit der Korpusfrequenz eines Lemmas zunimmt (Pham et al., 2025): Während hochfrequente Lemmata konsistente Definitionen liefern, zeigen sich die Grenzen der KI vor allem bei seltenen historischen Lexemen. Die Einbindung von Belegen erweist sich dabei als zentraler Qualitätsfaktor, da sie historisch authentische Bedeutungsdimensionen erschließt und Halluzinationen reduziert, jedoch nicht vollständig verhindert.

Menschliche Expertise bleibt daher unverzichtbar, insbesondere bei seltenen oder polysemen Lemmata sowie der redaktionellen Verdichtung der Definitionen. Für die Praxis empfiehlt sich ein abgestuftes Vorgehen: (1) KI-generierte Definitionen können bei gut belegten Lemmata als Entwurf dienen und bedürfen einer oberflächlichen Überprüfung, (2) bei seltenen oder semantisch komplexen Lexemen müssen sie jedoch konsequent von Lexikograph:innen anhand der Belege überprüft werden.

Die Studie ist als Pilotstudie zu verstehen. Einschränkungen ergeben sich aus der geringen Lemmazahl, der begrenzten Reproduzierbarkeit von LLM-Ausgaben (Kabatnik & Burkart, 2025), der Bewertung durch die Autor:innen sowie der Beschränkung auf ChatGPT-5. Eine unabhängige Evaluation der Interrater-Reliabilität und Untersuchungen mit größeren Datensätzen und weiteren Modellen stellen wichtige Aufgaben zukünftiger Forschung dar.

Literatur

De Schryver, G.-M. (2023). Generative AI and lexicography: The current state of the art using ChatGPT. *International Journal of Lexicography*, 36(4), 355–387.

Geeraerts, D. (2003). Meaning and definition. In P. van Sterkenburg (Hg.), *A Practical Guide to Lexicography* (S. 83–93). John Benjamins.

Geyken, A. & Lemnitzer, L. (2016). Automatische Gewinnung von lexikografischen Angaben. In A. Klosa & C. Müller-Spitzer (Hgg.), *Internetlexikografie. Ein Kompendium* (S. 197–247). De Gruyter.

Geyken, A. (2024). Generative AI in lexicography/Generative KI in der Lexikographie. In *Workshop am 16. Mai 2024 in der Berlin-Brandenburgischen Akademie der Wissenschaften*.

Jürviste, M. & Jakobson, J. (2025). Vision-Enabled LLMs in Historical Lexicography: Digitising and Enriching Estonian-German Dictionaries from the 17th and 18th Centuries. *arXiv preprint arXiv:2510.07931*.

Kabatnik, S. & Klee, A. (2024). Historischer Pandemiewortschatz am Beispiel der Cholera – eine korpusbasierte quantitativ-qualitative Untersuchung Historischer Korpora. *Sprachwissenschaft*, 49 (2/3), 297–328.

Kabatnik, S. & Burkart, R. S. (2025). "Ich habe 5 geeignete Belege zum Lemma 'Cholera' aus der historischen Korpusdatei ausgewählt." – Experimentelle Untersuchung zur KI-gestützten Belegextraktion für die digitale Lexikografie. In V. Harm & U. Kramer (Hgg.): *Beispiel, Beleg, Quelle: Exemplifizieren, Dokumentieren und Verifizieren im Wörterbuch* (S. 171–199). De Gruyter.

Kabatnik, S. & Klee, A. (2025). ”Choleraausbruch zwang zur Rückkehr“ – die Auswahl von Beispielen für das Pandemictionary-Projekt als lexikografisches Linked Open Data-Netzwerk. In V. Harm & U. Kramer (Hgg.): *Beispiel, Beleg, Quelle: Exemplifizieren, Dokumentieren und Verifizieren im Wörterbuch* (S. 131–148). De Gruyter.

Kabatnik, S. & Klee, A. (2026). Historischer Pandemiewortschatz und Linked Open Data (LOD): Lexikalische Veränderungen und kulturelle Reaktionen im Spiegel der Cholera. *Online-only Publikationen des Leibniz-Instituts für Deutsche Sprache*, 16.

Kilgarriff, A. (2006): Word senses. In E. Agirre & P. Edmonds (Hgg.): *Word Sense Disambiguation: Algorithms and Applications* (S. 29–46). Springer.

Kipfer, B. A. (1984). Methods of ordering senses within entries. In R. R. K. Hartmann (Hg.), *LEXeter ’83 Proceedings* (S. 101–108). Max Niemeyer.

Lew, R. (2023). ChatGPT as a COBUILD Lexicographer. *Humanities and Social Sciences Communications*, 10, 1–10.

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, 11, 1–8.

Liu, J., Yang, C., Yan, Z., Ma, X. & Pei, L. (2025). Leveraging generative AI through prompt engineering for corpus construction and in-depth intelligent interpretation of ancient texts. *Digital Scholarship in the Humanities*, 40(3), 846–862.

Marvin, G., Hellen, N., Jjingo, D. & Nakatumba-Nabende, J. (2024). Prompt engineering in large language models. In I. Jeena Jacob, S. Piramuthu & P. Falkowski-Gilski (Hgg.), *Data Intelligence and Cognitive Informatics. Proceedings of ICDICI 2023* (S. 387–402). Springer.

Mel’čuk, I. & Polguère, A. (2018). Theory and Practice of Lexicographic Definition. *Journal of Cognitive Science*, 19(4), 417–470.

Moerdijk, F. (2003). The codification of semantic information. In P. van Sterkenburg (Hgg.), *A Practical Guide to Lexicography* (S. 273–296). John Benjamins.

Pham, B., Wong, J., Kim, S., Yin, Y. & Skiena, S. (2025). Word Definitions from Large Language Models. In *19th International Conference on Semantic Computing (ICSC)*. https://arxiv.org/abs/2311.06362

Rundell, M. & Kilgarriff, A. (2011). Automating the creation of dictionaries. Where will it all end? In F. Meunier, S. De Cock, G. Gilquin & M. Paquot (Hgg.), *A taste for corpora* (S. 257–282). John Benjamins.

Sahoo, P., Singh, A. K., Saha, S., Jain, V., Mondal, S. & Chadha, A. (2024). A systematic survey of prompt engineering in large language models: Techniques and applications. https://arxiv.org/abs/2311.06362

Viehweger, D. (1982). Semantiktheorie und praktische Lexikographie. *Zeitschrift für Germanistik*, 3(2), 143–155.

Werner, R. (1989). Probleme der Anordnung der Definitionen im allgemeinen einsprachigen Wörterbuch. In F. J. Hausmann, O. Reimann, H. E. Wiegand & L. Zgusta (Eds.), *Wörterbücher. Ein Internationales Handbuch der Lexikographie* (S. 917–931). De Gruyter.

Wiegand, H. E. (1994). Elements of a Theory towards a So–called Lexicographic Definition. *Lexicographica*, 8 (S. 175–289).

Contact information

Rukiye Smilla Burkart
University of Trier
burkart@uni-trier.de

Prof. Dr. Susanne Kabatnik
University of Trier
kabatnik@uni-trier.de

Lian Chen, Damien Nouvel, Huy-Linh Dao &
Alexander Delaporte

FROM CORPUS TO DICTIONARY

OntoLex-Lemon-Based Modelling of Neology in Chinese–French and Vietnamese–French Computational Lexicography

Abstract This paper presents NeoLex, a multilingual electronic lexical resource dedicated to contemporary Chinese–French and Vietnamese–French neology. Based on diachronic media corpora (2015–2025) and pre-2015 reference lexicons, the project combines corpus linguistics, Natural Language Processing, and computational lexicography to detect emerging lexical units. Candidate neologisms were extracted using Out-of-Vocabulary detection, Vietnamese n-gram extraction, and a character-based method for Chinese, then manually validated according to lexicographic criteria. The validated entries were modelled using OntoLex-Lemon, SKOS, and vartrans to represent lexical entries, lexical senses, translation equivalents between Chinese–French and Vietnamese–French, semantic domains, and diachronic metadata. A Streamlit-based prototype was developed to explore and visualize the resulting resources. NeoLex demonstrates how corpus-based neology detection and ontology-driven lexicography can contribute to interoperable multilingual electronic dictionaries.

Keywords Neology; computational lexicography; OntoLex-Lemon; multilingual lexical resources; semantic web; electronic dictionaries; Chinese; Vietnamese

1 Introduction

Neology, understood as the process of lexical innovation, constitutes one of the most visible manifestations of language evolution (Pruvost et al., 2003; Cabré, 2004; Boulanger, 2010; Fan, 2015; Gérard et al., 2016; Che, 2017; Bernal et al., 2020; Díaz Hormigo, 2020; Lombard, 2021; Poix, 2021; Đỗ Phương & Nguyễn, 2022; Huang, 2024). As societies evolve through technological progress, globalization, digital communication, and cultural exchange, languages continuously generate new lexical units to designate emerging realities. Neologisms may take various forms, including newly created words, semantic innovations, borrowings, acronyms, blends, and other lexical or grammatical innovations (Newmark, 1988; Neveu, 2011). Beyond their lexical dimension, they reflect broader social, cultural, political, and technological transformations and constitute important indicators of linguistic vitality (Crystal, 1996; Quemada, 1971; Boulanger, 2010).

The study of Chinese and Vietnamese neology is particularly relevant because both languages have undergone rapid lexical expansion in response to profound socio-economic and technological transformations. In Chinese, the term 新 词 (xīncí, “new word”) emerged around the nineteenth century and became especially prominent after the Reform and Opening-Up period initiated in 1978. Contemporary Chinese neologisms (当代新词语) generally refer to lexical items that appeared after the late 1970s, reflecting modernization, digital culture, globalization, and social change. Lü Shuxiang (1984) was among the first scholars to call for a systematic study of Chinese neologisms in 1984, and subsequent decades saw the publication of numerous specialized dictionaries, including the Xinhua Dictionary of New Words (新 华 新 词 语 词 典, 2003). Chinese neology is characterized by several productive mechanisms, including compounding, homophony, abbreviations, semantic shifts, Internet slang, and borrowings from English, Japanese, and regional dialects (Jing-Schmidt, 2019). Examples such as

卷王 (juǎnwáng, "king of involution"), 凡尔赛 (fán'ěrsài, "humblebragging"), or 996, ICU illustrate the close relationship between lexical creativity and contemporary Chinese digital culture.

Vietnamese neology presents a similarly dynamic yet structurally distinct phenomenon. The Vietnamese expression từ mới ("new words"), according to Nguyễn Thiện Giáp (2008) and Cao Xuân Hạo (2002), refers to lexical units recently introduced into the language in order to designate new communicative or social realities. Vietnamese neologisms result from multiple sources: endogenous lexical creation, Sino-Vietnamese influence, French colonial heritage, and contemporary English borrowings. Words such as cà phê ("coffee"), bê tông ("concrete"), or hót ("hot, trendy") demonstrate how foreign lexical items are phonologically and morphologically adapted into Vietnamese. At the same time, Internet culture and youth slang have become major drivers of lexical innovation, generating highly expressive forms such as thả thính ("to flirt"), trẻ trâu ("immature youth"), or bó tay ("to give up helplessly"). These neologisms reflect not only linguistic creativity but also the influence of social media, globalization, and generational identity.

From the perspective of Natural Language Processing (NLP), Chinese and Vietnamese neology also present important computational challenges (Dinh, 2001; Le-Hong, 2015; Kang, 2002; Pan, 2021). Unlike inflectional Western languages, Chinese and Vietnamese are largely isolating languages characterized by high lexical productivity, compounding, and flexible semantic innovation. Consequently, tasks such as word segmentation, neologism detection, lexical normalization, and semantic interpretation require language-specific computational approaches (Drouin, 2021). While European projects such as ENEOLI Cost Action aim to develop multilingual resources and tools for lexical innovation, Chinese and Vietnamese remain relatively under-resourced languages in terms of open corpora, electronic dictionaries, and NLP tools specifically dedicated to neology. This lack of structured and interoperable resources motivates projects such as NeoLex, which seek to integrate Chinese and Vietnamese into an international comparative framework by combining descriptive linguistics, lexicography, and computational methodologies for the automatic detection and analysis of neologisms.

2 Corpus and reference lexicons

The dataset consists of two components: reference lexicons compiled from dictionaries published before 2015, and diachronic media corpora collected between 2015 and 2025.

For Vietnamese, we used several GitHub-based lexical resources, including the *vi-DauCu dictionary*, *the hongocduc dictionary*, *the tudienv dictionary*, and *Wiktionary*. These resources yielded 79,774 Vietnamese lexical entries. For Chinese, we used two frequency dictionaries: 现代汉语常用词表.txt (*List of Common Words in Modern Chinese*, 2014) and 现代汉语第七版.txt (*Modern Chinese Dictionary*, 7th edition, 2016). The resulting Chinese reference lexicon contains 237,741 entries. The reference lexicons were cleaned and normalized in order to remove duplicates, noisy forms, and non-lexical items. Vietnamese entries were filtered to exclude isolated letters and malformed strings, while Chinese entries were normalized into simplified Chinese characters.

The media corpora were built from online newspapers, news portals, and specialized websites, which constitute privileged environments for the emergence and diffusion of neologisms. Due to time constraints and anti-scraping limitations, data collection was restricted to three Chinese sources — 人民日报 (*People's Daily*), 澎湃新闻 (*The Paper*), and 新华网 (*Xinhua News*) — and three Vietnamese sources: *Kenh14*, *Cafebiz*, and *Genk*. Although the current corpus focuses on edited online media, future work will extend the data collection to additional

sources in order to capture a broader range of lexical innovations. Between 20 and 30 articles per year and per website were collected, resulting in 1,286 Chinese articles and 759 Vietnamese articles. The extracted data were cleaned by removing residual HTML tags, noisy characters, and unnecessary metadata, then normalized into UTF-8 plain text. Each article was stored in CSV format according to a unified schema containing the author, title, date, content, and URL.

3 Neologism detection methodology

This section presents the methodology used to detect candidate neologisms in the Chinese and Vietnamese media corpora. The workflow combines tokenization, OOV-based detection, language-specific extraction methods, and manual expert validation.

3.1 Tokenization and preprocessing

Since Chinese and Vietnamese are isolating languages with specific segmentation challenges, tokenization was a crucial preprocessing step. Chinese texts were segmented using THULAC, while Vietnamese texts were tokenized with PyVi. Before candidate extraction, the corpora were cleaned and normalized. This process included the removal of residual HTML tags, duplicate content, noisy characters, and non-textual metadata. The resulting corpora were then converted into a standardized UTF-8 format and prepared for frequency analysis, dictionary comparison, and neologism detection.

Following Sagot & Nouvel (2013), we initially explored an Out-of-Vocabulary (OOV) approach based on comparison with pre-2015 reference lexicons. However, because OOV detection proved insufficient for Chinese and Vietnamese due to segmentation ambiguity and substantial noise, we adopted language-specific extraction methods.

3.2 Vietnamese n-gram extraction

To improve the detection of Vietnamese multiword neologisms, we applied an n-gram extraction method ranging from unigrams to 5-grams. This approach is particularly suitable for Vietnamese, where lexical units are often composed of several syllabic words separated by spaces. The extraction process combined frequency thresholds, Pointwise Mutual Information (PMI), contextual entropy, and redundancy filtering. Candidate sequences were then compared with the pre-2015 Vietnamese reference lexicons in order to distinguish known lexical units from potential neologisms. Examples include sống ảo (living virtually on social media), ông lớn (“big boss”), đốt tiền (“burning money”), and Thung lũng Silicon (“Silicon Valley”). Compared with simple OOV detection, the n-gram approach proved more effective for Vietnamese because it better captures stable multiword lexical innovations.

3.3 Chinese character-based extraction

For Chinese, word segmentation remains a major source of uncertainty because lexical boundaries are not explicitly marked in writing. To address this issue, we developed a character-based extraction method relying on sliding windows of 1 to 5 Chinese characters within sentence boundaries. Unlike the initial THULAC-based approach, this method does not depend entirely on automatic word segmentation. It directly extracts continuous character sequences from the corpus and compares them with the Chinese reference lexicon. Only sequences absent from the reference lexicon were retained as potential neological candidates.

The reference lexicon was filtered to keep only entries composed exclusively of Chinese CJK characters and ranging from one to five characters. After filtering, the system generated 3,432,729 candidate sequences: 389 one-character sequences, 262,914 two-character sequences, 798,267 three-character sequences, 1,117,839 four-character sequences, and 1,253,320 five-character sequences. For each sequence, total frequency, document frequency, and year of first occurrence were calculated. The outputs were then exported into five separate CSV files according to sequence length. This method improved the identification of short Chinese neologisms such as 元宇宙 ("metaverse"), 人工智能 ("artificial intelligence"), and 新冠肺炎 ("Covid-19").

3.4 Manual evaluation

Due to the absence of exhaustive gold-standard resources for contemporary Chinese and Vietnamese neology, recall and F1-score could not be reliably computed. The evaluation therefore relied on manual expert validation and precision-oriented analysis. For each language, we manually evaluated the 200 most frequent candidate neologisms extracted by the system. The annotation was conducted by native-speaking specialists in Chinese and Vietnamese linguistics and computational lexicography. A candidate was considered a valid neologism if it met the following criteria: absence from the pre-2015 reference lexicon; semantic stability across contexts; attestation in multiple documents or contexts; exclusion of named entities, dates, symbols, segmentation errors, and typographic noise. The manual evaluation confirmed that OOV-based detection generated substantial noise, especially in Chinese, where segmentation ambiguity and morphographic fragmentation affect candidate quality. Nevertheless, meaningful lexical innovations were identified in both languages.

In Chinese, the improved character-based extraction method generated a large number of candidate sequences with genuine lexical significance. However, because the extraction relied on sliding windows of one to five characters, many truncated fragments were also produced. For example, segments extracted from longer expressions could appear independently as shorter sequences without carrying autonomous semantic value. Consequently, the validation process did not retain all extracted character combinations. Only candidates corresponding to semantically meaningful and lexically innovative forms were considered genuine neologisms. The selection therefore prioritized lexical units presenting both formal stability and semantic novelty within contemporary Chinese discourse. Among the 200 most frequent Chinese candidates extracted from the corpus, 23 forms ranging from one to five Chinese characters were validated as genuine neologisms.

TABLE 1 Validated Chinese neologisms

Longueur (characters)	Number of neologisms (?/200)	Examples (extraits)
1 character	0	—
2 characters	3	防控 (Prevention and control) 新冠 (COVID-19) AI (Artificial Intelligence)
3 characters	6	元宇宙 (Metaverse) 多样性 (Diversity) 可持续 (Sustainability/Sustainable development) 发展中 (Developing/In development) 马克思 (Marx) 消费者 (Consumer)

TABLE 1 (*CONTINUED*)

Longueur (characters)	Number of neologisms (?/200)	Examples (extraits)
4 characters	21	疫 情 防 控 (Epidemic prevention and control/COVID-19 prevention and control) 同比增长 (Year-on-year growth) 基础设施 (Infrastructure) 乡村振兴 (Rural revitalization) 酸罂粟碱 (Noscapine) 新冠肺炎 (COVID-19 pneumonia/COVID-19) 科技创新 (Technological innovation) 卫生健康 (Public health) 一带一路 (Belt and Road Initiative)
5 characters	6	生物多样性 (Biodiversity) 高质量发展 (High-quality development) 盐酸罂粟碱 (Papaverine hydrochloride) 命运共同体 (Community with a shared future) 新能源汽车 (New energy vehicle (NEV)) 人形机器人 (Humanoid robot)

For Vietnamese, the extracted candidates included heterogeneous lexical phenomena such as genuine neologisms, personal names, commercial brands, product references, institutional names, English borrowings, and emerging multiword expressions. Consequently, a manual validation protocol was established in order to distinguish actual neological units from non-neological proper nouns and corpus noise. The annotation process excluded: personal names, isolated commercial brands, product model references, newspaper and institutional names, ordinary proper nouns, and isolated movie or media titles. By contrast, the validation retained lexical candidates corresponding to: lexical innovations, socio-cultural expressions, stabilized lexical borrowings, technolects and technological terminology, emerging multiword expressions, semantic innovations, and hybrid lexical units combining Vietnamese and English elements. In this context, forms such as AI, smartphone, and CEO were retained not because they constitute newly created lexical items in English, but because they function as stabilized lexical borrowings that have become fully integrated into contemporary Vietnamese media and technological discourse. Their high frequency, recurrent usage, and semantic stabilization across multiple contexts suggest an ongoing process of lexical integration within Vietnamese. This distinction was particularly important for Vietnamese media discourse, where technological communication, platform economy, and digital culture generate numerous hybrid lexical forms combining local and globalized vocabulary. For Vietnamese, 32 candidates among the 200 most frequent extracted forms were considered valid neological units.

TABLE 2 Validated Vietnamese neologisms by n-gram length

Longueur (n-grammes)	Number of neologisms (?/200)	Examples (extraits)
1 gramme	7	AI CEO Pro game THPT smartphone cryptocurrency

TABLE 2 (*CONTINUED*)

Longueur (n-grammes)	Number of neologisms (?/200)	Examples (extraits)
2 grammes	15	sống ảo (living virtually on social media) lùm xùm (controversy/scandal) ông lớn (big player/major corporation) đốt tiền (burning money/heavy spending) mô-men xoắn (torque) bo mạch (circuit board) quạt tản (cooling fan) smartphone gập (foldable smartphone) rò rỉ (information leak) Chơi Thật (play for real) Siêu nhí (child prodigy) râu dế (thin moustache) hô biến (magically transform) deep learning machine learning . . .
3 grammes	8	Thung lũng Silicon (Silicon Valley) gã khổng lồ (giant/giant corporation) quạt tản nhiệt (cooling fan/heat dissipation fan) bút S Pen (S Pen stylus) giấc ngủ Earable (Earable sleep technology/smart sleep monitoring) mô-men xoắn 300 (300 torque/300 Nm torque) vùng Comfort Zone (comfort zone) cryptocurrency tiền số (cryptocurrency/digital currency)
4 grammes	1	xỉu ngang xỉu dọc (extremely shocked/overwhelmed with surprise)
5 grammes	1	xôi xoài xôi sầu riêng (mango sticky rice and durian sticky rice)

Future work will involve the construction of a manually validated gold-standard corpus to support larger-scale quantitative evaluation and inter-method comparison.

3.5 Methodological limitations

The proposed methodology shows that neologism detection is highly dependent on corpus quality, segmentation accuracy, and filtering procedures. While the combination of OOV detection and language-specific extraction methods successfully identifies relevant lexical innovations, several limitations remain. For Chinese, the absence of spaces between characters creates segmentation uncertainty and morphographic noise. Although the character-based method improves the detection of short forms, longer political, institutional, or technical expressions may still be extracted only as partial fragments. For Vietnamese, the main difficulties concern media redundancy, English borrowings, social-media noise, and the distinction between genuine neologisms and transient lexical usages. Multiword expressions can be better captured through n-gram extraction, but manual validation remains necessary to assess their lexical stability. These limitations highlight the need for a post-processing stage combining human validation and semantic structuring. Manual validation filters noisy candidates and confirms lexical stability, while OntoLex-Lemon modelling transforms validated neologisms into

structured and interoperable lexical resources suitable for multilingual lexicography and semantic-web applications.

4 From neologism detection to ontology-based lexicography: OntoLex-lemon modelling in NeoLex

The NeoLex workflow follows a corpus-to-ontology architecture combining neologism detection, OntoLex-Lemon lexical modelling (Liu et al., 2011; McCrae et al., 2017; Alcina, 2024), semantic- domain classification, multilingual alignment, and lexicographic visualization.



FIGURE 1 NeoLex workflow from corpus-based neologism detection to ontology-driven lexicographic publication

The automatic detection of neologisms constitutes only one stage in the construction of sustainable lexical resources. In order to transform extracted candidates into interoperable dictionary entries, NeoLex adopts the OntoLex-Lemon framework, combined with SKOS for conceptual categorization, vartrans for multilingual alignment, and corpus-based metadata for diachronic analysis. The objective of the NeoLex ontology is not to represent world knowledge as a whole, but to organize neologisms according to semantic domains, conceptual relations, multilingual correspondences, and diachronic properties.

4.1 From validated candidates to lexicographic entries

Before being represented in OntoLex-Lemon, validated neological candidates are transformed into structured lexicographic entries. Each entry contains lexical, semantic, and diachronic information extracted from the corpus and enriched through manual validation. The entry structure includes the lemma, language, transcription, definition, semantic domain, conceptual category, first attestation date, validation status, and external semantic links. Table 3 illustrates the unified metadata schema adopted in NeoLex through two example entries, one in Chinese and one in Vietnamese. Although the lexical values are language-specific, both entries follow exactly the same metadata structure.

TABLE 3 Unified Metadata Schema and Examples of Chinese and Vietnamese Neologism Entries

Field	Value in Chinese	Value in Vietnamese
Lemma	防控	sống ảo
Language	Chinese	Vietnamese
Pronunciation	fáng kòng	sống ảo
Definition	Prevention and control	Virtual life on social networks
Domain	Public Health	Digital Culture
Concept	PublicHealth	DigitalCulture
First attestation	2020	2016
Validation status	candidate	

These lexicographic entries constitute the intermediate layer between corpus-based neologism detection and ontology-based lexical modelling.

4.2 OntoLex-Lemon modelling

The validated lexicographic entries are automatically transformed into RDF resources compliant with the OntoLex-Lemon model. The following sections describe the different layers of the model.

4.2.1 Lexical layer

The lexical layer represents the linguistic properties of validated neologisms using the OntoLex-Lemon core model. Each neological unit is modelled as an *ontolex:LexicalEntry* associated with a canonical form (*ontolex:Form*) and one or more lexical senses (*ontolex:LexicalSense*). This layer records language-specific information such as the written representation, part of speech, and romanization. In NeoLex, the Chinese neologism 防控 (“*prevention and control*”) is modelled as follows:

```
:fangkong
a ontolex:LexicalEntry;
   dct:language “zh”;
   lexinfo:partOfSpeech lexinfo:noun;
   ontolex:canonicalForm :fangkong_form;
   ontolex:sense :fangkong_sense.

:fangkong_form
   a ontolex:Form;
   ontolex:writtenRep “防控"@zh;
   lexinfo:romanization “fáng kòng”.

:fangkong_sense
   a ontolex:LexicalSense.
```

The same modelling scheme is applied to Vietnamese neologisms:

```
:song_ao
   a ontolex:LexicalEntry;
   dct:language “vi”;
   lexinfo:partOfSpeech lexinfo:verb;
   ontolex:canonicalForm :song_ao_form;
   ontolex:sense :song_ao_sense.

:song_ao_form
   a ontolex:Form;
   ontolex:writtenRep “sống ảo"@vi.

:song_ao_sense
   a ontolex:LexicalSense.
```

NeoLex also represents multiword neologisms as *ontolex:MultiWordExpression*. For expressions such as *sống ảo* or brain–computer interface, the internal lexical structure can be explicitly described using the OntoLex decomp module when necessary.

4.2.2 Conceptual layer with SKOS

The conceptual layer distinguishes lexical forms from the concepts they lexicalize. Following the OntoLex-Lemon model, lexical senses are linked to conceptual entities represented as ontolex:LexicalConcept. These concepts are organized through SKOS semantic relations, allowing NeoLex to represent semantic domains and conceptual hierarchies independently from language-specific lexical forms. For example, the Chinese neologism 防控 lexicalizes the concept PublicHealth:

```
:fangkong_sense
  ontolex:isLexicalizedSenseOf
  :PublicHealth.

:PublicHealth
  a ontolex:LexicalConcept;
  skos:prefLabel “Public Health"@en;
  skos:broader :Health
```

Conceptual relations are represented through SKOS properties such as skos:broader and skos:related:

```
:ArtificialIntelligence
  skos:broader :Technology.

:Fintech
  skos:broader :DigitalEconomy.

:DigitalCulture
  skos:related :SocialMedia.

:PublicHealth
  skos:related :Covid19.
```

This distinction between lexical entries, lexical senses, and conceptual entities enables NeoLex to organize Chinese and Vietnamese neologisms within a coherent semantic network while preserving language-specific lexical realizations and multilingual correspondences.

4.2.3 Semantic-domain classification

The multilingual layer relies on the **vartrans** module to represent translation and cross-linguistic relations between the Chinese–French and Vietnamese–French lexical resources. Translation relations are established at the level of **lexical senses** rather than surface forms. Each validated neologism is represented by a language-specific `ontolex:LexicalEntry`, while its translation equivalent is represented as an independent lexical entry in the target language. The corresponding lexical senses are then linked through `vartrans:translation`, allowing NeoLex to preserve language-specific lexical information while explicitly modelling multilingual equivalence. For example, the Chinese neologism 防控 is linked to its French equivalent through corresponding lexical senses:

```
:fangkong
  a ontolex:LexicalEntry;
  ontolex:canonicalForm :fangkong_form;
  ontolex:sense :fangkong_sense.

:fangkong_sense
  vartrans:translation
    :prevention_controle_sense.

:prevention_controle
  a ontolex:LexicalEntry;
  ontolex:canonicalForm :prevention_controle_form;
  ontolex:sense :prevention_controle_sense.
```

Similarly, the Vietnamese neologism **sống ảo** is represented as a Vietnamese lexical entry aligned with its French equivalent through corresponding lexical senses:

```
:song_ao
  a ontolex:LexicalEntry;
  ontolex:canonicalForm :song_ao_form;
  ontolex:sense :song_ao_sense.

:song_ao_sense
  vartrans:translation
    :vie_virtuelle_sense.

:vie_virtuelle
  a ontolex:LexicalEntry;
  ontolex:canonicalForm :vie_virtuelle_form;
  ontolex:sense :vie_virtuelle_sense.
```

Both the Chinese and Vietnamese lexical senses may additionally be linked to shared conceptual entities through `ontolex:isLexicalizedSenseOf`, thereby separating multilingual translation relations from language-independent conceptual representation. This distinction avoids semantic conflation and enables multilingual alignment whilepreserving language-specific conceptual nuances.

It is particularly relevant for culturally marked neologisms such as 内卷 (involution), 躺平 (lying flat), and *sống ảo*, whose meanings cannot always be captured through direct lexical equivalence.

4.2.4 Metadata layer

In addition to lexical and conceptual information, NeoLex incorporates a metadata layer describing the corpus-based and diachronic properties of each neologism. This layer records information such as the year of first attestation, corpus frequency, source corpus, extraction method, and validation status.

```
:fangkong
  neolex:firstAttestation “2020";
  neolex:frequency “5234";
  neolex:sourceCorpus :ChineseMediaCorpus;
  neolex:validationStatus neolex:validated.
```

These metadata enrich lexical entries with information about their emergence and usage within the corpus. They transform *NeoLex* into a diachronic lexical observatory capable of tracing the appearance, diffusion, and stabilization of lexical innovations across time and domains.

4.2.5 Semantic-domain classification

The current semantic-domain vocabulary consists of eight high-level domains derived from corpus analysis and manual lexicographic annotation, including Public Health, Artificial Intelligence, Digital Culture, Fintech, Green Mobility, Economy, Society, and Environment. The hierarchy was manually constructed for the present study using SKOS broader/narrower relations. At this stage, it is domain-specific and does not directly reuse existing vocabularies, although future work will investigate alignment with resources such as EuroVoc or Wikidata.

```
:PublicHealth
  skos:broader :Health.

:ArtificialIntelligence
  skos:broader :Technology.

:Fintech
  skos:broader :DigitalEconomy.

:DigitalCulture
  skos:broader :Society.

:GreenMobility
  skos:broader :SustainableDevelopment.
```

Validated neologisms are associated with these domains through their lexical senses and conceptual representatons. For example, the Chinese neologisms 防控 (fángkòng, “prevention and control”) and 新冠 (xīnguān, “COVID-19”) are linked to the domain PublicHealth, while 脑机 (nǎojī, “brain–computer interface”) and 人工智能 (réngōng zhìnéng, “artificial intelligence”) are associated with ArtfcialIntelligence. Similarly, the Vietnamese neologisms sống ảo (“living virtually on social media”), Viettel Money (a digital payment and mobile fnancial service), and Taxi Xanh SM (a green electric taxi service) are linked respectvely to the domains DigitalCulture, Fintech, and GreenMobility. The semantc-domain layer enables NeoLex to represent neologisms not as isolated lexical units but as interconnected conceptual phenomena embedded within broader technological, economic, social, and cultural transformatons.

5 Visualization and lexicographic prototyping

To facilitate the consultation and exploration of multilingual neological resources, NeoLex includes an interactive prototype developed with Streamlit. Rather than functioning solely as a visualization dashboard, the platform serves as an ontology-driven lexicographic interface allowing users to search, filter, and explore validated neological entries together with their lexical, semantic, and diachronic information. The validated entries are modelled using OntoLex-Lemon and transformed into structured RDF/OWL resources. The Streamlit application retrieves this information and presents it through interactive lexicographic cards displaying the lemma, language, transcription, definition, semantic domain, conceptual category, validation status, first attestation date, and multilingual correspondences. The interface currently supports both Chinese–French and Vietnamese–French lexical resources. Users can browse entries through keyword search, semantic-domain filters, and validation categories.

Figure 2 presents selected views of the NeoLex interface, including lexicographic consultation pages and statistical dashboards for Chinese and Vietnamese neological resources.

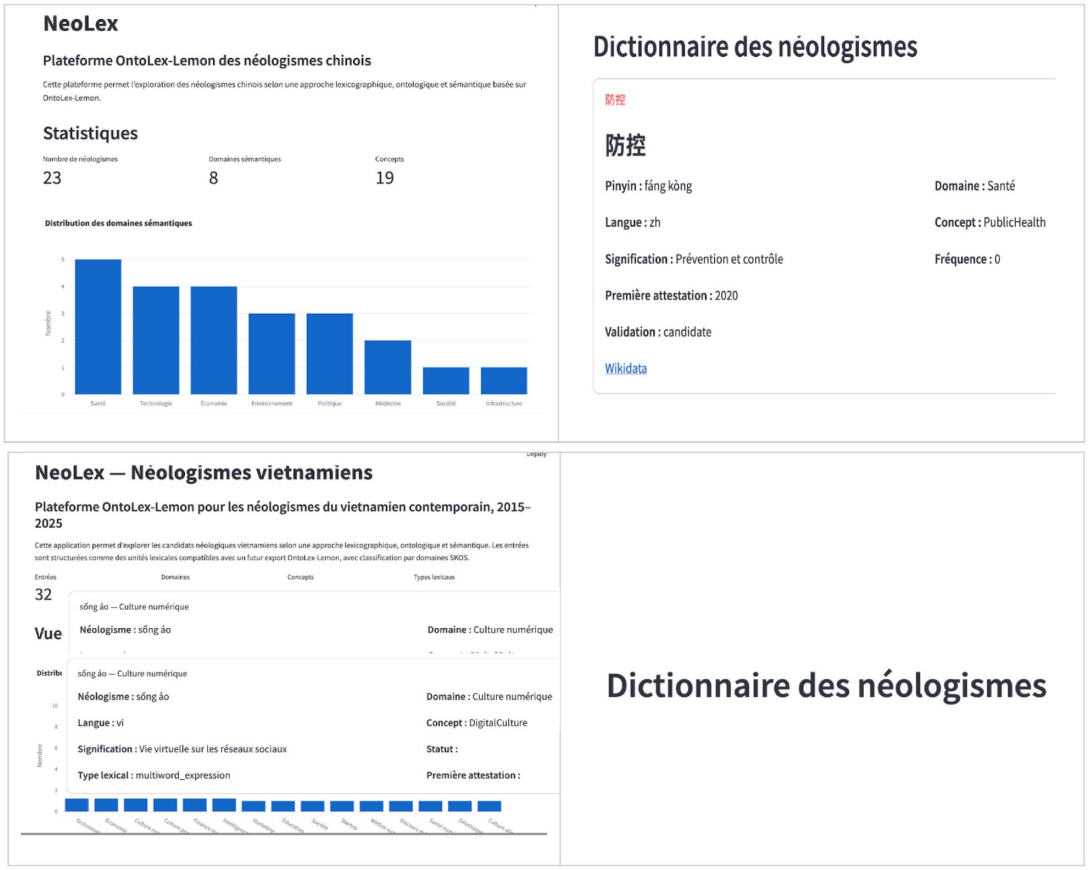


FIGURE 2 NeoLex Streamlit interface for multilingual neological lexicography and semantic exploration

Beyond lexicographic consultation, the platform provides statistical visualizations of semantic domains and conceptual categories extracted from the ontology. These dashboards offer an overview of the distribution of neologisms across technological, social, economic, environmental, and cultural domains. Streamlit was selected because it enables rapid prototyping, seamless integration of ontology-based resources, and lightweight web deployment. The resulting prototype demonstrates how OntoLex-Lemon modelling can be combined with interactive visualization techniques to support multilingual computational lexicography and the exploration of contemporary lexical innovation. The current prototype is available as an interactive Streamlit interface. Future releases will provide RDF exports and SPARQL access, while REST API support and downloadable datasets are planned to facilitate integration with external NLP and lexicographic applications.

6 Conclusion

This paper presented NeoLex, a multilingual electronic lexical resource dedicated to the automatic detection, semantic modelling, and lexicographic visualization of contemporary Chinese–French and Vietnamese–French neology. By combining corpus linguistics, Natural Language Processing, computational lexicography, and OntoLex-Lemon modelling, the project aims to transform automatically extracted neological candidates into structured and interoperable multilingual lexical resources. Unlike traditional neologism detection systems, NeoLex does not stop at candidate extraction. Validated candidates are transformed into structured

lexicographic entries, encoded in OntoLex-Lemon, and made accessible through an ontology-driven dictionary consultation interface. This corpus-to-dictionary workflow provides a complete framework linking neologism detection, lexical modelling, semantic representation, and electronic lexicography. The study demonstrated that Chinese and Vietnamese neology presents important linguistic and computational challenges related to lexical productivity, segmentation ambiguity, multiword expressions, and hybrid lexical borrowings. In response to these difficulties, several complementary extraction methods were implemented, including OOV-based detection, Vietnamese n-gram extraction, and a character-based extraction strategy specifically adapted to Chinese morphographic structures. The results confirmed that language-specific approaches are necessary in order to improve the identification of meaningful lexical innovations and emerging lexical usages.

The validated candidates were subsequently modelled using OntoLex-Lemon, SKOS, and vartrans, allowing the representation of lexical entries, semantic domains, multilingual correspondences, and diachronic corpus metadata within a unified ontology-driven framework. This modelling process transforms NeoLex into both a semantic-domain ontology for neology and a multilingual electronic lexical resource capable of tracing lexical innovation across contemporary media discourse. The Streamlit-based prototype further demonstrated how automatically extracted neologisms can be explored, validated, and visualized through ontology-driven lexicographic interfaces integrating semantic filtering, multilingual metadata, and conceptual categorization. Several limitations nevertheless remain. The absence of exhaustive gold-standard corpora prevents large-scale quantitative evaluation, while Chinese segmentation ambiguity and Vietnamese media noise continue to generate partial fragments and non-neological candidates. Although the present methodology integrates semantic validation and semantic-domain classification, it does not yet systematically detect semantic shifts associated with already existing lexical forms. Expressions such as 内卷 in Chinese or trà xanh in Vietnamese illustrate semantic evolutions that may emerge without the creation of entirely new lexical forms. In addition, the corpus remains limited to a restricted number of media sources and does not yet fully integrate conversational and social-media platforms, where neological creativity is particularly productive. The implementation scripts developed for corpus preprocessing, candidate extraction, and RDF generation will be released as open-source software after publication in order to facilitate reproducibility and further research.

Future work will therefore focus on the construction of manually validated gold-standard corpora, the integration of semantic-shift detection methods, RDF/OWL export and SPARQL querying, graph-based semantic visualization, and the extension of NeoLex toward larger collaborative multilingual lexicographic platforms.

The proposed framework was successfully applied to both Chinese–French and Vietnamese–French neological resources, demonstrating that OntoLex-Lemon provides a unified representation framework for multilingual neological dictionaries across typologically different languages. More broadly, NeoLex illustrates how NLP-based neologism detection can be integrated into a complete lexicographic workflow, transforming corpus evidence into structured multilingual dictionary entries through ontology-based modelling and semantic enrichment.

Availability of Resources

The NeoLex GitHub repository provides the corpora, Jupyter notebooks, Python scripts, Streamlit applications, and structured JSON lexical resources used in this study. The OntoLex-Lemon RDF resources described in this paper are currently under preparation and will be added in a future release. GitHub: https://github.com/Computational-Lexico/NeoLex

References

Alcina, A. (2024). ONTODIC: A model of linguistic knowledge representation based on description logic. *Terminology*, 30(1), 208–237. https://doi.org/10.1075/term.23030.alc

Bernal, E., Freixa, J., & Torner, S. (2020). Néologicité et dictionnarisabilité: Deux conditions inverses ? *Neologica*, 14, 47–60.

Boulanger, J.-C. (2010). Sur l'existence des concepts de « néologie » et de « néologisme »: Propos sur un paradoxe lexical et historique. In T. Cabré, O. Domènech, R. Estopà, J. Freixa, & M. Lorente (Eds.), *Actes del I Congrés internacional de neologia de les llengües romàniques* (pp. 31–73). Universitat Pompeu Fabra.

Cabré, M. T. (1994). Terminologie et dictionnaires. *Meta*, 39(4), 589–597.

Cabré, M. T. (2004). Introducció: La importància de la neologia per al desenvolupament sostenible de la llengua catalana. In *Llengua catalana i neologia* (pp. 17–45). Meteora.

Chao, Y. R. (1976). *Aspects of Chinese Sociolinguistics: Essays.* Stanford University Press.

Che, J. (2017). *Research on the New Nouns of Modern Chinese: Taking New Nouns from 2010–2014 as the Center* [Master's thesis, Yanbian University].

Crystal, D. (1996). *The Cambridge Encyclopedia of Language* (2nd ed.). Cambridge University Press.

Díaz Hormigo, M. T. (2020). Precisiones para una caracterización lingüística de la neología semántica. *ELUA*, 34, 73–94.

Dinh, D., Hoang, K., & Nguyen, V. T. (2001). Vietnamese word segmentation. In *Proceedings of the Sixth Natural Language Processing Pacific Rim Symposium*. Tokyo, Japan.

Đỗ Phương, T. & Nguyễn, N. T. T. (2022). Đặc điểm và phương thức cấu tạo từ tạo mới tiếng Hàn (Trọng tâm là từ tạo mới tiếng Hàn liên quan đến Covid-19). *Ngôn* ngữ & Đời sống, 4(324), 78–86.

Drouin, P. (2021). Repérage outillé de la néologie : Apports de la linguistique de corpus et du traitement automatique de la langue. In M. B. Villar Diaz, J. de Hoyos, P. Dury, J. Makri-Morel, & V. Renner (Eds.), *La néologie des langues romanes: Nouvelles approches, dynamiques et enjeux* (pp. 299–319). Peter Lang.

Fan, H. (2015). The interaction of Chinese translations of Western books and early modern Chinese and Japanese neologisms. *Japanese Language Learning Research*, 181(6), 27–33.

Gérard, C., Bruneau, L., Falk, I., Bernhard, D., & Rosio, A.-L. (2016). Le Logoscope: Observatoire des innovations lexicales en français contemporain. In J. García Palacios, G. De Sterck, D. Linder, N. Maroto, M. Sánchez Ibáñez, J. Torres del Rey (Eds.), *La neología en las lenguas románicas: Recursos, estrategias y nuevas orientaciones* (pp. 339–356). Peter Lang.

Huang, Y. (2024). Equivalence at word level: A study of the translation of Chinese neologism. *Journal of International Social Science*, 1(6), 4–10.

Jing-Schmidt, Z., & Hsieh, S. K. (2019). Chinese neologisms. In C.-R. Huang, Z. Jing-Schmidt & B. Meisterernst (Eds.), *The Routledge Handbook of Chinese Applied Linguistics* (pp. 514–534). Routledge.

Kang, S. (2002). Development and study of the "Modern Chinese New Words Information Electronic Dictionary". *Computational Linguistics and Chinese Language Processing*, 7(2), 89–100.

Le-Hong, P., Roussanaly, A., & Nguyen, T. M. H. (2015). A syntactic component for Vietnamese language processing. *Journal of Language Modelling*, 3(1), 145–184.

Liu, J., Li, S., & Gu, A. (2011). An overview of ontology research on Chinese neologism. *Journal of Wenjiao Ziliao*, 19, 29–30.

Lombard, A., Huyghe, R., & Gygax, P. (2021). Neological intuition in French: A study of formal novelty and lexical regularity as predictors. *Lingua*, 254, 103055. https://doi.org/10.1016/j.lingua.2021.103055

Lü, S. (1984). *汉语语法论文集* [Recueil d'articles sur la grammaire du chinois]. Maison d'Édition du Commerce.

McCrae, J., Bosque-Gil, J., Gracia, J., Buitelaar, P., & Cimiano, P. (2017). The OntoLex-Lemon model: Development and applications. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek & V. Baisa (Eds.), *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017* (pp. 587–597). Lexical Computing CZ.

Neveu, F. (2011). *Dictionnaire des sciences du langage (2nd ed.)*. Armand Colin.

Newmark, P. (1988). A Textbook of Translation. Prentice Hall.

Nguyễn, V. K. (2013). *Từ ngoại lai trong tiếng Việt* [Les mots empruntés dans la langue vietnamienne]. NXB Thành phố Hồ Chí Minh.

Pan, W., Liu, T., Sun, Y., Gong, B., Zhang, Y., & Yang, P. (2021). Finding Chinese new words by combining self-encoder and adversarial training. In *Proceedings of the 20th China National Conference on Computational Linguistics* (pp. 736–746).

Poix, C. (2021). Études francophones de néologie. Complexité terminologique. *Neologica*, 15, 97–115.

Pruvost, J., & Sablayrolles, J.-F. (2003). *Les néologismes*. Presses Universitaires de France.

Quemada, B. (1971). À propos de la néologie. Essai de délimitation des objectifs et des moyens d'action. *La Banque des mots*, 2, pp. 137–150.

Sagot, B., Nouvel, D., Mouilleron, V., & Baranes, M. (2013). Extension dynamique de lexiques morphologiques pour le français à partir d'un flux textuel. In *Proceedings of TALN 2013* (pp. 407–420). Les Sables d'Olonne, France. ATALA.

Contact information

Lian Chen
LLL, University of Orleans
loselychen@gmail.com

Damien Nouvel
ERTILM
damien.nouvel@inalco.fr

Huy-Linh Dao
CRLAO-CNRS
huy-linh.dao@inalco.fr

Alexander Delaporte
CNRS-CRLAO
alexander.delaporte@ehess.fr

Acknowledgements

This work was carried out within the NeoLex project, supported by CRLAO (Centre de Recherches Linguistiques sur l'Asie Orientale, UMR 8563, CNRS–EHESS–INALCO).

Flávia de Oliveira Maia-Pires & Silvio Licurgo Pires

USER-ORIENTED SPECIALIZED DIGITAL LEXICOGRAPHY AND TERMINOLOGICAL ACCESSIBILITY

The LexiC Platform Applied to the Glossary of Brazilian Migration Legislation

Abstract The technical and legal language of Brazilian migration legislation constitutes a significant barrier to information access for migrants and refugees in situations of social, linguistic, and legal vulnerability. This article presents the LexiC Platform, a web-based computational infrastructure for the collaborative management of terminological data, applied to the Computerized Terminological Glossary of Brazilian Migration Legislation. The research question guiding the proposal is how terminological data extracted through well-established corpus analysis tools can be systematized in a collaborative digital environment so as to preserve methodological consistency, traceability, and the integrity of conceptual relations across a distributed terminographic workflow. The proposal draws on the communicative approach to Terminology (Cabré, 1999, 2005), user-oriented functional lexicography (Tarp, 2008, 2012), and Corpus Linguistics methodologies (Biber, Conrad & Reppen, 1998). The Platform organizes terminographic work into interdependent workflows, covering structured data entry, methodological validation, collaborative review, and automated generation of terminographic products, operationalized through computerized records with more than 60 standardized fields. The specialized corpus comprises approximately 150,000 words, covers laws and decrees published between 2017 and 2025, and includes around 270 terms at different stages of description. Results indicate that terminological rigor, collaborative infrastructure, and user orientation produce resources that are accessible, and socially relevant.

Keywords Specialized lexicography; digital lexicography; user orientation; terminological accessibility; terminography; migration legislation

1 Introduction

The specialized language of Brazilian migration legislation constitutes a significant barrier to legal information access for migrants and refugees, particularly those in situations of social, linguistic, and legal vulnerability. The main difficulties stem from technical and legal vocabulary and the conceptual complexity inherent to migration law, which compromises the understanding of fundamental rights, duties, and administrative procedures. According to the World Migration Report (International Organization for Migration, 2024), there are approximately 281 million international migrants worldwide, equivalent to 3.6% of the global population. In Brazil, the National Committee for Refugees (CONARE) registered 58,628 applications for refugee status recognition in 2023 alone, totaling 406,695 applications filed since 2011. This growing volume of applications highlights the urgency of terminological resources that render migration-related legal language accessible to the very population that needs this information. The absence of systematized and accessible terminological resources geared toward non-specialist users contributes to the perpetuation of inequalities in access to justice and public services, directly affecting the full exercise of citizenship.

Recognizing this gap shows that the production of lexicographic and terminographic resources has historically been an activity of high methodological complexity and considerable labor demand, given that compiling terminological data, articulating microstructural entries, managing conceptual relations, and formatting publishable outputs make up a set

of interdependent tasks that have resisted full systematization for decades. Available tools, such as the *Sketch Engine* (Kilgarriff et al., 2014), are known to support only isolated stages of this workflow, without addressing the systematic organization of the data obtained, as will be detailed in Section 3. The result is a fragmented workflow in which the terminographer must move manually between different systems, which can generate methodological inconsistencies and loss of information or traceability of the decisions made throughout the process, also hindering teamwork on projects aimed at terminological documentation and the production of terminographic or lexicographic outputs.

Given this scenario, this article aims to present the LexiC Collaborative Terminological Data Management Platform (https://plataforma.lexic.com.br), also referred to as the LexiC Platform throughout this work, a web-based computational infrastructure for the collaborative management of terminological data, and to demonstrate its application in the development of the Computerized Terminological Glossary of Brazilian Migration Legislation. The proposal is situated within the field of digital Lexicography and Terminography and integrates three complementary theoretical pillars: the communicative basis of Terminology (Cabré, 1999, 2005), the functional principles of user-oriented lexicography (Tarp, 2008, 2012), and the methodological toolkit of Corpus Linguistics (Biber, Conrad & Reppen, 1998; Tognini-Bonelli, 2001). The article is organized as follows: Section 2 presents the theoretical foundations underlying the proposal, articulating communicative Terminology, specialized lexicography, and systematic terminography; Section 3 describes the LexiC Platform, including its methodological foundations, technical architecture, and positioning in relation to already-established terminology management tools; Section 4 presents the case study of the Computerized Terminological Glossary of Brazilian Migration Legislation, including the analysis of the term formation of the concept *visto* ('visa'); and Section 5 offers concluding remarks, summarizing the contributions and indicating future directions.

2 Terminological accessibility and specialized lexicography

The identification of the linguistic units used by specialists in fields of knowledge gave rise to the discipline known as Terminology, whose social function is to facilitate communication between specialists and the lay public, overcoming the terminological barriers created by contact between languages (Cabré, 1993, p. 11). From the standpoint of the communicative approach (Cabré, 1999, 2005), terms are linguistic units endowed with specialized value according to their context of use, which implies that the terminology of migration legislation must be studied in its actual contexts of circulation, taking into account the different user profiles that interact with it. Along these lines, Maia-Pires (2009, p. 34) states that the set of specialized words used to refer to knowledge in a given domain, whether scientific, technical, or specialized, constitutes a terminology, reaffirming the functional and communicative dimension of terminological resources.

User-oriented functional lexicography, developed by Tarp (2008, 2012), provides the theoretical framework needed to design terminological resources aimed at non-specialist audiences. From this perspective, information tools are defined as instruments designed to satisfy specific user needs in specific situations of use (Tarp, 2008), shifting the organizing axis of the lexicographic resource from the internal structure of the language to the communicative and cognitive needs of users. This conception is particularly relevant in the context of migration legislation, where non-specialist users, including migrants, refugees, and reception professionals, need clear, reliable, and functional access to legal and administrative concepts that directly affect their life trajectories.

Terminography, in turn, is concerned with the recording, processing, and presentation of terminological data acquired through terminological research (ISO 704, 2022; ISO 1087, 2019). The articulation between communicative Terminology, functional lexicography, and systematic Terminography (Faulstich, 2003; Almeida, 2006; Maia-Pires & Cruz, 2021) constitutes the theoretical and methodological basis on which the LexiC Platform and the glossary built with this tool rest, as resources designed from the communicative and cognitive needs of users, prioritizing terminological accessibility, definitional clarity, and orientation toward real use.

3 The LexiC Platform: foundations, architecture, and functionalities

The development of the LexiC Collaborative Terminological Data Management Platform is grounded in the principles of Terminography presented in Section 2, also taking into account the target audience and the macrostructural and microstructural aspects that make up a terminographic work. According to Barros (2004, p. 68), Terminography constitutes a scientific discipline that analyzes its object of study, terminological dictionaries, proposes new models for data treatment, reflects scientifically on its own practice, and consolidates a methodology for compiling terminographic works. It is from this perspective that the design of LexiC emerges, developed as a computerized resource customized to the specific needs of terminological research that support terminographic projects, rather than as an adaptation of generic data management instruments or of systems oriented primarily toward terminological consistency in translation workflows. The collaborative platform brings together functionalities such as the customization of the terminological record, structured management of the definitional base, assisted drafting of terminological definitions, and the editing of entries from previously selected fields. These elements set the platform apart as an instrument conceived from the very methodological demands of terminological research, rather than as a general-purpose tool adapted to this end.

It is worth situating the LexiC Platform in relation to already-established terminology management systems, such as SDL MultiTerm and TermWeb, which are widely recognized for their effectiveness in meeting the demands of terminological consistency in computer-assisted translation workflows and multilingual content management in corporate and institutional settings. The LexiC Platform occupies a distinct application domain, that of academic terminological research, in which the record for each term is structured around multiple definition sources, conceptual and cross-referencing relations, quantitative data extracted directly from the base corpus, and a record of the degree of specialized consensus obtained through documented human validation. This difference in purpose, rather than a hierarchy of technical sophistication between the systems, is what underlies LexiC's rationale. Likewise, the Platform does not aim to replace already-established term-extraction tools, such as the Sketch Engine (Kilgarriff et al., 2014), which each researcher is free to draw on according to their own corpus-mining methodology. The Platform's proposal addresses a distinct and complementary stage of the terminographic workflow: the systematization of terminological data after the extraction stage. Once term candidates have been identified and analyzed in their contexts of use through the available corpus tools, LexiC offers the structural support needed to organize, validate, and record this information in consistent terminological records, thereby overcoming the methodological fragmentation that characterizes the manual transition between extraction tools and terminology management systems.

The LexiC Collaborative Terminological Data Management Platform was conceived by researchers Maia-Pires and Vilarinho within the LexiC Research Group (Science, Projects, and Research on the Lexicon) at the University of Brasília, as a digital methodological instrument

for the systematic recording, organization, and documentation of terminological data. Technical development was carried out by Pires (2025), based on methodological guidelines grounded in Faulstich (2003) and on the group's accumulated experience in earlier projects, such as the COVID-19 Terminological Glossary (Maia-Pires & Cruz, 2021). The Platform integrates the principles of systematic Terminography with the contemporary demands of collaborative research in distributed digital environments, and it is registered as software under INPI number BR512026002753-5.

The Platform allows terminographic work to be carried out simultaneously and in a geographically decentralized manner by researchers, specialists, and coordinators, each operating under a clearly defined user profile with specific functions and permissions. Collaborative work is organized through a system of differentiated user profiles. The coordinator holds administrative and scientific management of the project, with the authority to create projects, set permissions, invite participants, validate records, and export terminographic outputs. The vice-coordinator performs support and monitoring functions and may be delegated authority by the coordinator. The researcher is responsible for entering, analyzing, and reviewing terminological data. The specialist acts as a subject-matter consultant for each project and is responsible for the technical and conceptual validation of records, issuing formal opinions that are recorded in the terminological record. The system further enables collaboration between researchers from different institutions and countries, ensuring, through systematized fields, methodological standardization of data regardless of participants' geographic location.

Each record brings together standardized fields, including: term entry, grammatical classification, variants, synonyms and antonyms, definitions with their respective sources, contexts of use extracted from the corpus, cross-references, terminological equivalents in other languages, an encyclopedic note, degree of consensus and specialist opinion, and the possibility of including images, audio, and video. The Multimedia Resources field allows the insertion of illustrations (in PDF, JPG, JPEG, and PNG formats), videos (via URLs from streaming or cloud storage services), and audio (also via URL). The system implements version control for each record, ensuring traceability and transparency throughout the research process.

The system's design is guided by three fundamental principles. The first is methodological adherence: all fields and operational workflows are structured in accordance with reference terminological theories, so that the platform does not constitute a generic data repository but a specialized terminographic instrument. The second principle is interinstitutional collaboration: the system's architecture allows multiple researchers to work simultaneously on the same project from different institutions and countries, with access control differentiated by user profile. The third is traceability: every modification made to a terminological record is automatically logged in the system's history, with identification of the person responsible, the date, and the nature of the change, which guarantees the scientific integrity of the terminological database throughout the project's entire life cycle.

With regard to technological infrastructure, the LexiC Platform was developed with the aid of Visual Studio Code as the development environment and Lucidchart for diagramming the system's workflows and data structures. The back-end layer was built in JavaScript and Python, while the front-end layer relies on HTML, CSS, and JavaScript, ensuring interoperability between the logic for processing terminological data and the interface made available to users. Data storage and management are handled through the MySQL database management system, responsible for ensuring the integrity and consistency of structured information throughout the entire terminographic workflow. The development methodology adopted is grounded in the MVC (Model-View-Controller) architecture, which separates data logic, presentation, and control of user interactions, and in the Object-Oriented Programming (OOP) paradigm,



FIGURE 1 Public interface of the LexiC Platform (Maia-Pires, Vilarinho & Pires, 2025)

which gives the platform modularity and supports its capacity for future expansion, both in terms of incorporating new functionalities and increasing the volume of terminological data processed.

The LexiC Platform is accessible through its public web interface, available at plataforma.lexic.com.br. Figure 1 illustrates the main terminographic workflows operationalized by the system: the registration of a terminological unit, the completion of the corresponding terminological record, the user-profile-based validation system, and the visualization of the database under construction. Remote and simultaneous access by multiple researchers, regardless of their geographic location, constitutes one of the structural principles of the platform's design.

4 Case study: Computerized Terminological Glossary of Brazilian Migration Legislation

The development of the Computerized Terminological Glossary of Brazilian Migration Legislation constitutes the case study that demonstrates the operationalization of the LexiC Platform in a terminographic project of direct social relevance. The specialized corpus, named Lexic.Legs.Imigração-UnB-BR, brings together approximately 150,000 words drawn from laws, decrees, and other normative acts issued between 2017 and 2025, including the Migration Law (Law 13,445/2017) and the Emergency Assistance to Migrants Law (Law 13,684/2018). The identification of terminological units follows a corpus-driven approach combined with qualitative analysis based on the Platform's terminological records, and currently totals approximately 270 terms at different stages of description, a number that reflects the project's ongoing expansion since its earlier phase (Maia-Pires, 2024).

From a methodological standpoint, the construction of concept maps constitutes a preliminary and complementary stage to the completion of the records, guiding the identification of simple and complex terms and the establishment of the conceptual relations that structure the specialized domain. According to Finatto and Krieger (2004, p. 134), a concept map is a hierarchical diagram composed of the key terms of a specialized field, similar to an organizational chart, an instrument that makes it possible to test the terminological relevance of each candidate term before its incorporation into the database. Networks of cross-references become apparent during this analysis, frequently leading to the identification of new candidate terms and to the revision of previously established hierarchical relations.

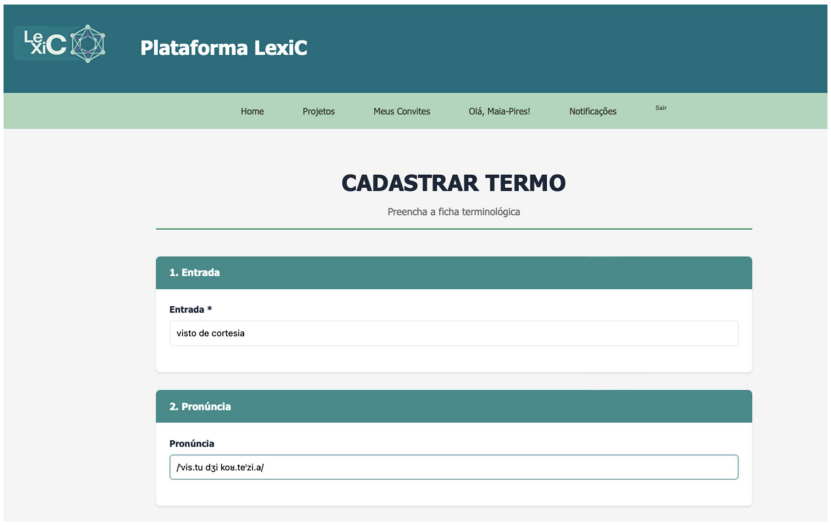


FIGURE 2 Terminological record: entry and pronunciation fields for the term "visto de cortesia," with phonetic transcription in the International Phonetic Alphabet (Maia-Pires, Vilarinho & Pires, 2025)

The term entry *visto de cortesia* ('courtesy visa') illustrates the integrated functioning of the platform's microstructure and makes it possible to demonstrate, on the basis of Faulstich's (2003) formal apparatus, the term-formation criteria that guide the distinction between simple and complex terms in the LexiC Platform. According to the author, the terminological construct corresponds to the equation C = <T(F), LT, R>, in which the terminological formative (F) may take the form of a simple term (F→A) or of a predicate (AB, AaB, ABC, etc.), according to the combination rules established between its components. From this perspective, *visto* ('visa') is configured as a simple term, a base unit (F→A), while the predications attached to this base form the corresponding complex terminological units (AB structure), whose components carry semantic, syntactic, and pragmatic functions defined according to the relations established between them. This formation pattern, base plus predication, regularly organizes the hyponyms of the concept *visto* (itself the base of other terms) documented in the record: *visto de cortesia*, in which the predication "de cortesia" ('courtesy') specifies the nature of the grant; *visto temporário*, in which the predication "temporário" ('temporary') specifies the duration of the authorization; *visto diplomático*, in which the predication "diplomático" ('diplomatic') specifies the status of the beneficiary; and *visto de visita*, in which the predication "de visita" ('visit') specifies the purpose of entry into the country. The platform's cross-reference field records this network of hierarchical relations, allowing the user to navigate between the hyperonym *visto* and its hyponyms, all of which were identified from the Lexic.Legs.Imigração-UnB-BR corpus.

5 Concluding remarks

Taken together, the results presented show that the articulation between user-oriented functional lexicography, collaborative digital infrastructure, and terminographic methodological rigor favors the production of resources that are at once accessible to the lay population and socially relevant. The LexiC Platform demonstrates how digital Lexicography and Terminography can function as an instrument of linguistic mediation, social inclusion, and the promotion of linguistic rights in the migration context. By systematically structuring the recording of terminological data and enabling the simultaneous work of multiple researchers in distributed environments, the Platform enhances the production of high-quality specialized glossaries, contributing to the standardization and dissemination of specialized knowledge.

The Computerized Terminological Glossary of Brazilian Migration Legislation, currently under development with the use of this resource, represents a concrete contribution to reducing inequalities in access to legal information for migrants and refugees in Brazil, aligning with the goals of the 2030 Agenda, particularly Sustainable Development Goal 10, Reduced Inequalities, and Target 7, which calls for facilitating orderly, safe, regular, and responsible migration through well-managed policies. The continuation of the project envisages expanding the corpus, increasing the number of terms described, and making the glossary freely available online, consolidating the LexiC Platform as a methodological reference for terminographic projects with social impact.

References

Almeida, G. M. de B. (2006). A Teoria Comunicativa da Terminologia e a sua prática. *ALFA: Revista de Linguística*, 50(2), 85–101.

Barros, L. A. (2004). *Curso básico de terminologia*. EDUSP.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Cabré, M. T. (1993). *La terminología: teoría, metodología, aplicaciones*. Editorial Antártida/Empúries.

Cabré, M. T. (1999). *La terminología: representación y comunicación*. Institut Universitari de Lingüística Aplicada.

Cabré, M. T. (2005). *Terminology: Theory, methods and applications*. John Benjamins.

Faulstich, E. L. de J. (2003). Formação de termos: do constructo e das regras às evidências empíricas. In E. Faulstich & S. P. de Abreu (Eds.), *Linguística aplicada à terminologia e à lexicologia: cooperação internacional — Brasil e Canadá*; (pp. 11–31). UFRGS, Instituto de Letras, NEC.

Finatto, M. J. B., & Krieger, M. G. (2004). *Introdução à Terminologia: teoria e prática*. Contexto.

International Organization for Migration (2024). *World Migration Report 2024*. IOM.

ISO 704 (2022). *Terminology work — Principles and methods*. International Organization for Standardization.

ISO 1087 (2019). *Terminology work and terminology science — Vocabulary*. International Organization for Standardization.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1(1), 7–36.

Maia-Pires, F. de O. (2009). *Brasília em termos: um estudo lexical do Plano Piloto de Brasília* [Master's dissertation, University of Brasília].

Maia-Pires, F. de O. (2024). Um glossário terminológico para a redução da desigualdade dos fluxos migratórios no Brasil. In *Anais do IX Simpósio Mundial de Estudos em Língua Portuguesa & VI Congresso da Associação Internacional de Linguística do Português*. Universidade da Madeira.

Maia-Pires, F. de O., & Cruz, C. L. S. (2021). Mapeamento terminológico da COVID-19: uma proposta de acessibilidade terminológica. In *Anais do XVII Simpósio Iberoamericano de Terminología (RITerm)*.

Maia-Pires, F. de O., Vilarinho, M. de O. M., & Pires, S. L. (2025). *Plataforma Colaborativa de Gestão de Dados Terminológicos — LexiC*. LexiC Research Group, University of Brasília. https://plataforma.lexic.com.br

Tarp, S. (2008). *Lexicography in the borderland between knowledge and information*. Max Niemeyer.

Tarp, S. (2012). Lexicography and user needs. *Lexikos*, 22, 1–18.

Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. John Benjamins.

Contact information

Flávia de Oliveira Maia-Pires
Universidade de Brasília (UnB), Departamento de Linguística, Português e Línguas Clássicas
maiapires@unb.br

Silvio Licurgo Pires
Grupo de Pesquisa LexiC, Universidade de Brasília (UnB)
silviolpires.dev@gmail.com

Rufus H. Gouws & Theo J.D. Bothma

OPTIMISING INTERFACE DESIGN IN ONLINE LEXICOGRAPHIC PRODUCTS

Abstract This paper focuses on certain aspects of interface design in online dictionaries. A discussion of a number of examples of interface design in existing online dictionaries is presented. It is emphasised that successful interface design in transformative lexicographic products should be characterised by a human-centred approach that enables intuitive access to the required data. Dictionary articles should offer users an overview of the available contents. The user should then be able to zoom in on specific data types whilst others could be filtered out so that information on demand can be retrieved. An interdisciplinary approach is needed to optimally utilise the possibilities offered by the online environment to design a dynamic article structure with an interface that gives the users rapid access to specific data. Lexicographic theory should offer an exposition of innovative procedures already prevailing in existing dictionaries. Lexicographic theory should also make proposals that can be applied in the practice to enhance the interface design in order to assist users in experiencing successful dictionary consultation endeavours.

Keywords Individualisation; interface design; screenshot structure; transformative lexicography

1 Introduction

Lexicographers may never assume that it will be unproblematic for their users to find the items they are looking for when consulting a given dictionary. Users differ in their dictionary and reference skills and, in addition to inaccessibility caused by the unnatural syntax prevailing in many dictionary articles, the design of articles is often not user-friendly. The appropriate planning and layout of dictionary articles are essential for successful dictionary use.

Online dictionaries have a dynamic article structure. The interface, which is the point of interaction between the user and the device being used (Tranquillo et al., 2023, p. 237; Du Plessis, 2025, p. 127), with its layout of data is the user's first encounter when consulting a dictionary article. For rapid and efficient access to dictionary content and successful retrieval of the required information, appropriate interface design is of the utmost importance (Du Plessis, 2025, p. 127). Principles of interface design and usability should therefore be integrated when planning a dictionary (Du Plessis, 2025, p. 204).

Focusing on interface design in online dictionaries, several existing general dictionaries are examined, and proposals are made for improvement. In this regard, the interaction between theory and practice is once again important. Lexicographic theory is not theory for the sake of theory; it is directed toward a scientific understanding of dictionaries and the improvement of lexicographic practice. One aim of the interaction between theory and practice in lexicography is the establishment of theory-based dictionaries as well as practice-based theory. The handling of interface design in existing dictionaries creates a point of departure for theoretical reflection on this topic, and this theoretical reflection should then be applied in improving existing dictionaries and planning new lexicographic products. In this article both a contemplative and a transformative approach (Tarp, 2008, p. 10) are followed: existing dictionaries are investigated (a contemplative approach), and proposals for change are made (a transformative approach).

2 Dictionary Users

In both the planning and evaluation of dictionaries, the target user is central. Lexicographers may not take their own preconceived ideas about the dictionary-using skills of target users to determine the presentation of dictionary content. They may not assume that users will know how to use a dictionary correctly and to understand the content of a particular article.

Because of space-saving considerations, traditional printed dictionaries often display unnatural syntax and a high degree of textual condensation. This negatively affects article layout as a basis for unambiguous information retrieval and frustrates users during dictionary consultation.

To solve problems caused by such layouts and data transfer, it could be argued that dictionary-using skills should be improved through appropriate instruction or through extensive user guide texts. This might help, but Tarp & Gouws (2020, p. 4) argue that responsibility for a good consultation experience should shift from the user to the compiler. In line with Norman's (2013) plea for human-centred design of everyday objects, Tarp and Gouws advocate dictionaries characterised by human-centred design enabling rapid and intuitive use. Such an approach prioritises users' needs, abilities, and behaviour. Dictionary design must adapt accordingly.

In the online environment, where dictionaries display dynamic article structures, human-centred design and intuitive use are essential. Contemporary lexicography therefore requires optimised and individualised interface design to ensure fast and successful dictionary consultation - without the obstacles often associated with traditional printed dictionary layouts.

3 Dictionary Content

Where earlier dictionary research focused on documenting the historical development of language and improving the presentation of linguistic items, contemporary research on online lexicographic products focuses more on access to relevant data and ways of satisfying users' information needs. When consulting a dictionary, users are confronted with the interface of the requested article as the gateway to retrieving information from the data presented in the article. Successful dictionary consultation therefore presupposes an optimised interface.

The design of such an interface must enable filtering and in-depth analysis so that users can delve deeper into the article and reach different levels and compartments (Gouws & Bothma, 2024, p. 358) where various data types are accommodated. An optimised interface allows information retrieval on demand, including individualised dictionary use.

Features of dictionaries discussed in this paper are illustrated by means of a number of screenshots. The specific dictionaries from which the screenshots are taken were selected purposively, i.e., in each case to illustrate a specific feature. The examples and choice of dictionaries are not exhaustive but are only exemplars for the purpose of illustration, and similar examples can be found in other dictionaries.

Some online dictionaries have article and microstructures with a layout that largely corresponds to that of printed dictionaries. This static nature of the article structure displays no innovative features. In longer articles, presenting the full article without selective viewing options

leads to findability problems. Online dictionaries must develop innovative methods enabling faster access to specific data through advanced interface design.

Figures 1 and 2 from the *Algemeen Nederlands Woordenboek* (ANW) illustrate a dynamic article structure.



FIGURE 1 From the *Algemeen Nederlands Woordenboek*

The opening screenshot does not display the full article. A short menu lists semantic distinctions of the polysemous word. Clicking on an entry here, e.g., "1.0", leads to a restricted article displaying only the treatment of the selected sense, as seen in Figure 2:

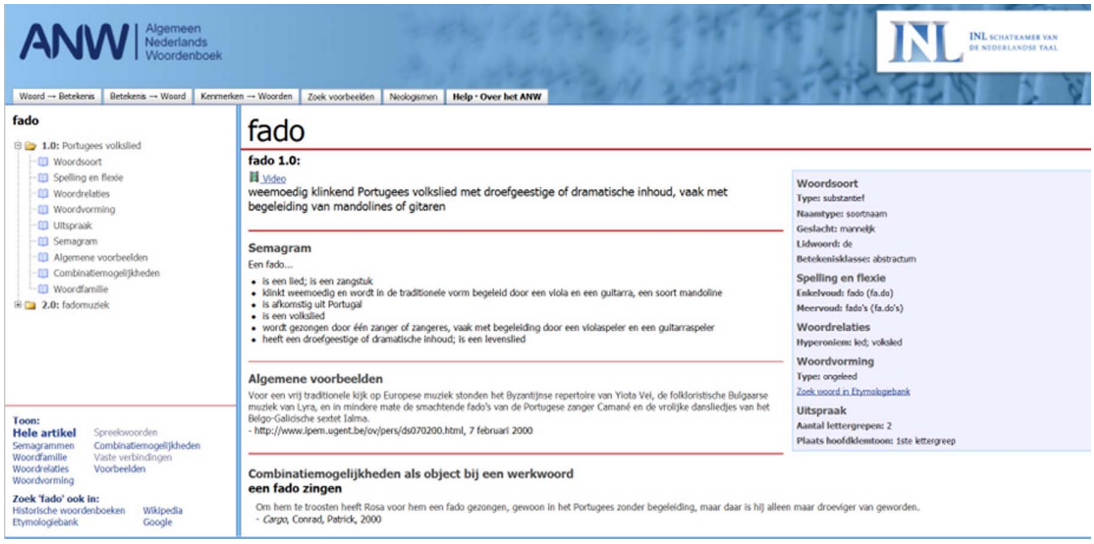


FIGURE 2 From the *Algemeen Nederlands Woordenboek*

Unlike traditional two-dimensional layouts, this interface design promotes rapid access to relevant lexicographic data and supports information retrieval on demand.

4 Interdisciplinarity

Although lexicography is now an independent discipline rather than a subdiscipline of linguistics (Bothma, Gouws & Prinsloo, 2016), an approach isolated from other disciplines is not conducive to the successful development of a lexicographic practice characterised by a transformative approach. Lexicographic products are no longer the result of a single lexicographer's efforts. Contemporary lexicography requires active collaboration across disciplinary boundaries, drawing on insights from information science, computer science, and linguistics to improve both the theory and practice of lexicography.

Influence should also come from broader sources. Norman's (2013) principles of human-centred design emphasise intuitive use and pleasurable interaction. These principles should inform interface design for online dictionaries, reinforcing dictionaries as utility tools.

Insights from Human-Computer Interaction are also relevant. Shneiderman's (1996) "Visual Information-Seeking Mantra" is particularly valuable:

- Overview: Gain an overview of the entire collection
- Zoom: Zoom in on items of interest
- Filter: Filter out uninteresting items
- Details-on-demand: Select an item or group and get details when needed

For lexicography, this means moving from a general overview to specific items, filtering out irrelevant information, and retrieving detailed information on demand. The interface design of a dictionary article should allow vertical navigation between levels and horizontal navigation between compartments (cf. Gouws & Bothma, 2024), enabling follow-up searches.

5 Overview, Zoom, Filter, Details on Demand

Printed dictionaries use typographic and non-typographic structural indicators to guide users through condensed text. Online dictionaries use fewer such indicators due to lower text density, but they employ explicit data identifiers. These identifiers provide both overview and direct access to search zones.

Gouws (2018) distinguishes between comprehensive and restricted articles in online dictionaries. A comprehensive article is an abstract structure that contains the full treatment of a given lemma, but this treatment is presented in different restricted articles - each with the lemma as guiding element but with different data types addressed to the lemma. The comprehensive article is not visible as a whole. Figure 3, part of the opening screenshot of the comprehensive article of the lemma *Maschine* in the German dictionary *elexiko*, does not present the full treatment of this lemma, but it provides access to restricted articles that each deal with a part of the treatment of this lemma.

The interface in this screenshot offers an overview of different data types to be found in the treatment of the word *Maschine*, including different senses. If the item of interest is that sense of *Maschine* indicating an airplane, a click on the element *weiter* (=further) next to the sense

marker *Luftfahrzeug* (=airplane) will enable the user to zoom in on a restricted article containing the treatment of only that sense, as seen in Figure 4.



FIGURE 3 From *elexiko*



FIGURE 4 From *elexiko*

The interface of this restricted article allows access to different data types directed at the specific sense. By clicking on the data-indicator *Konstruktionen* (=constructions) "uninteresting items"



FIGURE 5 From *elexiko*

are filtered out and only details on demand, i.e., different constructions with the word *Maschine*, are shown, as seen in Figure 5.

Shneiderman's mantra can be partially illustrated by the examples in Figures 3–5: In Figure 3, an overview is provided, with the option to drill down on any of the items in blue for more information, Figure 4 illustrates the principles of zoom and filter, and Figure 5 illustrates details on demand. Utilising this appropriate interface design, the user can progress via a vertical shaft structure (Gouws, 2025, p. 60) to a next level that is populated by a single subcomment on semantics. In this level the interface offers data-identifiers that can lead the user horizontally to different compartments in a search tunnel (Gouws & Bothma, 2024, p. 362) where the specific required information can be retrieved.

The human-centred interface design exploits dynamic article structures to provide intuitive access to specific dictionary content.

6 Innovation

An improved interface requires that the structures in the lexicographic database of a dictionary must be designed to ensure the appropriate granularity of the data so that it can be displayed in the required format and the correct search position. To assist users of online dictionaries in achieving a rapid retrieval of information, innovation is needed. Aspects of two innovative procedures are discussed. Firstly, the focus is on aspects in the lexicographic practice leading to changes in the interface design that still require a theoretical explanation.

Here practice is informing theory. Secondly, theoretically-determined suggestions are made for adaptations in interface design to enhance the lexicographic practice. This is theory informing practice.

6.1 From Practice to Theory: Screenshot Structure and Frame

Online dictionaries introduced a new lexicographic structure, namely the screenshot structure (Gouws, 2014, p. 165). The screenshot structure is determined by the genuine purpose of the dictionary and the resulting interface design. Typical dictionary screenshots contain an article, partial article, restricted article or partial article stretch. The screenshot structure is an ordering structure that accounts for the positioning of different data types within the domain of the screenshot.

Dictionary screenshots do not necessarily contain articles. In Figure 6 from *elexiko* a new type of frame structure prevails, with the article of the lemma *Dichter* partially framed by a front matter text, containing the partial article stretch in which the lemma is ordered.

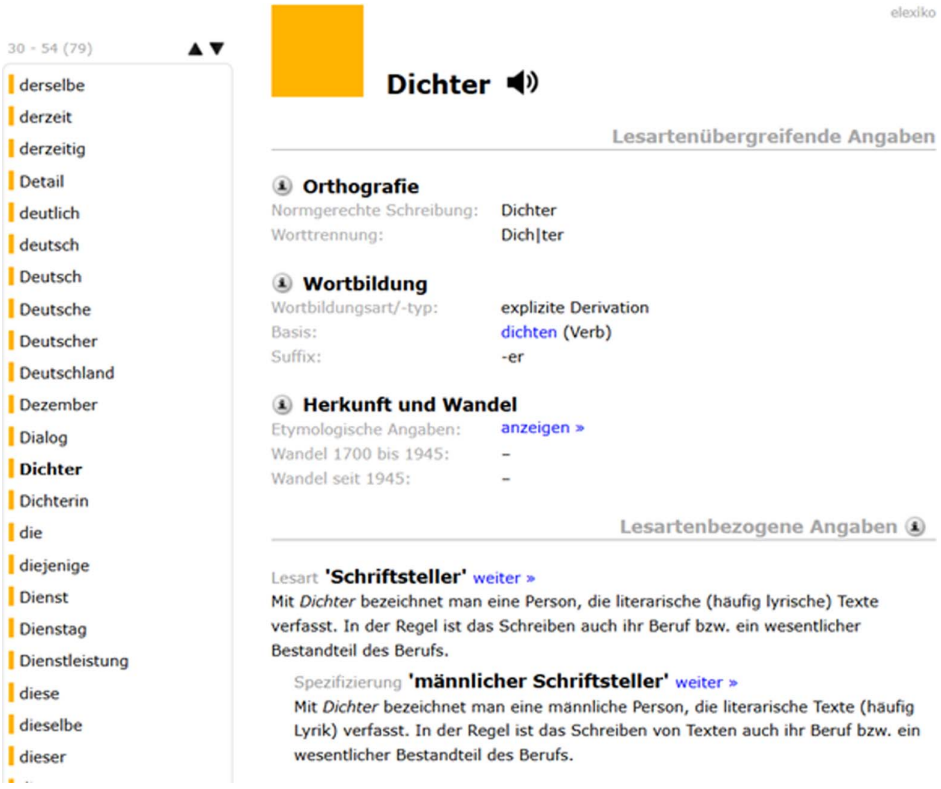


FIGURE 6 From *elexiko*

The contents of this frame text are not part of the article, but the data are relevant to the article because they provide the alphabetical context of the lemma within the macrostructure of the dictionary. This outer text constitutes a partial primary or article frame.

The screenshot structure must also make provision for the occurrence of frame texts that are lexicographically relevant although they are not relevant to the article. Figure 7 shows the opening screenshot when a user consults the *American Heritage Dictionary*:



FIGURE 7 From the *American Heritage Dictionary*

This screenshot also offers horizontally and vertically ordered frame texts that contain lexicographic data (e.g., the dictionary title and aspects regarding the use of the dictionary). These data are not relevant to the article and fall beyond the scope of the article frame. They constitute a secondary frame, also referred to as a search frame (Gouws, 2023, p. 398). A secondary frame can also complement a dictionary article, as seen in Figure 8, where outer texts that are lexicographically relevant but not relevant to the article, frame the article of the lemma *season* by means of a full horizontal frame (with outer texts to the left and the right of the article) and a partial vertical frame (with an outer text above the article):

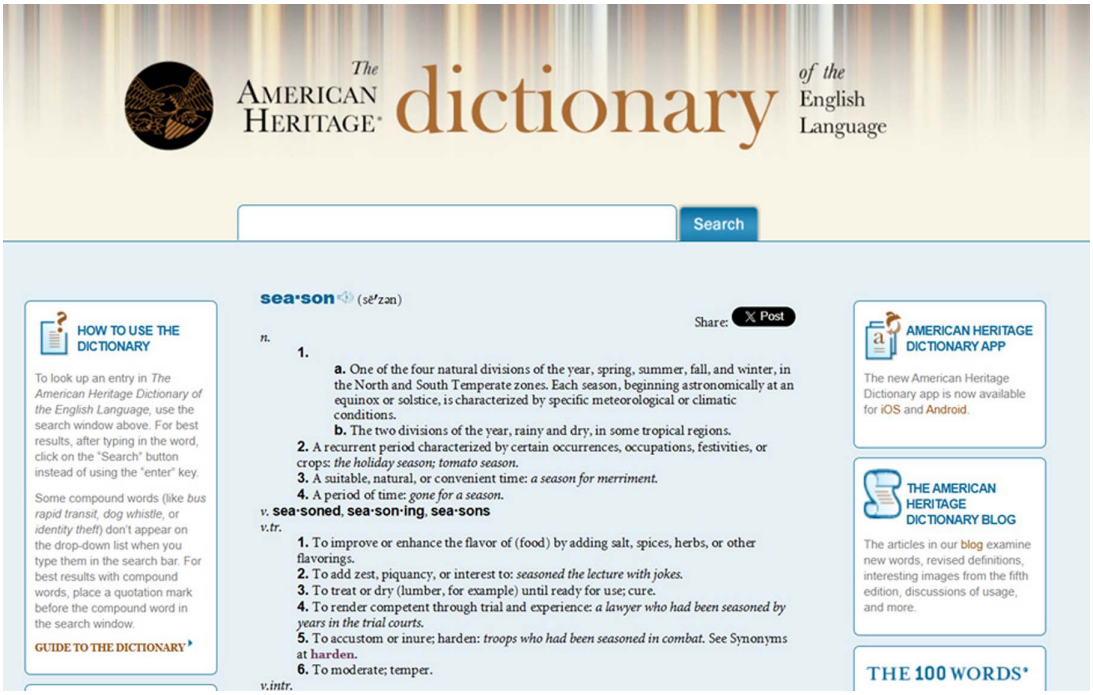


FIGURE 8 From the *American Heritage Dictionary*

Given the importance of sponsors to finance online dictionaries the screenshot structure and the interface design should be planned in such a way that advertisements could be accommodated. These entries may not occupy the primary and secondary frames that are lexicographic search positions, functioning as article frame and search frame respectively. Advertisements are allocated a place in the screenshot frame (Gouws, 2023, p. 401), as seen in Figure 9, a screenshot from *Merriam-Webster* (MW):

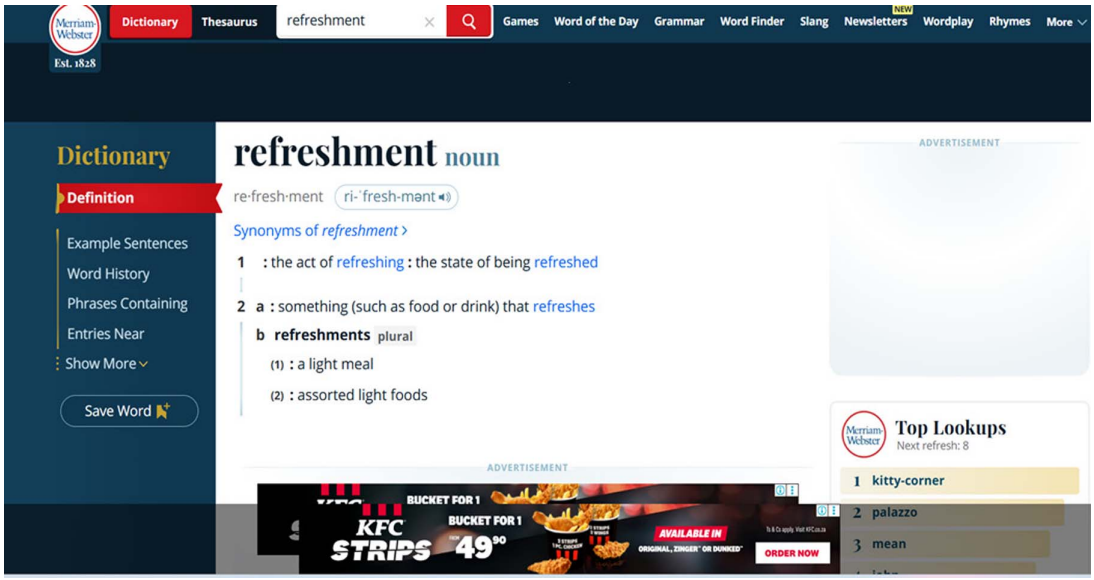


FIGURE 9 From *Merriam-Webster*

To optimally utilise the screenshot structure as an ordering structure the interface design employs new search positions and frame types to enhance user access to the lexicographic data on offer in the article. These innovative interface design features have already been introduced in the lexicographic practice and must now be integrated into lexicographic theory so that future lexicographers could use the adapted theory when planning new strategies for their dictionaries.

6.2 From Theory to Practice: Individualization

In addition to practice-driven innovation, theoretical developments can also serve as a source of innovation in lexicographic practice. One of the most important concepts in this regard is the idea of individualisation, which is closely linked to the principle of human-centred design referred to earlier (Norman, 2013) and intuitive dictionary use. Context-sensitive linking between the interface and the database is conducive to intuitive dictionary use, cf. Bothma & Gouws (2022). Individualisation could then enhance the adaptation of dictionary content and presentation to the specific needs, preferences, and characteristics of individual users.

In traditional printed dictionaries, all users are presented with the same content in the same format, regardless of their level of expertise or their specific information needs. Online dictionaries, by contrast, have the potential to offer a more personalised experience. By making use of the database as underlying source and a dictionary as individualised interface, it is possible to generate different versions of a dictionary from the same data, tailored to different user profiles. Bergenholtz & Bergenholtz (2013) show how four different dictionaries in the field of music can be retrieved from a single database. To go a step further, innovative planning in accordance with the profile of different target users can result in an interface that enables the retrieval of different items of the same data type from the same database. Users may,

for example, create their own profile, cf. Bothma (2011), by specifying their level of language proficiency, their primary purpose for consulting the dictionary (e.g., text reception, text production, or knowledge acquisition), and their preferences regarding the level of detail or the types of information displayed. Based on these preferences, the system can select and present the most relevant items, while filtering out less relevant information. This approach does not only improve usability, but it also enhances the efficiency of information retrieval.

Users differ widely in their information needs and information behaviour, and no user is always consistent in their approach. For a simple information need, a user can easily look up a word in a dictionary (but could be confronted by a long article that needs to be scanned to get to the data to solve the specific information need). An even easier option is to simply type the word into Google, optionally followed by "meaning" - no dictionary access is required. A user who needs highly specific data about a word would prefer a more complex search interface - an advanced interface is available in many online dictionaries, in which the user can specify a number of criteria. One of the most detailed advanced search interfaces by means of which data can be filtered to provide the user with the exact data in terms of their information need can be found in the OED, as illustrated in Figure 10. In addition to the filters provided in the screenshot in Figure 10, there are many additional filters, e.g., 11 parts of speech, 16 types of information, 16 languages of origin, 13 usage categories, 2 obsolescence categories, 6 regional English options and 22 subjects categories. Many categories have additional subcategories.

The idea of the dictionary as a static product is thus replaced by the concept of the dictionary as a dynamic interface to an underlying database. In this model, the database contains a rich set of lexicographic data, while the interface determines how that data are presented to specific users. Different interfaces can be generated from the same database, each tailored to a specific user or group of users or a specific usage situation. This represents a significant shift in the conceptualisation of dictionaries and opens up new possibilities for innovation in lexicographic design.

Recent work by Li & Tarp (2022) and Fuertes-Olivera (2023) further develops this perspective by emphasising the role of adaptive systems in lexicography. Such systems can learn from user behaviour and adjust their output accordingly, providing increasingly personalised and context-sensitive information over time. While these developments are still in an early stage, they point towards a future in which dictionaries are not static reference works, but dynamic and interactive tools that evolve in response to user needs.

Innovation in lexicography therefore involves a continuous interaction between theory and practice. Practical developments, such as new interface designs and data structures, challenge existing theoretical models and create a need for new conceptual frameworks. At the same time, theoretical insights, such as the principles of human-centred design and individualisation, provide guidance for the development of new lexicographic products. By maintaining this dynamic interaction, lexicography can continue to evolve and to meet the changing needs of users in a digital age.

It is to be noted that no specific user interface is recommended, and alternatives are illustrated, especially in Figures 3-6 and 10, where the interfaces of both dictionaries comply in different ways with the principles of Schneiderman's mantra.

User information behaviour studies are paramount in any lexicographic project. Pilot studies will be required to test to what extent any user interface, especially one that makes provision for individualisation and complex advanced search options, would satisfy the information needs of a specific user or user group with advanced information needs.

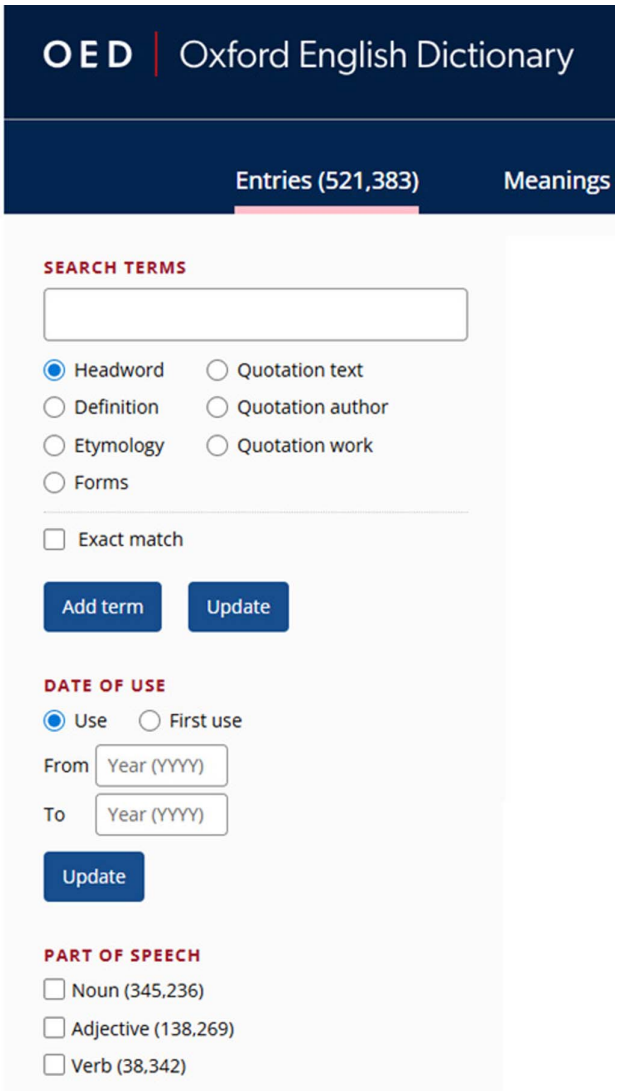


FIGURE 10 From the *OED*: advanced search interface (partial screenshot)

7 Conclusion

Lexicography must be characterized by continuous interaction between theory and practice. Theory should propose improvements to practice, while practical developments should expand theoretical description. Both theoretical and practical lexicography recognise the value of effective interface design. This article identifies innovations already introduced in practice that require theoretical description, as well as proposals from theory that can be applied in practice. A transformative approach – also regarding interface design – is necessary to ensure the successful future use of lexicographic products.

References

Dictionaries

Algemeen Nederlands Woordenboek, https://anw.ivdnt.org/

elexiko, https://www.owid.de/docs/elex/start.jsp

Merriam-Webster Dictionary, https://www.merriam-webster.com/

OED/Oxford English Dictionary, https://www.oed.com/

The American Heritage Dictionary, https://www.ahdictionary.com/

Other Sources

Bergenholtz, H. & Bergenholtz, I. (2013). One data base, four monofunctional dictionaries. *Hermes. Journal of Linguistics*, 50, 119-125.

Bothma, T.J.D. (2011). Filtering and adapting data and information in the online environment in response to user needs. In P.A. Fuertes-Olivera & H. Bergenholtz (Eds.) *e-Lexicography: The Internet, Digital Initiatives and Lexicography* (pp. 71-102). Continuum.

Bothma, T.J.D. & Gouws, R.H. (2022). Information needs and contextualization in the consultation process of dictionaries that are linked to e-texts. *Lexikos*, 32(2), 53-81.

Bothma, T.J.D., Gouws, R.H. & Prinsloo, D.J. (2016). The role of e-lexicography in the confirmation of lexicography as an independent and multidisciplinary field. In T. Margalitadze & G. Meladze (Eds.), *Lexicography and Linguistic Diversity: Proceedings of the XVII EURALEX International Congress* (pp. 109-116). Ivane Javakhishvili Tbilisi State University.

Du Plessis, A.H. (2025). *'n Teoretiese model vir leksikografiese produkte op selfone*. [Doctoral dissertation, Stellenbosch University].

Fuertes-Olivera, P.A. (2023). Structuring a collection of lexicographic data for different user and usage situations. *Lexikos*, 33(2), 22-42.

Gouws, R.H. (2014). Article structures: moving from printed to e-dictionaries. *Lexikos*, 24, 155-177.

Gouws, R.H. (2018). Accessibility, access structures and access procedures. In Jesenšek, V. & M. Enčeva (Eds.), *Wörterbuchstrukturen zwischen Theorie und Praxis* (pp. 35-56). De Gruyter.

Gouws, R.H. (2023). Nuwe raamstruktuurtipes in aanlyn woordeboeke. *Lexikos*, 33(1), 382-403.

Gouws, R. H. (2025). 'n Omvattender aanbod van semantiese data in aanlyn woordeboeke. *Lexikos*, 35(2), 46-74.

Gouws, R.H. & Bothma, T.J.D. (2024). Aspekte van inligtingsonttrekkingstrukture in aanlyn woordeboeke. *Lexikos*, 34(1), 355-370.

Li, Q. & Tarp, S. (2022). Differentiated treatment of cultural items in lexicographical products: a necessary adaptation to the digital environment. *Lexikos*, 32, 90-117.

Norman, D. (2013). *The Design of Everyday Things*. Basic Books.

Shneiderman, B. (1996). The eyes have it: a task by data type taxonomy for information visualizations. In *Proceedings 1996 IEEE Symposium on Visual Languages* (pp. 336-343). Boulder, CO.

Tarp, S. (2008). *Lexicography in the borderland between knowledge and non-knowledge. General lexicographical theory with particular focus on learner's lexicography*. Niemeyer.

Tarp, S. & Gouws, R.H. (2020). Reference skills or human-centered design: towards a new lexicographical culture. *Lexikos*, 30, 470-498.

Tranquillo, J., Goldberg, J. Allen, R. (2023). *Biomedical Engineering Design*. Academic Press.

Contact information

Rufus H. Gouws
Department of Afrikaans and Dutch, Stellenbosch University, South Africa
rhg@sun.ac.za

Theo J.D. Bothma
Department of Information Science, University of Pretoria, South Africa
theo.bothma@up.ac.za

David Lindemann

DEVELOPING A DIGITAL VERSION OF GIDOR BILBAO'S LATIN–BASQUE DICTIONARY

Abstract In this article, we describe several steps taken to convert Gidor Bilbao's *Oinarrizko Hiztegia Latina-Euskara* (Basic Latin-Basque Dictionary) from a paper-based resource and move toward presenting it as Linked Data. Starting from a document in MS Word format, we attempted to isolate and tag elements of the dictionary's macro- and microstructures. The goal of the experiments outlined here was to capture the content of those elements in an innovative format based on the Linked Data paradigm, using the Ontolex-Lemon model. Once integrated into Wikibase, the data are ready for final structuring or further enrichment.

Keywords Dictionary digitization; linked data; Wikibase; Latin; Basque

1 Introduction

In this article, we present several experiments carried out to extract, remodel, and link to other resources Gidor Bilbao's *Oinarrizko Hiztegia Latina-Euskara* (Basic Latin-Basque Dictionary, henceforth LBD). The dictionary was published in print after six years of work (Bilbao Telletxea, 2020) and has since been updated by the author, who has made the data available to us to carry out the experiments presented here. Our goals have been the following:

- To incorporate the dictionary into a software platform enbaling continued enrichment through collaborative work on the data;
- To give the dictionary text a machine-interpretable structure;
- To link the dictionary with the databases of the *Linking Latin* (LiLa) project, as well as with Wikidata.

In the remainder of this article, we describe the work carried out. Section 2 provides details on the restructuring of the original dictionary text according to its layout markup, and Section 3 outlines the steps for defining links with the LiLa database. In order to make the data accessible to anyone, and enable continued editing by authorized users, we have stored the entire dataset in a Wikibase platform, as explained in Section 4. We conclude with an outlook towards future work with this data. Manual curation could not be carried out at this initial stage where only a rule-based conversion has been applied; what is presented here is thus only a starting point.

Making dictionary contents openly available for third parties to experiment with is not at all common; our sincere thanks to the author for his generosity. We also extend our warmest thanks to our colleagues Marco Passarotti, Paolo Ruffolo, and Giovanni Moretti from the *Linking Latin* project for processing the lemmas with the project's tools.

2 Structuring the original dictionary

In January 2024, Gidor Bilbao provided us with the latest edited version for use in the experiments presented here. It is an MS Word file in *.doc* format of 5.7 megabytes, which would correspond to 547 pages if printed on A4 paper. As shown in the example in Figure 1, the dictionary has a layout made up of different colors and font styles.

It is evident that each dictionary entry takes up one paragraph, each paragraph beginning with the headword of a new entry. Thus, the first challenge in structuring—isolating entries from one another—is easily solved, since paragraph breaks clearly mark the boundaries between entries.

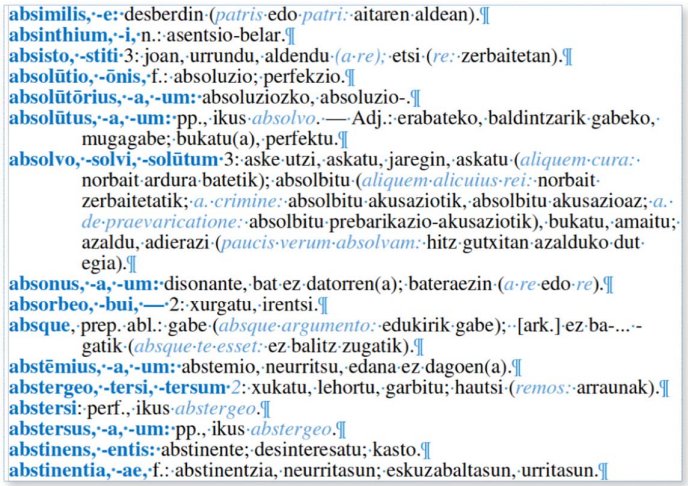

absimilis,-e: desberdin (*patris* edo *patri:* aitaren aldean).¶
absinthium,-i, n.: asentsio-belar.¶
absisto,-stiti 3: joan, urrundu, aldendu (*a re);* etsi (*re:* zerbaitetan).¶
absolūtio,-ōnis, f.: absoluzio; perfekzio.¶
absolūtōrius,-a,-um: absoluziozko, absoluzio-.¶
absolūtus,-a,-um: pp., ikus *absolvo*.— Adj.: erabateko, baldintzarik gabeko, mugagabe; bukatu(a), perfektu.¶
absolvo,-solvi,-solūtum 3: aske utzi, askatu, jaregin, askatu (*aliquem cura:* norbait ardura batetik); absolbitu (*aliquem alicuius rei:* norbait zerbaitetatik; *a. crimine:* absolbitu akusaziotik, absolbitu akusazioaz; *a. de praevaricatione:* absolbitu prebarikazio-akusaziotik), bukatu, amaitu; azaldu, adierazi (*paucis verum absolvam:* hitz gutxitan azalduko dut egia).¶
absonus,-a,-um: disonante, bat ez datorren(a); bateraezin (*a re* edo *re*).¶
absorbeo,-bui,— 2: xurgatu, irentsi.¶
absque, prep. abl.: gabe (*absque argumento:* edukirik gabe); [ark.] ez ba-...-gatik (*absque te esset:* ez balitz zugatik).¶
abstēmius,-a,-um: abstemio, neurritsu, edana ez dagoen(a).¶
abstergeo,-tersi,-tersum 2: xukatu, lehortu, garbitu; hautsi (*remos:* arraunak).¶
abstersi: perf., ikus *abstergeo*.¶
abstersus,-a,-um: pp., ikus *abstergeo*.¶
abstinens,-entis: abstinente; desinteresatu; kasto.¶
abstinentia,-ae, f.: abstinentzia, neurritasun; eskuzabaltasun, urritasun.¶

FIGURE 1 Original MS Word *.doc* format on screen

As the sample shows, in the blue bold text at the beginning of the entry, the canonical form of the lemma is given first, along with abbreviations indicating the inflection paradigm. From these inflected abbreviations, anyone with basic knowledge of Latin can deduce the grammatical category of that lemma. For example, *-i n.* or *-onis f.* correspond to nouns, while abbreviation pairs such as *-a, -um* or *-ens, -entis* appear in entries for participles (which are redirected to the entry presented under the infinitive) or for adjectives. Those with three forms and a number indicating the conjugation paradigm are verb entries. For these major grammatical categories, no explicit category marker is given. For other categories, however, the abbreviation corresponding to the grammatical category is provided, for example, *prep. abl.* (a preposition requiring the ablative). In addition, *perf.* (first person perfect) and, as mentioned, *pp.* (perfect participle) also appear as headwords, where the entire entry content is a redirect to the entry for the corresponding infinitive, unless an adjectival use is specified. This decision regarding the dictionary's macrostructure responds to the learner's need, as they may have difficulty determining the infinitive of verbs (especially irregular ones) and also expect the dictionary to indicate whether a participle can be used as an adjective. In general, in the case of participles, it may be questionable whether to provide a separate entry for the adjective when the adjective is semantically identical to the participle; this is another argument for including participles in the lemma list.

The overall entry structure is as follows, as seen in Figure 1: After the headword with the lemma and the components described above, the Basque equivalents are given by sense. Semicolons are used between senses, and within a single sense, commas are used to separate different equivalents. Additional notes appear in parentheses within a sense, sometimes including a translated usage example (e.g., *absque argumento* – "edurikirik gabe", *absque te esset* – "ez balitz zugatik"), and other times notes on proper usage, providing collocations (e.g., *remos* – "arraunak" in the entry *abstergeo* under the sense translated with the verb "hautsi"), or indicating the required preposition or case (e.g., "*a re* or *re*" for the adjective *absonus* in the sense translated with "bateraezin").

The task described here is to obtain a tagging of the dictionary's structure based on the layout features —from layout markup to structural markup— so that the dictionary's structure is

machine-interpretable, where each component of an entry tells the machine what it is: lemma, equivalent, example, etc. Regarding the macrostructure, as mentioned, we have no problem thanks to the paragraph breaks between entries. Isolating the lemma is also straightforward: In the blue bold text at the beginning of the paragraph, the lemma itself appears together with the previously mentioned inflection markers, sometimes complemented with an additional note in brackets, and then followed by a colon. Following that, and inside the entry, semicolons serve as a clear boundary for separating senses, and colors are essential for interpreting the additional information within parentheses.

To make use of this layout markup present in the *.doc* file, we first converted it to LibreOffice format, as the layout markup code can be displayed and reused in the latter but not in the former. The *.odt* file produced by LibreOffice is a *.zip* archive, and the format in which the application saves the document appears inside the archive under the name *content.xml*. The layout markup code appears as shown in Figure 2. There are many layout changes within an entry, and each segment using the same font appears within an XML element called <text:span>. Font commands (bold, blue, italic . . .) are not visible; instead, text-style codes appear. For example, the style with code "T93" (the bold blue style of the headword) is declared and defined earlier in the document, to which this code is a reference. We see that style "T93" is used in the headword; however, we have already located the beginning of each entry thanks to the paragraph break, since the <text:p> element contains the paragraph.

```
name="T93">absimilis, -e: </text:span><text:span text:style-
name="T50">desberdin </text:span><text:span text:style-
name="T51">(</text:span><text:span text:style-name="T385">patris</
text:span><text:span text:style-name="T50"> edo </
text:span><text:span text:style-name="T385">patri: </
text:span><text:span text:style-name="T51">aitaren aldean).</
text:span></text:p><text:p text:style-name="P24"><text:span
text:style-name="T93">absinthium, -i, </text:span><text:span
text:style-name="T50">n.: asentsio-belar.</text:span></
text:p><text:p text:style-name="P24"><text:span text:style-
name="T93">absisto</text:span><text:span text:style-name="T92">, </
text:span><text:span text:style-name="T128">-</text:span><text:span
text:style-name="T92">stiti </text:span><text:span text:style-
name="T52">3:</text:span><text:span text:style-name="T50"> joan,
urrundu, aldendu </text:span><text:span text:style-name="T385">(a
re);</text:span><text:span text:style-name="T50"> etsi </
text:span><text:span text:style-name="T51">(</text:span><text:span
text:style-name="T385">re: </text:span><text:span text:style-
name="T51">zerbaitetan).</text:span></text:p><text:p text:style-
name="P24"><text:span text:style-name="T93">absolūtio, -ōnis, </
```

FIGURE 2 LibreOffice xml format, inside the *.odt* archive

It is not clear to us what the remaining style codes are associated with, at least from looking at Figure 2. Indeed, LibreOffice defines hundreds of styles at the beginning of this document. This is because, while editing the dictionary, styles were assigned manually and individually —for example, adding blue color and then italics to black text— rather than reusing an already known style from the styles section. Thus, MS Word (the original program used) treated the manually applied layout commands as innovations. Therefore, identic styles appear associated to several codes. For instance, headwords have not only style "T93" but also others, such as "T92" and "T135". However, despite the duplication of styles, we know that the same code does not appear with Basque and Latin content together; that is, some codes will appear with Basque and some with Latin content within the same entry.

Having reached this point in the analysis, we transformed the XML format provided by LibreOffice to clarify what we know so far and to begin interpreting the style codes. Figure 3 shows the remodeled version.

```
<entry><element type="lemma">absimilis, -e: </element><element
type="T50">desberdin </element><element type="T51">(</element><element
type="T385">patris</element><element type="T50"> edo </element><element
type="T385">patri: </element><element type="T51">aitaren aldean).</element></
entry>
<entry><element type="lemma">absinthium, -i, </element><element type="T50">n.:
asentsio-belar.</element></entry>
<entry><element type="lemma">absisto</element><element type="T92">, </
element><element type="T128">-</element><element type="T92">stiti </
element><element type="T52">3:</element><element type="T50"> joan, urrundu,
aldendu </element><element type="T385">(a re);</element><element type="T50">
etsi </element><element type="T51">(</element><element type="T385">re: </
element><element type="T51">zerbaitetan).</element></entry>
<entry><element type="lemma">absolūtio, -ōnis, </element><element
type="T50">f.: absoluzio; perfekzio.</element></entry>
<entry><element type="lemma">absolūtōrius, -a, -um:</element><element
type="T50">absoluziozko, absoluzio-.</element></entry>
<entry><element type="lemma">absolūtus, -a, -um: </element><element
type="T265">pp., ikus </element><element type="T406">absolvo</element><element
type="T265">. </element><element type="T497">– Adj.: erabateko, baldintzarik
gabeko, mugagabe; bukatu(a), perfektu.</element></entry>
<entry><element type="lemma">absolvo</element><element type="T92">, </
element><element type="T128">-</element><element type="T92">solvi, </
element><element type="T128">-</element><element type="T92">solūtum </
element><element type="T51">3</element><element type="T52">:</element><element
type="T50"> aske utzi</element><element type="T473">, askatu, jaregin,
askatu</element><element type="T51">(</element><element type="T385">aliquem </
element><element type="T470">cura: </element><element type="T475">norbait
ardura bateti</element><element type="T51">k);</element><element type="T50">
absolbitu </element><element type="T51">(</element><element
type="T385">aliquem alicuius rei</element><element type="T470">: </
element><element type="T475">norbait zerbaitetatik; </element><element
type="T470">a.</element><element type="T385"> crimine</element><element
type="T470">: </element><element type="T475">absolbitu akusaziotik, absolbitu
akusazioaz; </element><element type="T470">a.</element><element type="T470">de
praevaricatione: </element><element type="T475">absolbitu prebarikazio-
```

FIGURE 3 Remodeled xml

After examining which style code is associated with which type of content, we were able to determine certain structural aspects: language changes within the microstructure, and the beginnings and ends of cross-references. Additionally, we assigned a number to each entry in order not to lose reference to each entry when recording its parts one by one in Wikibase.

The following structural elements have been isolated at this point: the entry itself (now with its number), the initial lemma and inflection paradigm indicators within the entry, and, within the entry, the Latin and Basque parts separated, grouping together consecutive segments of the same language. We have preserved the style codes corresponding to each part, as shown in Figure 4, just in case the nature of the part might be determined through them.

```
<entry entry_id="83">
    <lemma type="lemma">absolūtus, -a, -um: </lemma>
    <xref_intro type="T265">pp., ikus </xref_intro>
    <xref_lemma type="T406">absolvo</xref_lemma>
    <xref_intro type="T265">. </xref_intro>
    <basque type="T497">– Adj.: erabateko, baldintzarik gabeko, mugagabe; bukatu(a),
    perfektu.</basque>
</entry>
<entry entry_id="84">
    <lemma type="lemma T92 T128 T92 T128 T92">absolvo, -solvi, -solūtum </lemma>
    <basque type="T51 T52 T50 T473 T51">3: aske utzi, askatu, jaregin, askatu(</basque>
    <latin type="T385 T470">aliquem cura: </latin>
    <basque type="T475 T51 T50 T51">norbait ardura batetik); absolbitu (</basque>
    <latin type="T385 T470">aliquem alicuius rei: </latin>
    <basque type="T475">norbait zerbaitetatik; </basque>
    <latin type="T470 T385 T470">a. crimine: </latin>
    <basque type="T475">absolbitu akusaziotik, absolbitu akusazioaz; </basque>
    <latin type="T470 T470">a.de praevaricatione: </latin>
    <basque type="T475 T51 T50 T473 T50 T473">absolbitu prebarikazio-akusaziotik),
    bukatu, amaitu; azaldu, adierazi (</basque>
    <latin type="T469">paucis verum absolvam: </latin>
    <basque type="T473">hitz gutxitan azalduko dut egia)</basque>
</entry>
```

FIGURE 4 Style codes linked to the language of the content

3 Alignment with LiLa

The ERC project *Linking Latin* (LiLa), based at the Catholic University of the Sacred Heart in Milan, is a pioneer in capturing corpus content, dictionary content, and other lexical data according to the Linked Data paradigm.[1] They have linked a large set of canonical classical Latin texts and several Medieval Latin texts, on the one hand, and several bilingual Latin dictionaries, on the other, with the LiLa Lemma Bank as the central resource (Passarotti et al., 2020; Passarotti & Mambrini, 2022). The Lemma Bank is a collection of lemmas and hypolemmas equipped with URIs.[2] An hypolemma is "a citation form that belongs to a word's regular inflectional paradigm but receive a different PoS tag or degree of comparison than their 'most canonical' lemma, including participles, gerundives, deadjectival adverbs, and comparatives" (Passarotti et al. 2025). For example, the adjective *abstersus*, which is morphologically the participle of the verb *abstergeo*, is a hypolemma of that verb.[3]

We are in contact with members of the LiLa project (see Lindemann et al., 2023, 2025), and they have carried out an effort to link the list of headwords we isolated to the LiLa Lemma Bank using the algorithms used in the project. This set of algorithms has been used to link several textual and lexical resources (including dictionaries) to the Lemma Bank (see, for example, Mambrini et al., 2021). These algorithms start from a spelling standardization: the letter *v* becomes *u*, and vowel length markers and other diacritics (often present in dictionary lemmas) are removed (*aestīvus* > *aestiuus*). Since the LiLa team have before processed other dictionaries that present lemmas together with inflection paradigm indicators like *-a, -um*, the algorithm is able to convert inflection paradigm indicators into tags from the Universal PoS tagset as grammatical category indicators. This linking effort, based on lemma and part grammatical category, applied to our list of 17,134 headwords, has had the following results:

(1) 15,329 headwords (89.4%) were linked unambiguously to the Lemma Bank;

(2) 1,690 headwords (9.9%) were linked ambiguously to the Lemma Bank, meaning the linking algorithm returned more than one candidate;

(3) 126 headwords (0.7%) could not be linked to the Lemma Bank.

Group (1) has an equivalent in the LiLa Lemma Bank with the same spelling and a single category; group (2), however, has more than one equivalent according to the same criteria. For group (3), the algorithm proposed no equivalent; 72 of these 126 entries contain the word *ikus* ("see") in the entry text —that is, they are redirect entries— and most of these (43) are generic redirects, such as "adc- . . . : *ikus* acc . . . ".

4 Transforming the dictionary's contents into Linked Data

4.1 Recording entries in Wikibase

Wikibase[4] is a software platform for storing, editing, and publishing datasets according to the Linked Data paradigm. Lexical data are modeled in a Wikibase according to the Ontolex-Lemon standard (McCrae et al., 2017); details are given in Lindemann (2025). This same

1 Project home page accessible at https://lila-erc.eu/

2 Uniform Resource Identifier, the type of identifier used in the Linked Data paradigm, see https://en.wikipedia.org/wiki/URI

3 The identifier for the adjective *abstersus* is http://lila-erc.eu/data/id/hypolemma/242

4 See https://wikiba.se

standard is used by the LiLa project and is also the basis for Wikidata.[5] The standard allows lexical datasets to be linked to each other, as it requires the use of URIs at three levels: the entry level, the sense level, and the form level. That is, a lexical entry, a sense within an entry, and an inflected form can have direct relations set within the same dataset or from one resource to another. These links must be typed and identified themselves with URIs; one type of link can indicate synonymy, another translation equivalence, another one links a lexical entry to its corresponding inflected forms, and so on.

Remodeling the dictionary's contents into the format required by a Wikibase enables linking to other resources, even if those resources are physically stored in different databases. Furthermore, the Wikibase software allows its content to be edited manually or programmatically, with the possibility of reviewing and eventually undoing changes made by users. That is, this software is very suitable for our case, as it provides options for performing and for reviewing the editing steps that need to be done after the rule-based transformation of the content, either algorithmically or manually.

We have previously carried out several experiments in a Wikibase instance called *Monumenta Linguae Vasconum (Lindemann & Alonso 2024)*[6] in the same instance, we have created an OntoLex lexical entry for each entry in Gidor Bilbao's dictionary. As lemma at the entry level, we applied normalization according to the criteria of the LiLa project.

As shown in Figure 5, the entries we created have the following three parts:

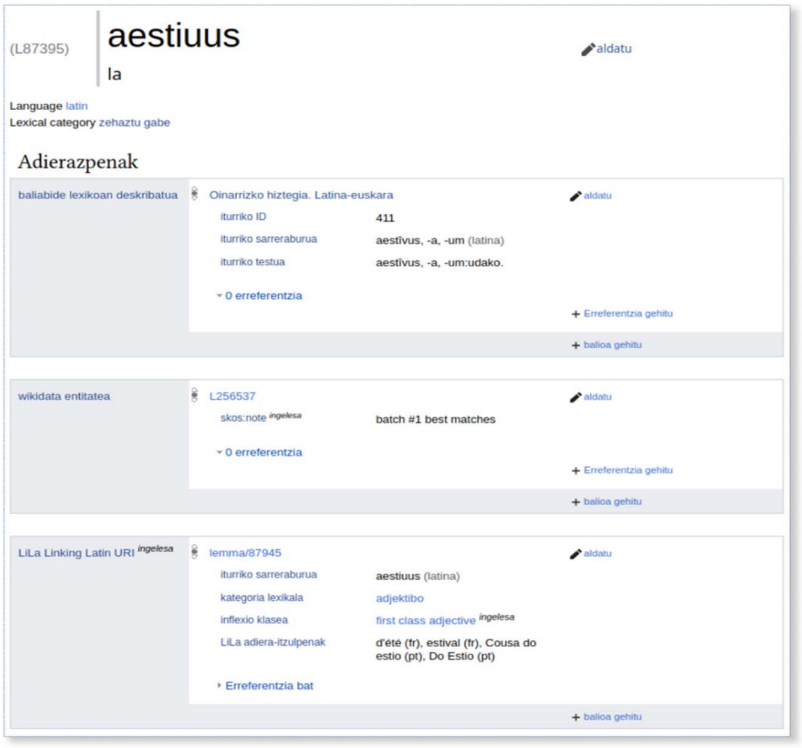


FIGURE 5 Cross-resource entry relation statements on *aestiuus*[9]

[5] See https://wikidata.org

[6] See https://monumenta.wikibase.cloud

(1) A statement corresponding to the LBD dictionary itself. Here, we applied the criterion used in Wikidata,[7] i.e., we created a property in Wikibase called *described in resource*, because a lemma can in principle be described not only in this dictionary but also in other dictionaries, and this would most probably be the appropriate representation in Wikidata, if the statement were transferred to the equivalent Latin lexeme in that database. In this part of the entry, we therefore indicate that the above lemma is mentioned in Gidor Bilbao's dictionary, and we add several qualifiers to this statement: the entry number, the sequence of signs used as the headword in the source, and the full text of the entire entry. Thus, all 17,134 entries of the dictionary have a representation in Wikibase.[8]

(2) A statement for reference to LiLa, through a link to the LiLa URI. As a consequence, one can directly access the information that the LiLa database holds about the lemma, either manually by clicking on the URI link, or programmatically. LiLa identifiers share the same namespace prefix (http://lila-erc.eu/data/id/), which is followed by the identifier itself, which changes from lemma to lemma; clicking on it takes the user to the LiLa portal. The content of LiLa can also be accessed programmatically, and in this way we have obtained the values of the qualifiers added in this statement section: the LiLa lemma and grammatical category, inflection classes, and translation equivalents in several languages.[10] The latter are those appearing in dictionary entries that the LiLa project has linked to the Lemma Bank; we have got translation equivalents in Spanish, Portuguese, English, and French, in order to support the review process of ambiguously linked lexemes, where users decide which one is the correct link, and which one is to be set as deprecated.

(3) A statement for a reference to Wikidata. Most of the LiLa Lemma Bank entries are aligned with Wikidata[11], meaning that many of them are also described on that platform. This enables us to enrich the Wikibase entry with additional content or reach that content through federated database queries. In addition, in the Wikibase interface, as described for the LiLa identifier, the user can navigate to the Wikidata portal by clicking. One of the special features of Wikidata's lexical collections is that senses appear linked to ontological entities, which to date is not the case in LiLa.[12]

4.2 Approaches to sense modeling

As explained above, in the entries, Basque equivalents appear grouped by sense, with semicolons separating senses and commas separating different equivalents within a sense. Additional notes appear within parentheses. In principle, this layout structure provides a solid basis for isolating senses and adding them as Ontolex-Lemon sense objects to the lexical entries in Wikibase. These objects will each have a URI, allowing both internal and external objects to be linked to them.

7 See the links to descriptions in different resources on the Danish adjective *rød on Wikidata at* https://www.wikidata.org/wiki/Lexeme:L2310#P1343

8 See https://tinyurl.com/23uzp3rp for a query that lists all entries.

9 See https://monumenta.wikibase.cloud/entity/L87395, *aestiuus*.

10 We have used a query like https://tinyurl.com/2868bj96 for retrieving those data from LiLa.

11 An *external-id* property pointing to LiLa is used in Wikidata; see www.wikidata.org/wiki/Property_talk:P11033

12 See, for example, the link of the first sense of the noun *nux*, pointing to the entity describing the edible fruit of the walnut tree: https://www.wikidata.org/wiki/Lexeme:L279629#S1

We have defined an algorithm to isolate senses and translation equivalents from the dictionary text. As mentioned, semicolons mark boundaries between senses, but semicolons can also appear within parenthetical note elements. Therefore, to isolate senses, parenthetical elements must be considered first, so that semicolons appearing within them are not treated as sense separators. The elements isolated by this set of rules appear as shown in Figure 6 in Wikibase:

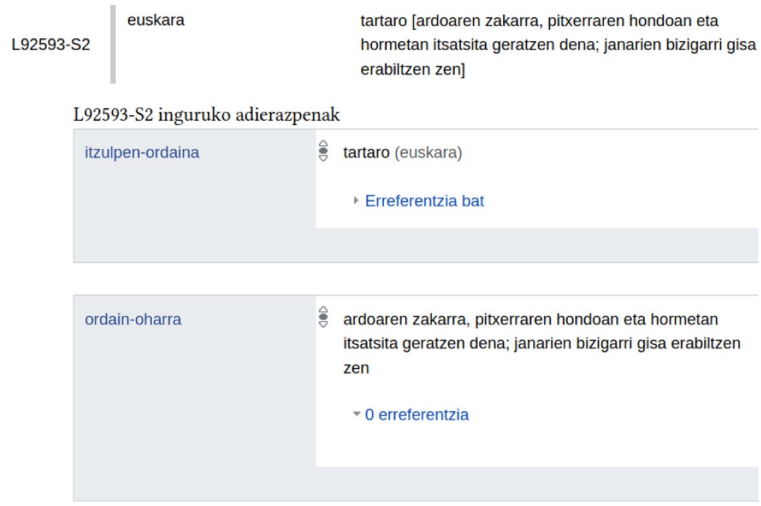


FIGURE 6 Second sense of the entry *faex, faecis, f.*[13]

The results[14] show that we have not succeeded for all entries with a more complex structure in fully and correctly structuring the content using rules alone. For 17,134 entries, we have been able to model 29,443 sense objects. 5,922 entries remain without sense object; the vast majority of these are redirect entries which also in the original dictionary text contain no sense description. One problem is the ambiguous application of layout within an entry, such as the fact that additional notes given alongside translation equivalents are not always placed in parentheses. For correctly interpreting such cases, it may be necessary to refine the applied rule set further. We face the following decision: How far will we go by further refining our programmatic rule set, and how far will we go by manually correcting the errors? These are questions to be resolved collaboratively, in a collaboration that will follow this report.

4.3 Preliminary results and future work

All entries of Gidor Bilbao's dictionary have been stored in the *Monumenta Linguae Vasconum* Wikibase instance.[15] We now have the entire text of each entry stored, and by entering either the Latin lemma or one of the Basque equivalents into the search field, entries can be found and viewed.[16] The vast majority of entries have been unanmbiguously linked to

[13] Address of the sense in Wikibase: https://monumenta.wikibase.cloud/entity/L92593-S2

[14] See the results of the database query at https://tinyurl.com/26zfz7v4 to view the sense objects.

[15] See the page that gives access to the data we present here: https://monumenta.wikibase.cloud/wiki/Oinarrizko_hiztegia._Latina-Euskara

[16] To search in Wikibase, go to https://monumenta.wikibase.cloud/w/index.php?ns146=1&search= Entering a lemma or equivalent in the search field returns a list of entries containing it.

the LiLa Lemma Bank. Whether these links are correct can be easily reviewed in the Wikibase interface: In a table provided by a specific query,[17] the data brought from LiLa (grammatical category, inflection paradigm, equivalents in third languages) can be compared with the original dictionary entry, and if a link is found to be incorrect, it can be invalidated by going to the corresponding dictionary entry page and removing the link, or setting it as deprecated, and replacing it with the correct one. Similarly, for those linked ambiguously,[18] one can go to the entry page and remove incorrect links or mark them as deprecated. According to the second option, the invalidated link will not disappear but will be expressly marked as invalid. In this way, we can help evaluate the LiLa project's algorithm for linking lemma signs to the Lemma Bank by providing these annotations.

On the other hand, LiLa and Wikidata content can be used to enrich Gidor Bilbao's dictionary. For example, to decide whether entries describing participles should be described as adjectives or not, one can review the decision taken for other dictionaries linked to LiLa.[19] Wikidata provides, for example, sense descriptions attached to Latin lexemes that are connected to ontological items,[20] and through both of these, links to other lexical or encyclopaedical resources.

For entries that have sense objects, those can be linked to senses of Basque lexemes recorded in Wikidata. One method for this is to search Wikidata for Basque translation equivalents using a federated database query, and then verify the correct links from the results.[21] The senses of Basque lexemes described in Wikidata come from resources managed by the Elhuyar Foundation, a major dictionary publisher in the Basque Country, and from those Wikidata entries users may jump to the Elhuyar dictionary web portal.[22] At the same time, various senses of Basque lemmas in Wikidata are linked to Wikidata's ontology, and so it is in the case of Latin lexemes, as discussed above. That is, linking Gidor Bilbao's translation equivalents at the sense level to Wikidata brings data enrichment in several regards.

Figure 7 shows several links added from the entry *abolitio* to the LiLa and Wikidata databases; these links can be used to import data. The LiLa database offers a wide range of content, such as corpus attestations of lemmas. Furthermore, as an extension of the LiLa project, the LiLaMorph Wikibase instance describing Latin verb forms is available (Lindemann et al., 2025); it also uses LiLa Lemma Bank identifiers, which makes the linking straightforward. Wikidata also adds value to the dataset at hand, and the way to do this is to define specific queries. Overall, the enrichment of entries can be described as follows:

- via the lemma's identifier in LiLa, one can access the lemma's corpus attestations, as well as entries in several linked dictionaries;

17 The query at https://tinyurl.com/22v3aegn returns a list of entries that have a single LiLa link.

18 The query at https://tinyurl.com/25jfe5gp returns a list of entries that have more than one LiLa link.

19 The query at https://tinyurl.com/24awrx2o brings up these cases: entries for participles that are described as adjectives in LiLa.

20 The query at https://w.wiki/QVt5 shows Wikidata Latin lexemes with links of type *item for this sense*, which is equivalent to `ontolex:reference`.

21 For example, see the results of the federated query at https://w.wiki/CwWf

22 The Basque equivalents in the LBD come from the monolingual learner's dictionary *Ikasleen Hiztegia* (Elhuyar, pers. comm.).

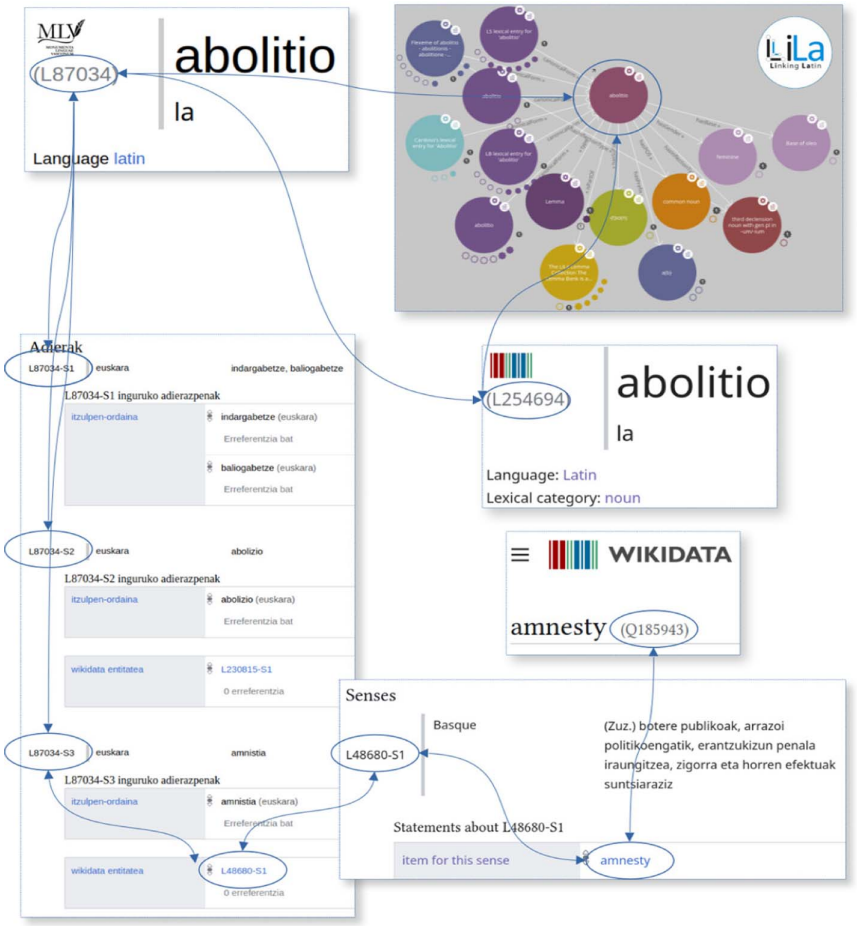


FIGURE 7 Entry *abolitio* and its relations[23]

- we can bring senses and inflected forms corresponding to a lemma from Wikidata;[24]
- we can bring the inflected forms corresponding to a lemma from the newly developed LiLaMorph Wikibase;[25]
- the link to a Basque Wikipedia article explaining the concept denoted by a sense associated to a translation equivalent can be brought from Wikidata.[26]

Both the structuring of the LBD dictionary text and the alignment with LiLa resources have been carried out using rules. As explained, in both areas we are left with content that needs manual correction. Not all content in this rich dictionary text could be fully structured by only taking the layout markup into account, and the ambiguities of the lemma set with respect to LiLa's lemma set also remain to be reviewed manually. For reference in any case, the original entry text has been recorded in each Wikibase entry.

By choosing Wikibase, a platform suitable for collaborative work, the review tasks can be done collaboratively in a peer-review process, as all users have the option to be notified of any change they might want to oversee.

23 The entry *abolitio* in Wikibase: https://monumenta.wikibase.cloud/entity/Q131

24 For inflected forms, see the results of the query at https://tinyurl.com/29rl9wgh

25 See the results of the query at https://labur.eus/56pa8sqv for an entry. The inflected verb forms in the LiLaMorph Wikibase collection have links pointing to *Index Thomisticus* corpus tokens.

26 See the results of the query at https://tinyurl.com/2byrfjn3

References

Bilbao Telletxea, G. (2020). *Oinarrizko hiztegia latina-euskara* (1. ed.). Larousse Editorial; Euskal Herriko Unibertsitateko Argitalpen Zerbitzua.

Lindemann, D. (2025). Ontolex-Lemon in Wikidata and other Wikibase instances. In K. Gkirtzou, I. Žitnik, J. Gracia, D. Gromann, M. Pia di Buono, J. Monti & M. Ionov (Eds.), *Proceedings of the 5th Conference on Language, Data and Knowledge: Workshops* (pp. 287–297). Unior Press. https://doi.org/10.5281/zenodo.15471514

Lindemann, D., Ahmadi, S., Khan, A. F., Mambrini, F., Iurescia, F. & Passarotti, M. C. (2023). When OntoLex Meets Wikibase: Remodeling Use Cases. *CEUR Workshop Proceedings*, 2773. https://doi.org/10.5281/zenodo.18098511

Lindemann, D. & Alonso, M. (2024). Linking Historical Corpus Data and Annotations using Wikibase. In K. Š. Despot, A. Ostroški Anić & I. Brač (Eds.), *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress* (pp. 743–748). Institute for the Croatian Language. https://doi.org/10.5281/zenodo.13932190

Lindemann, D., Pellegrini, M., Mambrini, F. & Passarotti, M. (2025). Wikidata and LiLa for Latin: Enabling Interoperability and Access to Inflected Forms and Corpus Attestations. *Journal of Open Humanities Data*, 11(1). https://doi.org/10.5334/johd.464

Mambrini, F., Litta, E., Passarotti, M. & Ruffolo, P. (2021). Linking the Lewis & Short Dictionary to the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin. In E. Fersini, M. C. Passarotti & V. Patti (Eds.), *Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-It 2021)* (pp. 214–220). https://doi.org/10.5281/zenodo.5773783

McCrae, J., Bosque-Gil, J., Gracia, J., Buitelaar, P. & Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek & V. Baisa (Eds.), *Electronic lexicography in the 21st century: Lexicography from scratch. Proceedings of eLex 2017* (pp. 587–597). Lexical Computing CZ s.r.o. (http://lexbib.elex.is/entity/Q6744). https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf

Passarotti, M., Iurescia, F. & Ruffolo, P. (2025). Harmonizing Divergent Lemmatization and Part-of-Speech Tagging Practices for Latin Participles through the LiLa Knowledge Base. In S. Peng & I. Rehbein (Eds.), *Proceedings of the 19th Linguistic Annotation Workshop (LAW-XIX-2025)* (pp. 103–114). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.law-1.8

Passarotti, M., Mambrini, F., Franzini, G., Cecchini, F. M., Litta, E., Moretti, G., Ruffolo, P., & Sprugnoli, R. (2020). Interlinking through Lemmas. The Lexical Collection of the LiLa Knowledge Base of Linguistic Resources for Latin. *Studi e Saggi Linguistici*, 58(1). https://doi.org/10.4454/ssl.v58i1.277

Passarotti, M. C. & Mambrini, F. (2022). Linking Latin: Interoperable Lexical Resources in the LiLa Project. In E. Biagetti, C. Zanchi & S. Luraghi (Eds.), *Building new resources for historical linguistics* (pp. 103–124). Pavia University Press. https://zenodo.org/records/5994271

Contact information

David Lindemann
EHU University of the Basque Country
david.lindemann@ehu.eus

Acknowledgements

The author would like to warmly thank the author of the Latin-Basque Dictionary, Gidor Bilbao, for providing its content for the experiments described here. Many thanks as well to Marco Passarotti, Paolo Ruffolo, Giovanni Moretti, and Francesco Mambrini from the LiLa

project, for kindly processing the Latin lemma list extracted by us. This work has been supported by Diachronic Linguistics, Typology, and the History of Basque (DLTB) research group at EHU, under grant agreement IT1877-26 (Basque Government).

Peter Meyer

AN AI-DRIVEN NATURAL LANGUAGE INTERFACE FOR GRAPH DATABASE QUERIES IN A LEXICOGRAPHICAL RESOURCE

Abstract Complex digital lexicographical resources often feature rich data structures that are difficult for non-specialist users to query. This paper presents a pilot study on an LLM-driven natural-language interface for the *Lehnwortportal Deutsch*, a graph-based database of German loanwords in other languages. Rather than translating user questions directly into the database query language, the proposed architecture maps natural-language requests to the JSON configuration format already used by the portal's visual query builder. This intermediate representation makes it possible for users to inspect and modify query builder configurations generated from their questions. We discuss the advantages of this approach and present a pilot study that evaluated it on a set of 310 German questions that, due to lack of real user questions, had to be AI-generated using several different strategies to cover a large range of usage scenarios as well as available querying options. Using one-shot prompting, a state-of-the-art commercial-grade LLM produced valid JSON in almost all cases; 260 outputs were rated perfect and 23 acceptable by human assessment under a low reasoning budget. A high-budget rerun improved most initially bad cases, though with impractical latency for real-time use. The paper also discusses methodological challenges in generating and evaluating synthetic query sets, handling unsupported user requests, and extending the approach through larger datasets, comparing multiple language models, varying output parameters, and using fine-tuning as an alternative strategy.

Keywords Loanword lexicography; semantic parsing; natural-language interface; graph database; text-to-JSON; large language models; query generation

1 Introduction

Lexicographical resources have become increasingly complex digital infrastructures, a development that creates new possibilities for research and exploration, but also produces a familiar accessibility problem: the richer the underlying data structures become, the more difficult it is for non-specialist users to formulate queries against it.

The *Lehnwortportal Deutsch* is a case in point. It is a freely accessible database on lexical borrowings from German into other languages. It provides a uniform interface to digital and digitised dictionary resources currently covering 16 recipient languages (Leibniz-Institut für Deutsche Sprache, 2026; Meyer, 2025). Building on the lexicographic data contained in these resources, a cross-resource graph database has been compiled, partly through manual curation. Its core entities (node types) represent words with their associated grammatical and linguistic properties, sense definitions, and dictionary entries. Words can be linked by relations, modelled as edges in the graph, such as borrowing and derivation.

The database's user interface offers a visual query builder that gives access to this graph structure and enables end users to express complex questions like "which German nouns have changed gender in the process of being borrowed into French or Portuguese?" (cf. Figure 1). Users can assemble such queries by adding words to the builder configuration, defining their desired properties and specifying relationships between them through simple mouse clicks. Properties can be nested and combined using Boolean operators. The browser serializes the query builder configurations to JSON (cf. Figure 2), which is translated into the graph database (Neo4J) query language, Cypher, on the server side.

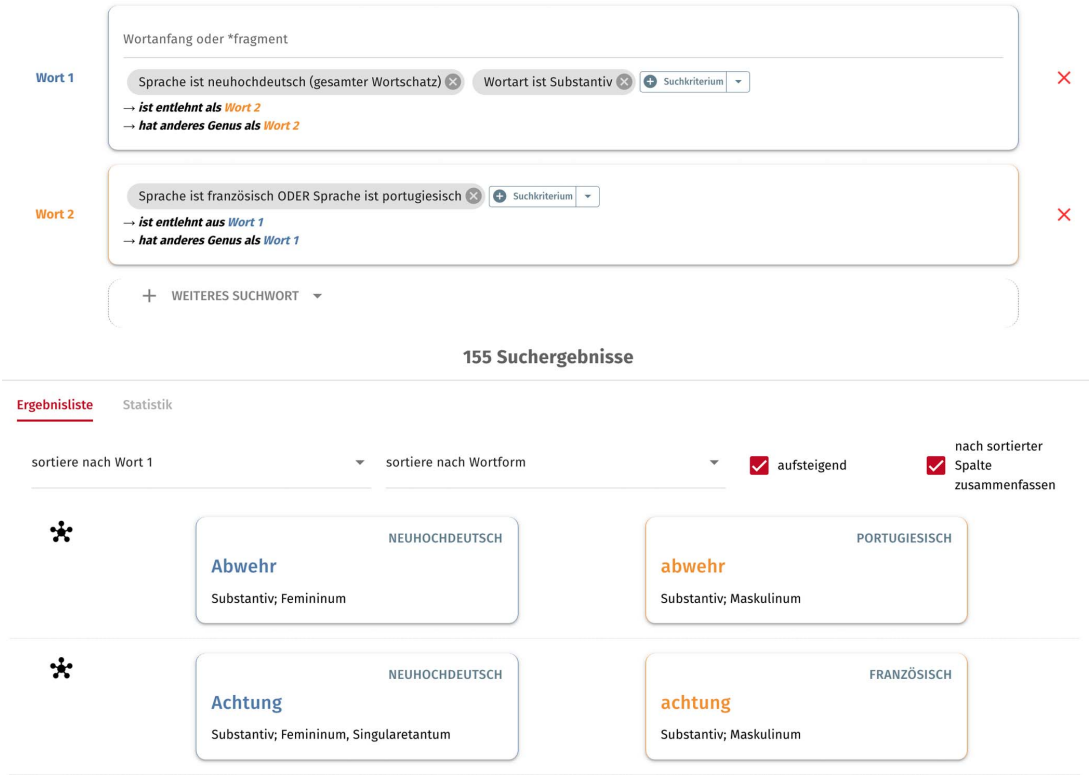


FIGURE 1 Screenshot of a query builder configuration corresponding to the natural language question "which German nouns have changed gender in the process of being borrowed into French or Portuguese?" in the *Lehnwortportal Deutsch* (https://lehnwortportal.ids-mannheim.de/suche)

```
{
  "type": "wordquery",
  "words": [
    { "uuid": "w1", "props": [
      { "name": "dia", "relation": "isdia", "args": ["neuhochdeutsch (gesamter Wortschatz)"] },
      { "name": "pos", "relation": "ispos", "args": ["Substantiv"] }
    ]},
    { "uuid": "w2", "props": [
      { "name": "BOOL", "relation": "or", "args": [
        { "name": "dia", "relation": "isdia", "args": ["französisch"] },
        { "name": "dia", "relation": "isdia", "args": ["portugiesisch"] }
      ]}
    ]}
  ],
  "rels": [
    { "from": "w2", "to": "w1", "name": "borrowedfrom" },
    { "from": "w1", "to": "w2", "name": "differentgen" }
  ]
}
```

FIGURE 2 Slightly simplified JSON representation of the query builder configuration in Figure 1

The resulting output is a list of n-tuples of words (lexical units with lexical indications) that satisfy the graph pattern thus defined, such as pairs of German nouns and their French/Portuguese loanword counterparts with different gender. The list can also be represented as simple bar charts that show how the words at each position in the tuple are distributed according to language, part of speech, gender (in the case of nouns), and first year of attestation.

However, the same expressive power that makes the query builder useful also makes it difficult to use. With its large number of options and possibilities it remains a substantial usability barrier for potential users, including students, teachers, or interested laypersons, and typically requires a non-trivial period of familiarisation despite extensive help materials. The main challenge for users is to transform their question into the formal framework of a graph-based conceptual model.

To get a feeling for the complexities involved in formulating queries, it suffices to have a look at a sample question of the study described below, viz. "German words that have been borrowed into both Polish and Czech, but which only occur as verbs in Polish". The builder configuration proposed by an LLM for this question can be mimicked verbally roughly as follows: "search for word triples with words w1, w2, w3, such that w1 has language (i.e. diasystematic variety) 'New High German' and it is not the case that there exists any word w such that w has both language 'Czech' and part of speech 'verb'; w2 has language 'Polish' and part of speech 'verb'; w3 has language 'Czech' and does not have part of speech 'verb'; w2 is borrowed from w1; w3 is borrowed from w1". This builder configuration is a bit redundant since it is not necessary to restrict w3's part of speech to non-verbs. Constructing this query presupposes a good understanding of the builder's options and input system and the ability to reformulate the original question in the logically structured way described above, using a negated existence statement. Even without the redundant bit, the actual manual input procedure to configure the query builder manually requires a sequence of 60 mouse clicks.

This paper proposes an LLM-driven natural-language interface to the *Lehnwortportal* database. In the following sections, we first provide a concise overview of our basic approach, which is semantic parsing into the query builder configuration language via one-shot prompting, and discuss the central problems to be solved in the testing and evaluation stages. We then present a small-scale pilot study that offers preliminary quantitative and qualitative insights into the viability of our approach. Finally, methodological considerations regarding further, more comprehensive tests and alternative strategies such as fine-tuning are put forward.

2 Natural-language interfaces and semantic parsing

In computational linguistics, the general approach of mapping natural-language utterances to executable formal representations such as logical forms, SQL, SPARQL, Cypher, or other domain-specific query languages is commonly referred to as semantic parsing.

Natural-language interfaces to databases have a long history, but recent progress with generative LLMs has renewed interest in them. In (Koutrika, 2024) current systems are framed as part of a longer "data access odyssey" in which database access is progressively shifted from formal query languages toward more accessible interaction modes. Recent work in the well-researched field of text-to-SQL (cf. Shi et al., 2025) likewise highlights the role of LLMs in lowering barriers to database access, while also emphasising persistent challenges such as ambiguity, underspecification, schema linking, evaluation, and robustness (Liu et al., 2024).

Graph databases introduce further complications. Text-to-Cypher approaches (Ozsoy et al., 2024) attempt to translate natural language directly into Cypher, the query language used by Neo4j and related graph-database systems. Similarly, experiments with Text-to-SPARQL parsing have been used in knowledge-graph question answering, cf. (Perevalov and Both, 2025).

The approach discussed in this paper differs from such direct text-to-Cypher and text-to-SPARQL approaches in one important respect: The target representation is not some

general-purpose database query language but the domain-specific JSON configuration language already used by the *Lehnwortportal* query builder that is subsequently translated into such a query language, viz. Cypher, by the server. The natural-language interface therefore does not bypass the formal builder interface; it provides an additional entry point into it.

2.1 Query builder configurations as an intermediate representation

This intermediate-representation strategy is central to the project. Direct translation to Cypher would require the model to learn implementation-specific details of the graph database that ordinary users never need and are not even supposed to see. By targeting the query-builder JSON instead, the model operates at the same level of abstraction as the visual interface. The JSON data schema thus functions as an easily interpretable, well-constrained domain-specific formal encoding of possible *Lehnwortportal* queries that sharply restricts the otherwise vast space of possible Cypher queries.

The approach also enables a smooth transition between natural-language and formal querying. Since query builder configurations correspond one-to-one to JSON objects, the JSON created by the LLM simply instantiates a state of the visual query builder that ideally represents the user's question. Users may then inspect, modify, simplify, or extend the generated query in the builder. This has obvious didactic value: users can learn how their informal question corresponds to formal query-builder operations. In this respect, the natural-language interface may reduce the initial usability barrier while still preserving the query builder as an explicit, inspectable formal tool.

The same architecture can support feedback for ill-posed questions. Some user requests will be contradictory, underspecified, unsupported by the database, based on some sort of confusion, etc. In such cases, the system should return a structured response indicating whether the request is only translatable with certain additional assumptions or not supported at all. In suitable cases, it may offer a simplified query that captures the supported part of the request while explicitly specifying which aspects of the query had to be discarded.

A general-purpose LLM can handle aspects of requests that are not available as options in the query builder and whose expression requires additional skills. Examples include generating appropriate regular expressions, e.g. for morphological or repetition-pattern searches, or selecting the relevant disjunction or range of available options when users ask e.g. about languages neighbouring German, or the Middle Ages as a borrowing period.

2.2 Generating user queries

A systematic evaluation of the proposed interface requires a test set of natural-language user queries paired with expected outputs. In a mature natural-language interface, such a test set might be derived from server logs or user studies. In the present case, this is not possible. Almost no material on actual user queries with the *Lehnwortportal* query builder is currently available as the builder has –so far– been used by only a small number of people. Even where queries exist, they do not form a sufficiently large or representative corpus in any sense.

This lack of 'naturalistic' data is of course part of the problem the LLM interface is trying to solve. If the query builder is difficult to use, then query logs from the query builder cannot be assumed to represent the full range of information needs that users would express if a natural-language interface existed. A corpus based only on existing query-builder usage would

therefore be biased toward expert users, known use cases, and queries that the existing interface made sufficiently easy to formulate.

Testing the interface therefore requires generated data. However, query generation must be methodologically controlled. If the researcher simply invents a small number of plausible queries, the evaluation risks being subjective and incomplete. If an LLM is asked to generate both user queries and gold-standard JSON without constraints, the evaluation may become circular. If only queries expressible in the JSON language are generated, the test set will ignore important aspects of real user interaction: users frequently ask for things that are vague, unsupported, or based on misconceptions.

The methodology proposed here therefore distinguishes two complementary tracks. One track begins with imagined user needs and deliberately hides the formal query schema from the generating LLM, so that the resulting requests may fall inside or outside the supported space of requests. Here, we further distinguish between two types of imagined users with different levels of prior knowledge about both the *Lehnwortportal* and linguistics. The other track starts from the JSON query builder representations and systematically explores the space of formal possibilities.

3 Pilot study: Generating and executing natural-language queries

This section outlines the methodology and results of a small-scale pilot study on evaluating the text-to-JSON approach outlined above. The study has the very modest aim of demonstrating the feasibility of the approach in principle and identifying methodological challenges. It is meant as a precursor to a larger project that further develops the methodology described here and extends it to larger datasets and a larger number of language models as well as prompting and fine-tuning techniques. The study will be presented as a sequence of six steps. For each step, we will discuss its results and possible strategies to overcome observed shortcomings.

3.1 Generating natural-language questions

In a first step, a set of 310 German natural language questions for the *Lehnwortportal* was generated using the Claude Sonnet 4.6 model by Anthropic, with automatic adaptation of the amount of reasoning used for each prompt (“Adaptive Thinking”). Three different strategies were used to generate questions, corresponding to three different prompts to generate the questions:

- The LLM-based conversational AI system (‘system’ for short) was asked to invent 50 natural-language questions laypersons might pose, assuming they only know that the *Lehnwortportal* is some kind of online resource about German loanwords in other languages and have no deeper knowledge about linguistics in general or about loanwords or Germanisms in particular. For half of the questions, the model should assume a ‘confused’ user who e.g. believes the database is about loanwords borrowed from other languages into German, or loanwords in general, or something else that has to do with language; or whose questions are vague, contradictory, too broad, underspecified, requiring explanation rather than retrieval, or are based on some kind of misunderstanding (or at least terminological confusion) about borrowing, etymology, or language in general. The system turned out to be quite imaginative, producing e.g. a range of meta-questions about the database (such as on the number of words available or on the availability of lexicographical indications), but showed subpar performance in emulating ‘confused’ users, mixing

legitimate information requests ("If a word looks the same in German and English, is it automatically a loanword?") with almost nonsensical questions ("Is 'jardín de niños' a Germanism in Spanish because it means 'kindergarten'?").

- The system was then asked to invent 100 questions that would be typical for various kinds of more informed persons who have rudimentary knowledge (specified explicitly in the prompt) about the graph-based way the *Lehnwortportal* database is organized, the structure and capabilities of the query builder, and what kinds of query output (word n-tuples; bar charts) it offers. The system was asked to consider the possibility that questions might reflect misconceptions about the purpose, functionality or data structures of the database and was given a list of different kinds of users that might invoke the interface, such as historical linguists, undergraduate linguistics students, lexicographers, journalists, etc. A range of topics to be covered in the questions was also given, including e.g. history, culture, semantic change, geography, chronology, language contact, productivity, and comparison across languages. The prompt also named some challenging features of natural-language questions, such as difficult-to-understand specification of the direction or the chain of derivation and borrowing, hidden negation, scope ambiguity, multiple possible source languages, vague comparative terms, contradictory constraints, requests that mix database querying with explanation, etc., and explicitly requested the system to use such features in 30 of the invented questions. A preliminary qualitative assessment of the quality and diversity of the output showed that the system was a bit biased towards asking about chain-like configuration patterns like "Which German nouns ending in *-ung* were borrowed into Polish before 1850 and were used to form adjectives there?" since such a pattern was the showcase example of what can be achieved with the *Lehnwortportal* query builder in the prompt.

- In a third prompt, the system was given a complete and precise textual specification (approx. 2,500 words) of the structure (schema) of the JSON query builder representation and the meaning of its various keys and values from a user's point of view. It was asked to generate, on the basis of the formal possibilities of the schema, a basic set of 40 natural-language questions that can be formally translated into the specified JSON format, not paying attention to whether the question is in any sense plausible, instead just taking arbitrary combinations of possible word features and relations between words. The system was asked to vary complexity (from queries about single words up to complex configurations of multiple words), and to produce a mixture of expert, semi-expert, and layperson formulations. Each of these 40 questions should then be modified in three different ways: (i) by paraphrasing it; (ii) by turning it into a 'near-miss' question, replacing a part of the original question by something not formally queryable or adding an unsupported part to the query; (iii) by turning it into an 'adversarial' variant that uses one of the challenging linguistic features mentioned above to make translation to JSON more difficult. The system was subjectively good at constructing questions and paraphrasing them, but struggled considerably with the 'adversarial' reformulation, which often led to highly implausible questions in which a perfectly reasonable request for information was accompanied by a rather odd, somewhat out-of-place follow-up question on a meta-level, like in "German nouns that gave rise to Polish verbs – where the German word is the source and the Polish verb is the result; but could the German word itself also have been a verb?". This seemed to be a consequence of insufficient prompt engineering and might also reflect idiosyncrasies of the model used.

Overall, the set of 310 questions displayed a remarkable diversity of information needs. Subjectively, there was a predilection for certain patterns, however; for example, the English word *kindergarten* was clearly overrepresented in the question set, possibly because it appears as a typical example in a lot of introductory texts on Germanisms in the training material

of the LLM. It has become clear, though, that in a larger study, the generation of potential questions should be spread across a larger number of prompts, each of which is tailored to a specific group of users and clearly defines the usage context. Experimentation should also be carried out with the number of example questions to be included in the prompt, for instance to illustrate potential interests or misconceptions of users.

3.2 Semantic parsing into query builder JSON

In the crucial second step the 310 LLM-generated questions were 'translated' into the query builder JSON representation used by the *Lehnwortportal*, to the extent that this was at all possible. For each question created in the previous step, a separate one-shot prompt was constructed by concatenating an invariant instructional part of about 2,850 words (which contains the complete JSON structure description already mentioned above and one nontrivial concrete example of a text/JSON pair) and the natural language question itself. The language model used for the task was 'gpt-5.5-2026-04-23', invoked for each prompt in an independent and stateless text-generation request via OpenAI's Responses API with the 'reasoning effort' option set to 'low' and no web access allowed.

The prompt explicitly considered cases in which it is not possible to "translate" the human language question into a JSON object. Typical reasons might be that (a) the requested properties or relations are not available, i.e. they are not part of the JSON schema/specification; (b) there are contradictions in the question or the question is ill-posed in some way; (c) the question is off-topic (does not concern the database data or is not interpretable as a request for database information) or (d) incomprehensible; or (e) the requested information cannot be represented as a list of n-tuples of words or through the kind of visual statistics the *Lehnwortportal* offers. In such cases, the central parts of the JSON object were required to be empty, but a short plaintext explanation (max. 20-30 words) should be provided in German as value of a special 'userFeedback' key in the object. If there were issues with the question but it was still possible to 'translate' the query at least partially (by removing certain parts or by including certain extra assumptions not explicitly formulated in the question) then the JSON query object was expected to reflect this partial translation: The issues found should be commented on in the 'userFeedback' value, and the problematic chunks of the user's question should be set as the value of a separate 'unsupportedParts' key, as an array of strings. In addition, a special 'classification' key was added to the JSON object whose value was one of the following strings, depending on whether there were issues with translating: FULLY_TRANSLATABLE, TRANSLATABLE_WITH_ASSUMPTIONS, PARTIALLY_TRANSLATABLE, NOT_TRANSLATABLE, NOT_A_DATABASE_QUERY; where NOT_A_DATABASE_QUERY should be chosen in cases (c), (d) and (e) above.

The average time for invoking the API and obtaining the response was 8.9 seconds, which would still be a little bit too long for an acceptable user experience, although latency can reasonably be expected to decrease with the rapid progress in LLM development. On average, the API produced 208 output tokens, of which 162 tokens were internal reasoning tokens. Although the prompt requested a pure serialized JSON object as a response, neither OpenAI's 'Structured Output' API option of submitting an explicitly formalized JSON schema alongside the natural-language prompts nor the 'JSON mode' setting which enforces well-formed JSON as sole output were used. Still, all 310 responses consisted exclusively of well-formed JSON, and all but one were confirmed to be valid with respect to a formal JSON schema that was developed from the informal JSON specification in the prompt. On the evaluation of the results returned, see below, step 5.

In the following two steps, information was gathered to provide additional guidance for the final human assessment of the quality of the semantic parsing.

3.3 Executing and verbalizing the query builder configurations

As a third step, the *Lehnwortportal* query builder was programmatically invoked with the semantic parsing results, i.e. with the JSON query objects, in all cases where such an object could be generated from the natural language question in the first place. The search results or error messages based on these query configurations were recorded and a German-language verbalization of the queries (which is used in the *Lehnwortportal* query builder interface and also in the context of its bookmarking functionality) was retrieved.

3.4 LLM comparison of original question and query builder verbalization

In a fourth step, OpenAI's Responses API was used again to compare the *Lehnwortportal* verbalizations with the original question as generated in step 1. To this end, a short invariant prompt text was concatenated with each original question from step 1 and the *Lehnwortportal* verbalization of its corresponding query builder configuration from step 3; the resulting text was submitted with the same API request settings as in step 2. In the prompts, the system was asked to judge the semantic similarity of the original question to the query builder verbalization as PERFECT, ACCEPTABLE, PROBLEMATIC or INSUFFICIENT, and to give a short German-language justification for all cases except the PERFECT ones. Importantly, this classification cannot be used and was not used as independent evidence for evaluation, but as a helpful diagnosis tool for human evaluation in borderline cases, especially in comparison to the 'userFeedback' comments generated in step 2.

3.5 Evaluation of the semantic parsing quality

In a fifth step all 310 cases were evaluated by the author with respect to the quality of the LLM output in step 2 as PERFECT, ACCEPTABLE (i.e. with very minor issues) or BAD. A fourth category, PROMPTING_ISSUE, had to be introduced to cover cases where an unacceptable parsing result turned out to be due to imprecise or lacking information in the prompt. For example, the relation of 'derivation' between words was not defined sufficiently, such that a question like "give me Greek words for which there is a Latin derivative" were constructed using the derivation relation instead of the borrowing relation.

For the human assessment, all information available was presented to the evaluator, i.e. the original question, the type of question (i.e. by which prompt it was generated and to which subclass, as defined in that prompt, it belongs), the 'userFeedback', 'unsupportedParts', and 'classification' values of step 2; and in all cases where a formal query was generated, the *Lehnwortportal* verbalization of the query, the number of hits in the database, one sample hit where available, the LLM comparison of original question and query builder verbalization (classification and justification). It goes without saying that this approach potentially induces various biases. In a rigorously controlled experimental setting, the evaluator(s) must not be presented with anything else than the original question, the 'userFeedback', and the query builder configuration. For the pilot study, the additional information served to speed up evaluation, making the arduous and intellectually demanding task of trying to find flaws in the constructed queries easier.

Out of the 310 questions, 260 parsing responses were rated as PERFECT, 23 as ACCEPTABLE, 9 as bad due to a PROMPTING_ISSUE, and 18 were rated BAD. The distinction between PERFECT and ACCEPTABLE in particular turned out to be difficult, as many questions left plenty of room for interpretation. The example question quoted in the introduction, "German words that have been borrowed into both Polish and Czech, but which only occur as verbs in Polish", may serve as an illustration. The LLM originally proposed a query builder configuration that looks for word triples consisting of a German etymon as well as a Polish verb and a Czech non-verb borrowed from the etymon. Upon close inspection, this is not strictly speaking a correct solution (and was therefore rated as BAD) since a word of origin might have led to multiple borrowings in another language, a phenomenon not uncommon in the *Lehnwortportal* database; so the existence of a non-verb loanword in Czech does not preclude the existence of a Czech verb borrowed from the same etymon. Upon repeating the semantic parsing with the 'reasoning effort' parameter set to 'high' (see step 6 below), the system indeed came up with the formally correct, if a bit redundant, solution mentioned in the introduction. Since the redundancy does not change the result set, this was rated as PERFECT. Even then, questions as to correct interpretation and handling remained – if a German word "occurs as a verb" in Polish, couldn't this also mean that some Polish Germanism has formed a derivative verb? And shouldn't users be warned that the non-occurrence of a phenomenon in the database does not imply its non-existence?

Subjectively, the user feedback generated in step 2 in the case of ill-posed or problematic questions was excellent. To give an example, for the question "Are 'Haus' and 'house' related because they sound so similar?" the user was informed that "Phonetic similarity and a causal explanation for it cannot be retrieved; the search is limited to a direct or indirect lexical relationship", the problematic part of the question was correctly identified as "because they sound so similar", and instead of just refusing to handle this question a search for a generic relationship between German *Haus* and English *house* was constructed.

More generally, the handling of questions part of which did not correspond to available query builder options worked consistently well. To give some examples, the system translated the concept of "non-European language" by a plausible disjunction of available languages in the database, rendered the concept of "a word with less than five letters" through the appropriate regular expression, interpreted morphological conversion (zero derivation) as derivation between two homographic words, correctly rendered the temporal qualification "during the two World Wars" in terms of years ("[since 1914 and until 1918] or [since 1939 and until 1945]"). The system was able to 'invent' strategies to circumvent restrictions of the query system. Thus, for the question "How many Germanisms in the database have more than one possible grammatical gender in the target language?" the user feedback was: "Multiple genders cannot be queried directly; they are approximated using Germanisms in the same language that are homographs but have different genders. The number is indicated as the result list size."

3.6 Redoing the experiment with problematic cases

Finally, the procedure of steps 2 to 5 was repeated for all 18 questions where the semantic parsing result of step 2 was classified as BAD in step 5. As said above, in one of these cases the JSON was well-formed, but not valid, and thus could not be interpreted by the *Lehnwortportal* due to an undefined value. Nine questions were misinterpreted by the system, e.g. "Which German prefixes have been particularly productive in the formation of loanwords in other languages?" was understood as a search for German prefixes borrowed into other languages. In the remaining eight cases, the system failed to make proper use of the query builder options. For example, the question "Which German words have been recorded as loanwords in

more than five European languages at the same time?" was rejected as NOT_TRANSLATABLE with the user feedback "The query requires counting across more than five languages as well as concurrency; such aggregated conditions cannot be represented." In the repetition, the system came up with a close approximation by looking for borrowings of a German word into six other languages, using 15 different-language-relations for the six target languages and correctly commenting on its solution as follows: "It is not possible to display simultaneous attestations within the same time period; 'European languages' is interpreted, for all intents and purposes, as the main European target languages available for selection."

In the repetition step, the 'reasoning effort' parameter of OpenAI's large language model was set to 'high', which led to an average server round trip latency of almost 82 seconds per question with an average number of 3400 internal reasoning tokens. Since such latencies would of course be useless for a real-world user interface, this part of the experiment was conducted mainly to see whether the system was, at least in principle, able to improve on bad output. In the repetition, the handling of 14 questions was classified as PERFECT or ACCEPTABLE. Four problematic cases remained. Of these, two concerned rather subtle issues with the interpretation of slightly ill-posed questions, and it is debatable whether the initial solutions were acceptable after all. The other two concerned the 'multiple borrowings in another language' issue discussed in the previous section that requires the use of negated existence statements. Such constellations were solved successfully in three other cases, which indicates that this is not a fundamental issue with the LLM's capabilities.

4 Outlook: Refining and extending the methodology

The study presented here is meant as a precursor to some larger-scale probes into semantic parsing, with the ultimate goal of adding a natural language interface to the *Lehnwortportal Deutsch*. Even though the results have been very encouraging, many shortcomings and open questions need to be addressed; the evidence is obviously anecdotal due to the small question set and does not allow for valid statistical inference. First and foremost, using real user queries instead of synthetic ones would clearly be the preferred methodology for the evaluation procedure. As long as no sufficient amount of such queries is available, experimentation with more principled and diverse prompting strategies for typical classes of users and information needs would be a necessary next step for the generation of questions. In particular, the capability-driven generation strategy that starts from the JSON schema has been explored in a very naïve fashion in the pilot, prompting an LLM to make up arbitrary valid query builder configurations and turn them into natural language questions. A more controlled way to do this would be to generate such arbitrary JSON configurations programmatically and use an LLM only for the translation process. Finding objective criteria for assessing the quality of synthetic question sets remains an open research topic, however.

The centrepiece of the proposed architecture is semantic parsing through a one-shot prompted LLM-based conversational AI system. Further exploration should use considerably larger question sets, compare the performance of a variety of language models (with a strong focus on open source models) instead of just using one, comparatively evaluate zero- and many-shot prompting strategies, and experiment with the models' prompting parameters such as the chain-of-thought reasoning budget, dynamic adaptation of reasoning effort, and structured output options that enforce output validity against a formal schema.

Operationalizing and controlling the evaluation of the quality of semantic parsing results has turned out to be surprisingly difficult as human evaluation results from a subtle, complex interplay of interpreting both the natural language questions and the exact semantics of query builder options, knowledge about linguistics in general and lexical borrowing in particular,

and the capability of developing "sufficiently approximate" formalizations of textual questions. A more rigorous approach would need multiple trained human evaluators who were not involved in setting up the experiment, measurement of inter-annotator agreement etc.

Another possible avenue is, of course, fine-tuning. The artificially generated and positively evaluated datasets from prompting experiments like the one discussed here could, if sufficiently large, be reused for supervised fine-tuning. In theory, natural-language requests would serve as inputs, and gold JSON objects as outputs. In principle, this could reduce dependence on very long prompts as a fine-tuned model might internalise recurring fragments like "borrowed into X" as a recipient-language constraint expressed in JSON.

Fine-tuning has several potential advantages. It may reduce latency and allow a smaller model to approximate the behaviour of a larger prompted model. It may also improve robustness for common query patterns if sufficient high-quality examples are available.

However, fine-tuning also introduces significant problems. First, it requires a carefully validated training corpus. Synthetic data may encode the biases of the generation procedure, and a model fine-tuned on such data may perform well on generated examples but less well on real user queries. The very idea of a canonical gold JSON representation is highly problematic: Complex query configurations have a large number of semantic equivalents due to possible permutations of search words and multiple equivalent Boolean expressions. More challenging are non-formal equivalences that need some sort of external knowledge to verify. The strategy of just checking against a gold set of correct *Lehnwortportal* search results (execution-based evaluation) does not work since most synthetic queries will yield no results anyway. Second, if the query builder options or their server-side translation to the native database query language changes, the model must be retrained (or carefully prompted with updated information). Third, fine-tuning may reduce transparency. Errors can become harder to diagnose because the model has internalised conventions that are no longer fully visible in the runtime prompt.

5 Conclusion

The proposed natural-language interface to the *Lehnwortportal Deutsch* addresses a practical problem in digital lexicography and elsewhere: how to make a complex graph-database resource accessible to users who cannot be expected to master a powerful visual query builder. In the small pilot study presented here, only some 6% of the 310 synthetically generated questions were not translated in a satisfactory way into a query builder configuration on a low reasoning budget (an additional 3% yielded unacceptable results due to insufficiently precise and exhaustive information in the prompt). Only about 1% of the questions were not treated adequately even on a higher budget (assuming that correct lower budget solutions carry over to the higher budget scenario).

The main methodological challenge is the absence of a corpus of real user queries. Since almost no material on actual query-builder usage is available, possible queries must be generated. We propose a two-track generation methodology. Capability-driven generation starts from the JSON schema and systematically explores the formal query space, generating natural language questions from the schema and varying them by paraphrase or intentional distortion. Desire-driven generation starts from imagined user needs while hiding the schema, thereby producing not only answerable queries but also vague, unsupported, ambiguous, contradictory, or misconception-based requests.

The aim of the approach sketched here is not to replace curated databases with language-model guesses, but to use language models as mediators between ordinary user language and structured resources. If successful, such interfaces can broaden access to specialised lexicographical databases while preserving the precision, inspectability, and reliability of formal querying. The pilot study presented in this paper may serve as a blueprint for similar investigations for other resources.

References

Koutrika, G. (2024). Natural language data interfaces: A data access odyssey. In G. Cormode & M. Shekelyan (Eds.), *27th International Conference on Database Theory (ICDT 2024)* (Leibniz International Proceedings in Informatics, Vol. 290, Article 1, pp. 1:1–1:22). Schloss Dagstuhl – Leibniz-Zentrum für Informatik.

Leibniz-Institut für Deutsche Sprache (2026). *Lehnwortportal Deutsch* (Version 2.0.0) Retrieved July 24, 2026, from https://lehnwortportal.ids-mannheim.de/

Liu, X., Shen, S., Li, B., Ma, P., Jiang, R., Zhang, Y., Fan, J., Li, G., Tang, N. & Luo, Y. (2024). A survey of NL2SQL with large language models: Where are we, and where are we going? *arXiv*. https://doi.org/10.48550/arXiv.2408.05109

Meyer, P. (2025). Germanismen reloaded – das neue Lehnwortportal Deutsch (Aus der Rubrik "Wörter und Wörterbücher"). *Sprachreport*, 41(1), 18–28. https://doi.org/10.14618/sr-1-2025-mey

Ozsoy, M. G., Messallem, L., Besga, J. & Minneci, G. (2024). Text2Cypher: Bridging natural language and graph databases. *arXiv*. https://doi.org/10.48550/arXiv.2412.10064

Perevalov, A. & Both, A. (2025). Text-to-SPARQL goes beyond English: Multilingual question answering over knowledge graphs through human-inspired reasoning. In S. Tramp, M. Gôlo, E. Marx & P. V. do Carmo (Eds.), *Proceedings of the First International TEXT2SPARQL Challenge co-located with Text2KG at ESWC25* (CEUR Workshop Proceedings, Vol. 4094). CEUR-WS.

Shi, L., Tang, Z., Zhang, N., Zhang, X. & Yang, Z. (2025). A survey on employing large language models for text-to-SQL tasks. *ACM Computing Surveys*, 58(2), 1–37. https://doi.org/10.1145/3737873

Wang, Y., Berant, J. & Liang, P. (2015). Building a semantic parser overnight. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (pp. 1332–1342). Association for Computational Linguistics.

Contact information

Peter Meyer
Leibniz-Institut für Deutsche Sprache, Mannheim
meyer@ids-mannheim.de

Krzysztof Nowak & Dorota Mika

POLILEX: MODELLING HETEROGENEOUS LEXICAL RESOURCES

Lessons from the Dictionary of Polish Dialects

Abstract The Institute of the Polish Language (Polish Academy of Sciences) maintains a body of dictionaries and card-file archives assembled over more than a century, whose differences in structure, editorial convention, and descriptive metalanguage have long kept them difficult to search together. This paper reports on PoliLex, an effort to bring such material under a common, machine-queryable representation without erasing the editorial character of each source. We describe an encoding and publication workflow that moves corrected dictionary text through TEI into an OntoLex-Lemon Linked Data representation, and a layered project ontology whose load-bearing decision is to keep a dictionary, as a published text, separate from the language system it records. We illustrate the approach with the Dictionary of Polish Dialects, the first resource processed through the full pipeline, and show how the same encoded source is served both as browsable full text and as a SPARQL-queryable graph. We then examine what such harmonisation is designed to achieve across resources, what it achieves today on a single resource, and where it stops — notably at cross-dictionary sense alignment.

Keywords Linguistic linked open data; OntoLex-Lemon; retro-digitised dictionaries; lexicographic resource integration; Polish dialectology

1 Introduction

The Institute of the Polish Language (Polish Academy of Sciences) holds a collection of lexical resources developed over decades, covering Polish and Latin used in Poland from the Middle Ages to the 21st century in both general and dialectal varieties. These include monumental dictionaries — among them the *Dictionary of Polish Dialects* (SGP), the *Dictionary of Medieval Latin in Poland* (SLSP), the *Dictionary of Old Polish Personal Names* (SSNO), and *Place Names of Poland* (NMP) — as well as card files (collections of handwritten and typewritten slips excerpted from source texts) totalling approximately 6.5 million slips, which constitute the primary source material for lexicographic work. However rich and varied, this very variety presents major challenges to unified storage and retrieval: the resources differ not only in scope and subject matter but also in their internal structure, editorial conventions, and metalanguage.

In this paper we present the workflow, data model, and current state of **PoliLex**, a project aimed at providing integrated access to these resources through harmonised modelling and semantic data publication.[1] We discuss the specific modelling choices involved, drawing on the *Dictionary of Polish Dialects*, which is the first resource for which the full **PoliLex** pipeline has been executed and for which a data release has been produced. Two further resources — the *Electronic Dictionary of 17th- and 18th-century Polish* (S17) and the *Conceptual Dictionary of Old Polish* (SPJS) — are scheduled for release as the project continues.

PoliLex addresses three questions:

(i) how can resources with incompatible structures, conventions, and metalanguages be encoded so they remain faithful to their editorial origins yet become jointly queryable;

[1] The project PoliLex operates in two phases: Phase I under Dariah.lab (2021–2023) and Phase II as statutory research (2024–).

(ii) what data model supports retrieval across resources at the level of forms, definitions, and the geographic localisations attached to them; and

(iii) how can the result be published so it is reusable on the Semantic Web.

The main contributions of the project described here are:

(i) a scans-to-TEI workflow for heterogeneous legacy dictionaries;

(ii) a layered ontology (PoliLex v0.1) extending OntoLex-Lemon and its lexicography module, built on an explicit dictionary-as-artefact vs. language-system distinction;

(iii) an operational single-resource access layer a full-text interface and a SPARQL endpoint).

Underlying all three is an architectural decision to expose the continuously-edited relational database as live virtual RDF via OBDA, so that ongoing editorial work and Linked Open Data publication coexist without forcing lexicographers onto RDF editors.

2 Background: linked lexical data, TEI encoding, and dictionary integration

The *de facto* standard for publishing lexical data as linked open data is OntoLex-Lemon (McCrae et al., 2017), whose lexicography module (lexicog) models a dictionary as a textual artefact, as distinct from the OntoLex core, which models the language system it describes (Bosque-Gil et al., 2019; for a recent overview, Khan et al., 2022). More recently the model has been extended to the frequency and attestation information that is central to documentation-oriented historical dictionaries (OntoLex-FrAC; Chiarcos et al., 2020), with which PoliLex's planned attestation layer is intended to align (§4.2, §7). PoliLex adopts OntoLex-Lemon as its publication target and builds precisely on this artefact/language-system distinction, which we develop in §4.2.

Publishing lexical data as linked data presupposes that the source dictionary has first been made machine-readable. For retro-digitised dictionaries, TEI and its customisation TEI Lex-0 have become a widely adopted target format, designed specifically to make heterogeneously encoded resources interoperable (Romary & Tasovac, 2019), and still under active development (Tasovac et al., 2023, 2026). PoliLex instead retains a per-resource ODD for each dictionary (One Document Does-it-all, see Rahtz & Burnard, 2013) and defers cross-resource reconciliation to the OWL/OBDA layer (§4.2), preserving editorial idiosyncrasies that a shared schema would force together prematurely. Approaches to the structuring step itself range from machine-learning extraction of TEI structure from digitised layout, as in GROBID-Dictionaries (Khemakhem et al., 2017), to the XProc/XSLT transformation of already-corrected text that PoliLex adopts (§4.1).

Encoding individual dictionaries faithfully is, finally, distinct from making many of them queryable together. At European scale, ELEXIS has built an infrastructure to harmonise and interlink national lexicographic resources, both modern and historical, around a common data model and shared standards (Krek et al., 2018). At the level of individual projects, the closest end-to-end analogue to PoliLex is MORDigital, which publishes the *Diccionário da Lingua Portugueza* (1789) through a TEI Lex-0 → OntoLex-Lemon pipeline (Costa et al., 2021; for the usage-information modelling, Almeida et al., 2023), while the LiLa knowledge base interlinks Latin resources around a shared lemma bank and models the Lewis & Short dictionary with lexicog, cleanly separating the lexicographic record from the lexical entry (Mambrini et al., 2021). Closest to PoliLex in material, Abgaz (2020) converts a collection of Bavarian

dialect data — a dictionary together with its questionnaire and paper-slip archive — to OntoLex-Lemon and interlinks the two.

PoliLex differs in the resource scope it must accomodate: a national collection spanning the Middle Ages to the present, semasiological and onomasiological dictionaries, broad diatopic variation, and a card-file base of roughly 6.5 million slips — a combination of diachronic, subject, and structural heterogeneity that comparable projects do not confront together (§7).

3 The Lexical Resources of the Institute of the Polish Language

The materials integrated by PoliLex were developed at the Institute of the Polish Language (IJP PAN) to cover Polish, in both general and dialectal varieties, from the Middle Ages to the present, as well as the Latin used in Poland in the Middle Ages (Table 1).

This paper focuses on the printed and electronic dictionaries in the active PoliLex pilot — SGP as the first deployed resource (see below), S17 and SPJS configured for releasing next, and NMP as the comparator for the discussion of geographic reconciliation in §6.3. The card-file archives constitute the primary source material for the lexicographic work and will be addressed by the planned HTR (handwritten text recognition) pipeline (§7). The IJP PAN text corpora are out of scope for the present paper and reserved for later integration.

TABLE 1 The IJP PAN lexical resources integrated by PoliLex

Resource (abbreviation)	Type	Language/variety	Period
SGP	printed dictionary	dialectal Polish	19th–20th c.
S17	electronic dictionary	Early Modern Polish	17th–18th c.
SPJS	electronic dictionary	Old Polish	up to 15th c.
NMP	printed dictionary	toponyms	modern
SLSP	printed dictionary	Latin	10th–16th c.
SSNO	printed dictionary	antrophonyms	up to 16th c.
SK	printed dictionaries	Polish of the Eastern Borderlands	20th c.
Card-file slip archives	archives	Polish, Latin, dialectal	various
Corpora (x2)	corpus	various	various

The *Dictionary of Polish Dialects* (SGP; Karaś et al., 1979–present) serves as the running example throughout this paper. It is the first resource processed through the complete PoliLex pipeline and the first for which a data release has been produced, comprising approximately 812 entries and 657 senses The dictionary documents dialectal material collected in the nineteenth and twentieth centuries across the major dialect regions of Poland (Silesian, Lesser Polish, Masovian, Greater Polish, and Kashubian), as well as in Polish-speaking enclaves in Lithuania, Belarus, Ukraine, the Czech Republic, Slovakia, Romania, and Hungary. Publication began in the 1960s and is ongoing; by 2025, it comprised more than 6,000 pages in double-column format and ten completed volumes. In 2015, editorial estimates suggested that the published material represented approximately one sixth of the dictionary's projected scope (Kucharzyk, 2015, p. 164).

Among the resources addressed by the project, SGP constitutes a particularly demanding case because of both its scale and the coexistence of multiple editorial conventions. The dictionary

incorporates material documented over more than 150 years and represented using diverse transcription systems and descriptive practices, resulting in a high degree of internal heterogeneity. From the perspective of data modelling and encoding, these layers of variation have direct consequences for digital processing, as changes in entry structure and geographical referencing require different segmentation, encoding, and reconciliation procedures.

4 Methods

4.1 Workflow

The text of the paper-born dictionaries is obtained from scans through OCR and post-correction — for SGP, OCR is carried out with Tesseract (Smith, 2007) using a Polish language model extended with additional phonetic characters, and the output is corrected by annotators through the PoCoTo web tool (Vobl et al., 2014). The hOCR, an OCR output format, is then converted to PAGE XML and progressively structured into TEI-conformant XML through an iterative XProc/XSLT pipeline (Figure 1). The encoding aims to balance preserving editorial idiosyncrasies against the needs of planned cross-resource harmonisation.

The TEI-structuring stage builds the entry tree through successive grouping passes — entries, sub-entries, punctuation handling, and overlap resolution. A subsequent sparse-tree extraction generates two per-volume TSV files (one of entries, one of definitions with parent-sense context) together with an intermediate sparse XML representation, ready for loading into the relational store described in §4.2.

```
<entryFree>
  <form type="lemma">
    <orth norm="ABOWAĆ">ABOWAĆ</orth>
  </form>
  <sense>
    <def>'mocno falować – o przypływie i odpływie morza':</def>
    <cit>
      <quote>ẑis to abu̯uje, včoro to ńö abu̯yvalo </quote>
      <usg type="geo">Jastarnia wej</usg>
      <bibl>
        <title type="abbr">S</title>
        <biblScope unit="volume">I</biblScope>
        <biblScope unit="page">1</biblScope>
      </bibl>
    </cit>
  </sense>
  <pc>.</pc>
  <persName>AN</persName>
</entryFree>
```

ABOWAĆ ‘*mocno falować — o przypływie i odpływie morza*’: Ẑis to abu̯ůje, včoro to ńö abu̯yvalo *Jastarnia wej* *S I 1.* *AN*

FIGURE 1 Entry ABOWAĆ ‘to surge (of the tide)’ in the Dictionary of Polish Dialects with TEI XML encoding

Each resource is documented by its own ODD specifying its annotation choices and their semantics. A per-resource ODD is justified by the heterogeneity of the dictionaries: similar typographic devices encode different features across resources, and even within a single resource the same abbreviation may denote different entities, so a shared schema would force premature reconciliation between editorial practices that the project explicitly seeks to preserve. Each ODD serves three functions in this pipeline: it documents the dictionary’s encoding conventions, drives the schemas used for validation against those conventions, and configures the rendering of the dictionary in the TEI Publisher presentation face described in §5. Adding a new XML resource means writing a new ODD; reconciliation across resources is deferred to the OWL/OBDA layer described in §4.2.

In SGP, entries fall into four types:

(i) <entryFree type="main"> for full entries,

(ii) <entryFree type="hom"> for homonyms,

(iii) <entryFree type="ref"> for cross-references, and

(iv) <entryFree type="empty"> for dialectal entries whose form and meaning are identical to those of the corresponding standard Polish lexeme.

Their microstructure may include a lemma, blocks of phonetic variants with grammatical information and geographic localisation, sense definitions, quotations in phonetic transcription (precisely localised), sub-entries, geographic references, and cross-references.

Two aspects of this microstructure are particularly important because both directly influence the RDF model presented in §4.2 (see Figure 2). The first is the pervasive role of geographic localisation. The element <usg type="geo"> is not confined to the entry level but recurs throughout the structure. It may apply to a block of forms, where several <orth> variants share the same localisation, as well as to individual senses, including nested ones. The second characteristic is the structural nesting of senses. Sense descriptions may be organised hierarchically, with subordinate <sense> elements retaining their own definitions and geographic localisations.

This nesting reflects the organisation of the lexicographic content rather than a semantic relation between senses. Subordinate senses are not interpreted as hyponyms of their parent senses but as lexicographically defined subdivisions within an entry. The relation therefore captures the internal microstructure of the entry rather than a conceptual hierarchy. In the RDF model (§4.2), such nesting is represented by lexicog:subComponent, while recurring geographic localisations form a separate queryable layer. Definitions are represented as skos:definition literals associated with individual senses.

```
<form>
    <orth>baźgro</orth>
    <lbl>||</lbl>
    <orth>baźgze</orth>
    <usg type="geo">
        <settlement>Mszana G</settlement>
        <region>lim</region>
    </usg>
</form>

<sense>
    <def>'oddać na własność – o ziemi przyznanej
      w zamian za poprzednie prawo do użytkowania
      lasu'</def>
    <usg type="geo">
        <settlement>St i N Siołkowice</settlement>
        <region>opol</region>
    </usg>
    <sense>
        <def>'zabronić używania lasu'</def>
        <usg type="geo">
            <settlement>St Siołkowice</settlement>
            <region>opol</region>
        </usg>
    </sense>
</sense>
```

FIGURE 2 Spatial information in TEI-XML-encoded dictionary entries. In the upper snippet, the geographic reference Mszana G[órna] (a village in Limanowa County) applies to both preceding orthographic variants. In the lower snippet, the superordinate sense is attested in St[are] i N[owe] Siołkowice (villages in Opole County), while the subordinate sense is attested in St[are] Siołkowice only.

Particularly challenging aspects of the SGP encoding include non-standard phonetic transcription characters (the dictionary uses a version of the Slavic phonetic alphabet represented in UTF-8), complex bibliographic references, and a geographic localisation system that has changed substantially over time. Geographical references in SGP are not uniform: depending on the source material and editorial practice, a <usg type="geo"> element may refer to a settlement, an administrative unit, a geographical region, or a country. Settlement names are usually given in full (with occasional abbreviations in compound names), whereas higher-level geographical entities are represented by conventional abbreviations. From Volume VI onward, settlement names were replaced by numerical identifiers referring to a separate geographical index (e.g. *Domaniewek łęcz* became *łęcz 3* where *łęcz* is the abbreviation of *Łęczyca County*; Okoniowa, 2001, pp. V–VI). In addition, the 1999 administrative reform altered the boundaries of many units referenced in the dictionary, requiring the reconciliation of geographic references with the current administrative geography. To address this issue, a batch geocoding step maps each abbreviation–settlement pair to coordinates using a public gazetteer service; the resulting coordinates are then merged back into the TEI source. The auto-geocoding procedure is deliberately simple and remains subject to ongoing lexicographic review.

Following automated processing, the structured TEI output is passed to annotators for proofreading and correction. The quotation strings receive particular attention, since they contain non-standard phonetic characters and represent language varieties that automatic correctors do not handle; annotators also correct over-aggressive entry, definition, or attestation splits introduced by the structuring pipeline. Annotators work in Oxygen and validate their output against the ODD-derived schema.

4.2 Data model

The resulting TEI XML files serve as the basis for data extraction: the sparse-tree transformation (§4.1) emits two TSV files (entries and definitions) alongside a sparse XML file. The entries and definitions are imported into an Entity–Attribute–Value (EAV) schema in SQLite, transferred to a PostgreSQL store, and then mapped to RDF through OBDA mappings implemented with Ontop (Bagosi et al., 2014), which read the resulting eav.* tables (entity types Entry/Sense/Form/Example).

The complexity of the data model reflects how heterogeneous the underlying resources are. To address this, we developed a project ontology (PoliLex, v. 0.1) structured in three layers:

(i) a lexicographic layer extending the W3C lexicog module (currently implemented);

(ii) a mapping layer for cross-dictionary alignments via SKOS and owl:sameAs (planned);

(iii) a semantic (lexicon) layer extending OntoLex-Lemon proper (McCrae et al., 2017; also planned).

Resource-specific ontologies (**sgp/**, with **s17/** and **spjs/** planned) import a shared **base/** ontology that in turn extends OntoLex-Lemon and its lexicography module (Bosque-Gil et al., 2019; Khan et al., 2022), so a new resource can be integrated by writing its resource-specific layer.

We distinguish between the dictionary as a textual artefact (lexicog:LexicographicResource) and the language system it describes (ontolex:Lexicon) (Bosque-Gil et al., 2019). Entries are meant to be represented at both the lexicographic level (lexicog:Entry, already in the live

graph) and the lexical level (ontolex:LexicalEntry, planned and linked via the base:linkedTo-Lemon property). Microstructural components — definitions, quotations, usage labels — are declared as subclasses of lexicog:LexicographicComponent in the base ontology, and the live SGP graph types sgp:Sense and sgp:Definition accordingly. Definitions themselves, however, are emitted not as sgp:Definition instances but as skos:definition literals attached to senses — a choice conformant with the W3C lexicog recommendation (Bosque-Gil et al., 2019). This dual representation preserves the specificity of each resource's editorial practice while enabling abstraction across resources (Almeida et al., 2023).

Choosing EAV for the relational stage is itself a response to heterogeneity: a single generic schema (eav.entities, eav.entity_types, eav.attributes, eav.entity_values, eav.entity_relationships) must absorb dictionaries with very different microstructures without per-dictionary table design. The EAV-to-RDF mapping is consolidated in a single canonical OBDA

```
# OBDA: one atomic mapping per triple (SQL SELECT over eav.* → RDF)

mappingId   SGP-Entry-Class
target      :entry/{entity_id} a sgp:Entry .
source      SELECT DISTINCT e.entity_id FROM eav.entities e
            JOIN eav.entity_types et ON e.type_id = et.type_id
            WHERE e.resource_name = 'sgp' AND et.type_name = 'Entry'

# remaining mappings, targets only (mappingId and source SQL omitted)
target      sgp:dictionary a base:LexicalResource .
target      :entry/{entity_id} lexicog:lemma {lemma}^^xsd:string .
target      :sense/{entity_id} skos:definition {definition}^^xsd:string .
target      :entry/{entry_eid} lexicog:subComponent :sense/{sense_eid} .
```

FIGURE 3 Sample mappings from the canonical OBDA file; SQL source clauses omitted except for the entry-class mapping

file, itself generated from the ontology and configuration so that the mapping stays in sync with schema changes (Figure 3).

A separate evidence vocabulary is declared in the base ontology but not yet populated. It keeps slips, source documents and facsimiles distinct from the lexical microstructure, on the principle that a slip is evidence for an entry or sense, not a lexical component inside it: the two are linked by a dedicated relation.[2] Populating this layer from the card-file HTR pipeline is future work (Chiarcos et al., 2020).

URIs follow a predictable scheme under the base namespace http://data.ijppan.pl/:

(i) instances at /dataset/lexres/{resource}/{entry|sense|definition}/{id} — where {id} is the EAV entity_id, the original TEI xml:id being retained as a base:hasEntryId literal;

(ii) vocabularies at /ontology/{base|sgp|s17|spjs}/, and

(iii) reserved per-resource named-graph slots at /graph/{lexres|mappings|ontolex}/{resource} (declared but not yet activated — see §5).

[2] The relevant classes are base:Attestation (with base:Slip ⊑ base:Attestation and base:SourceDocument ⊑ dct:BibliographicResource), linked to lexical components via base:supportsComponent, with IIIF and source-citation properties (base:hasIiifCanvas, base:hasIiifRegion, base:hasSourceDocument, . . .) attached to attestations.

5 Current State and Access

The first batch of data is exposed through two complementary access faces:

(i) a SPARQL/Linked-Open-Data endpoint and

(ii) a TEI Publisher full-text presentation, both fed from the same encoded SGP source.

The entry points are polilex.pl (redirecting to the LOD host data.ijppan.pl) and tei.polilex.pl (redirecting to the TEI Publisher instance); a dedicated browser application is in development (see below). The SPARQL endpoint allows structured queries — for instance, retrieving all orthographic forms or dialectal senses attested in a given region. In the present deployment only the *Dictionary of Polish Dialects* is loaded, so the query pattern already supports retrieval across resources but the results will not appear until we load further dictionaries (§7).

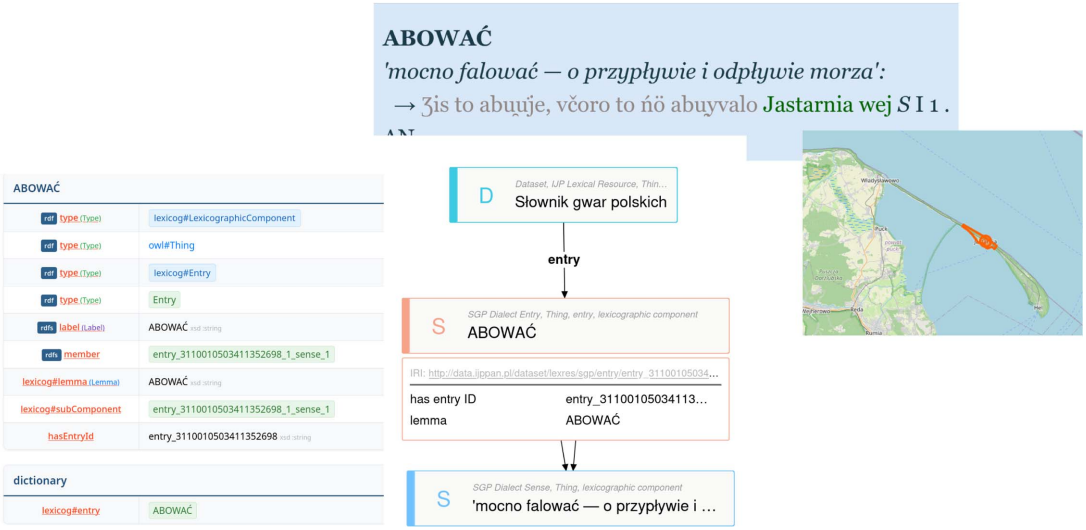


FIGURE 4 One SGP source, two faces: TEI Publisher full text with geolocated attestations (top right) and Trifid-based Graph Explorer view (bottom) of the entry ABOWAĆ 'to surge (of the tide)'

In concrete terms, the deployed graph currently holds 812 SGP entries — a pilot subset of the ca. 32,000 published to date — with 657 senses and 606 skos:definition literals. In total it contains 13,558 triples, all in the default graph (S17 and SPJS are configured but contain no instance data). The endpoint is backed by two stores over the same data:

– an RDF4J triplestore holding the materialised triples (the primary query backend) and

– an ONTOP instance exposing a live virtual view over the PostgreSQL eav.* schema through the canonical OBDA mapping (§4.2), periodically re-materialised from the ONTOP view into the RDF4J store.

A Trifid gateway[3] provides a simple frontend with the usual navigation, query-proxy and pl/en plugins, a lexicog-aware entity renderer and a YASGUI workbench; a query-type router and a Fuseki backup are present but not yet in production.

[3] https://github.com/zazuko/trifid

The full-text face uses TEI Publisher (Figure 4), configured for SGP through the lexica_sgp.odd schema. We chose it to get the text online quickly, not to build a custom presentation layer. Apart from text display, each <usg type="geo"> is plotted from the simple batch auto-geocoding (§4.1), a provisional but already usable solution. The dedicated browser application, built on a React stack (Figure 5), reads the materialised store and offers lemma search with autocomplete, a Polish-alphabet index, paginated results filtered by entry type and letter, and per-entry detail pages backed by a small JSON API; map and geographic visualisation are not yet part of this client.



FIGURE 5 React-based client application: autocomplete search (left) and entry page (right) for the entry ABOWAĆ 'to surge (of the tide)'

6 Discussion

6.1 Editorial fidelity vs cross-resource harmonisation

The tension between resource idiosyncrasies — which need to be preserved for future analysis, enrichment, and citability — and the cross-resource harmonisation we aim at is addressed first by distinguishing the text-centric lexicog representation (dictionary-as-artefact) from the data-centric ontolex representation (language-as-system). The workflow, data shape, and modelling choices aim to preserve both perspectives, reflecting the needs of the community involved in their elaboration and future reuse. They also reflect the heterogeneity of the resources — in genre, coverage, and scope — and the fact that editing proceeds incrementally: the shape of the data should not be pre-defined early on, since in-depth analysis is still needed to understand what is being modelled.

The storage architecture contributes to the same goal. Because the dictionaries are continuously edited and the lexicographic teams maintain their work in relational databases rather than RDF stores, a materialised triplestore would require a re-load on every edit and would force teams to switch to RDF editors. Virtual RDF via OBDA lets the editorial database remain the single source of truth while exposing its content as live linked data.

Two details of the definitions-as-skos:definition choice (§4.2) are worth noting here: language — the only definition-level metadatum currently in scope — may be attached to the literals via language tags; and the unused sgp:Definition class is kept as a forward placeholder (Bosque-Gil et al., 2019), for future definitions that must carry their own provenance,

dating, or cross-resource links. The ontolex side is currently empty: base:linkedToLemon is declared as the link between the two levels but has no instances yet. Populating it requires more than additional OBDA rules — it depends on sense-level alignment across dictionaries, which is different from data ingestion (see §6.2).

6.2 Data harmonisation

Harmonisation already reaches forms, definitions, and the geographic localisations attached to them; cross-dictionary sense alignment does not; it is an open research problem rather than a matter of loading more data. The SPARQL queries may already ask about lexicog:Entry and lexicog:LexicographicComponent rather than about SGP-specific classes. The same query would return results from all three resources without modification, because every entry carries the shared lexicog:Entry typing — though today, with only SGP loaded, it returns SGP results alone. The ontology stubs and OBDA mapping skeletons for S17 and SPJS are configured, with ingestion pending once their curation is finalised.

Cross-resource queries cannot, however, return sense-aligned results, and would not even if all three resources were loaded today, because the dictionaries do not share a common-sense inventory. Cross-dictionary sense alignment is itself an actively studied problem — the ELEXIS Monolingual Word Sense Alignment task and system (NAISC) tackle it directly — but the problem resists a clean solution: even dedicated methods produce matches that are often incomplete and rarely one-to-one, in part because dictionaries divide senses at different levels of detail (McCrae & Buitelaar, 2018; McCrae et al., 2021).

Linking senses across Old Polish (SPJS), 17th–18th-century Polish (S17), and 20th-century dialectal Polish (SGP), and other resources requires us both to represent word forms carefully and to model a shared conceptual space across nearly a millennium of language change and across distinct geographical and social varieties. Two natural entry points exist: S17 and SPJS carry explicit lexicographic cross-references to other dictionaries, and SGP links dialectal forms to their standard-language meanings through empty form-only entries. Systematic sense alignment is therefore planned as separate, research-intensive work.

6.3 Two publication faces

The two deployed publication faces correspond to the dual goal of the project: preserving the lexicographic, text-oriented perspective on the original resource — for human reading, citability, and printing — and exposing a separate data-centred LOD view for integrated querying, discovery, and reconciliation. The split is deliberate: no single front-end we tested reads an encoded dictionary as a text and queries its harmonised content as data equally well.

Geographic information is the layer where this dual ambition is most concretely visible. The relevant resources encode places at different levels of granularity and through different labelling systems — partly because the dictionaries serve different scholarly purposes (dialectal attestation in SGP, toponymic description in NMP), and partly because within a single long-running dictionary the encoding has drifted over decades of compilation, as both the conventions and the administrative geography they referred to changed over the course of compilation. In the text-centred face, the resource-specific abbreviations and labels are preserved exactly as written; in the data-centred face, automated geocoding produces a queryable layer that the editorial team treats as a proposal to review and correct, rather than as the authoritative reading.

7 Conclusion and Future Work

The PoliLex project aims to model lexical data and provide integrated retrieval across resources developed over decades at the IJP PAN, through TEI conversion and OntoLex-Lemon publication. In this paper we have presented the design choices and current state: the distinction between the dictionary as a published artefact and the language system it describes; the per-resource ODDs and ontologies; and the EAV + OBDA architecture that reconciles continuous editorial work with publishing the data as linked data. A first batch — 812 entries and 657 senses of the Dictionary of Polish Dialects — has been deployed through the full pipeline, queryable through a SPARQL endpoint and exposed as dereferenceable URIs.

Work still in progress includes semantic linking between senses across dictionaries — moving from the lexicographic (text-centric) level toward the lexicon (language-system) level, and systematic linking to external resources; a first step has been made by connecting source information to VIAF and Wikidata for author and work identification. The geographic layer will likewise move from simple coordinate geocoding toward Wikidata-anchored reconciliation of historical administrative units, following the OpenRefine workflow already piloted for Place Names of Poland (NMP); we treat this line of work more fully elsewhere. The continuation of this work, together with the integration of additional resources (S17 and SPJS are already configured), will enable richer cross-resource queries and support research on the development of Polish vocabulary across time and space.

Supporting research on Polish lexis across time and space is the project's principal aim. Dispersed across separate editions and formats, the resources have not so far supported an integrated view of the history and variation of Polish lexis. At a fine-grained level, linked data will support a richer characterisation of the lexical and phonetic properties of individual regional varieties. Taken together, the same data will make it possible to ask how and in what domains dialects lexicalise senses that the standard language leaves unlexicalised, how widely such properties are distributed in space and time, and what phonetic, grammatical, and semantic patterns emerge from their combined analysis.

References

Abgaz, Y. (2020). Using OntoLex-Lemon for Representing and Interlinking Lexicographic Collections of Bavarian Dialects. In M. Ionov, J. P. McCrae, C. Chiarcos, T. Declerck, J. Bosque-Gil & J. Gracia (Eds.), *Proceedings of the 7th Workshop on Linked Data in Linguistics (LDL-2020)* (pp. 61–69). European Language Resources Association. https://aclanthology.org/2020.ldl-1.9/

Almeida, B., Costa, R., Salgado, A., Ramos, M., Romary, L., Khan, F., Carvalho, S., Khemakhem, M., Silva, R., & Tasovac, T. (2023). Modelling Usage Information in a Legacy Dictionary: From TEI Lex-0 to OntoLex-Lemon. In Y. Rochat, C. Métrailler & M. Piotrowski (Eds.), *CEUR Workshop Proceedings*, 3602, Paper 1 (pp. 5–21). https://ceur-ws.org/Vol-3602/paper1.pdf

Bagosi, T., Calvanese, D., Hardi, J., Komla-Ebri, S., Lanti, D., Rezk, M., Rodríguez-Muro, M., Slusnys, M., & Xiao, G. (2014). The Ontop Framework for Ontology Based Data Access. In D. Zhao, J. Du, H. Wang, P. Wang, D. Ji, & J. Pan (Eds.), *The Semantic Web and Web Science. Communications in Computer and Information Science*, Vol. 480 (pp. 67–77). Springer. https://doi.org/10.1007/978-3-662-45495-4_6

Bosque-Gil, J., Gracia, J., Montiel-Ponsoda, E., Frontini, F., & Kernerman, I. (2019). *The OntoLex-Lemon Lexicography Module: Final Community Group Report. World Wide Web Consortium*. https://www.w3.org/2019/09/lexicog/

Chiarcos, C., Ionov, M., de Does, J., Depuydt, K., Khan, A. F., Stolk, S., Declerck, T., & McCrae, J. P. (2020). Modelling Frequency and Attestations for OntoLex-Lemon. In I. Kernermann, S. Krek, J. P. McCrae,

J. Gracia, S. Ahmadi & B. Kabashi (Eds.), *Globalex Workshop on Linked Lexicography (LREC 2020)* (pp. 1–9). European Language Resources Association (ELRA). https://doi.org/10.5281/zenodo.3842633

Costa, R., Salgado, A., Khan, A. F., Carvalho, S., Romary, L., Almeida, B., Ramos, M., Khemakhem, M., Silva, R., & Tasovac, T. (2021). MORDigital: The Advent of a New Lexicographical Portuguese Project. In I. Kosem, M. Cukr, M. Jakubíček, J. Kallas, S. Krek, & C. Tiberius (Eds), *Electronic lexicography in the 21st century. Proceedings of the eLex 2021 conference (5–7 July 2021)* (pp. 312–324). Lexical Computing CZ.

Khan, A. F., Chiarcos, C., Declerck, T., Gifu, D., González-Blanco García, E., Gracia, J., Ionov, M., Labropoulou, P., Mambrini, F., McCrae, J. P., Pagé-Perron, E., Passarotti, M., Ros Muñoz, S., & Truică, C.-O. (2022). When linguistics meets web technologies: Recent advances in modelling linguistic linked data. *Semantic Web*, 13(6), 987–1050. https://doi.org/10.3233/SW-222859

Khemakhem, M., Foppiano, L., & Romary, L. (2017). Automatic Extraction of TEI Structures in Digitized Lexical Resources using Conditional Random Fields. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek, & V. Baisa (Eds.), *Electronic Lexicography in the 21st Century: Lexicography from Scratch. Proceedings of eLex 2017* (pp. 598–613). Lexical Computing CZ. https://hal.science/hal-01508868

Krek, S., Kosem, I., McCrae, J. P., Navigli, R., Pedersen, B. S., Tiberius, C., & Wissik, T. (2018). European Lexicographic Infrastructure (ELEXIS). In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (Eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts (EURALEX 2018), Ljubljana, Slovenia* (pp. 881–892). https://doi.org/10.5281/zenodo.1435470

Kucharzyk, R. (2015). "Słownik gwar polskich" PAN: Nowe możliwości, nowe wyzwania. *Rozprawy Komisji Językowej*, 61, 163–172.

Mambrini, F., Litta, E., Passarotti, M., & Ruffolo, P. (2021). Linking the Lewis & Short Dictionary to the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin. In E. Fersini, M. Passarotti & V. Patti (Eds.), *Proceedings of the Eighth Italian Conference on Computational Linguistics (CLiC-it 2021)* (pp. 216–222). CEUR Workshop Proceedings. https://aclanthology.org/2021.clicit-1.34/

McCrae, J. P., Bosque-Gil, J., Gràcia García, J., Buitelaar, P., & Cimiano, P. (2017). The OntoLex-Lemon Model: Development and Applications. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek, & V. Baisa (Eds.), *Electronic lexicography in the 21st century: Lexicography from scratch. Proceedings of eLex 2017* (pp. 587–597). Lexical Computing CZ. https://elex.link/elex2017/wp-content/uploads/2017/09/paper36.pdf

McCrae, J. P., & Buitelaar, P. (2018). Linking datasets using semantic textual similarity. *Cybernetics and Information Technologies*, 18(1), 109–123. https://doi.org/10.2478/cait-2018-0010

McCrae, J. P., Ahmadi, S., Yim, S.-b., & Bajčetić, L. (2021). The ELEXIS system for monolingual sense linking in dictionaries. In I. Kosem, M. Cukr, M. Jakubíček, J. Kallas, S. Krek & C. Tiberius (Eds.), *Electronic lexicography in the 21st century: post-editing lexicography. Proceedings of the eLex 2021 conference* (pp. 542–559). Lexical Computing CZ. https://doi.org/10.5281/zenodo.5772026

Okoniowa, H. (Ed.). (2001). *Słownik gwar polskich* (Vol. 6, fasc. 16). Kraków: Instytut Języka Polskiego PAN.

Rahtz, S., & Burnard, L. (2013). Reviewing the TEI ODD system. In *Proceedings of the 2013 ACM Symposium on Document Engineering (DocEng '13)* (pp. 193–196). Association for Computing Machinery. https://doi.org/10.1145/2494266.2494321

Romary, L., & Tasovac, T. (2019). *TEI Lex-0: A Target Format for TEI-Encoded Dictionaries and Lexical Resources* [Conference presentation]. TEI 2018: TEI as a Global Language (TEI2018), Tokyo, Japan. https://doi.org/10.5281/zenodo.2613594

Smith, R. (2007). An overview of the Tesseract OCR engine. In *Proceedings of the Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)* (pp. 629–633). IEEE Computer Society. https://doi.org/10.1109/ICDAR.2007.4376991

Tasovac, T., Bański, P., Herold, A., Lehečka, B., Romary, L., & Salgado, A. (2023). TEI Lex-0: Recent Developments and New Directions. *TEI-MEC 2023*. https://teimec2023.uni-paderborn.de/contributions/174.html

Tasovac, T., Romary, L., Bański, P., Engler, V., Herold, A., Khan, F., Lehečka, B., Moerth, C., Salgado, A., Schopper, D., & Sramó, K. (2026). *TEI Lex-0: A baseline encoding for lexicographic data (Version 0.9.5)*. *DARIAH Working Group on Lexical Resources*. https://lex-0.org

Vobl, T., Gotscharek, A., Reffle, U., Ringlstetter, C., & Schulz, K. U. (2014). PoCoTo: An open source system for efficient interactive postcorrection of OCRed historical texts. In *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage (DATeCH '14)* (pp. 57–61). Association for Computing Machinery. https://doi.org/10.1145/2595188.2595197

Dictionaries

NMP: Rymut, K., Czopek-Kopciuch, B., Bijak, U., & Swoboda, P. (Eds.). (1996–2024). *Nazwy miejscowe Polski: Historia, pochodzenie, zmiany* [Place Names of Poland: History, Origin, and Changes] (Vols. 1–17). Instytut Języka Polskiego PAN.

SGP: Karaś, M., Reichan, J., Okoniowa, H., Wronicz, J., & Kucharzyk, R. (Eds.). (1979–present). *Słownik gwar polskich* [Dictionary of Polish Dialects] (Vols. 1–11). Instytut Języka Polskiego PAN.

SLSP: Plezia, M., Weyssenhoff-Brożkowa, K., & Rzepiela, M. (Eds.). (1953–2021). *Słownik łaciny średniowiecznej w Polsce* [Dictionary of Medieval Latin in Poland] (Vols. 1–8). Zakład Narodowy im. Ossolińskich – Wydawnictwo Polskiej Akademii Nauk.

SPJS: *Słownik pojęciowy języka staropolskiego* [Conceptual Dictionary of Old Polish]. Retrieved June 1, 2026, from https://spjs.ijppan.pl

SSNO: Taszycki, W. (Ed.). (1965–1983). *Słownik staropolskich nazw osobowych* [Dictionary of Old Polish Personal Names] (Vols. 1–6). Zakład Narodowy im. Ossolińskich – Wydawnictwo Polskiej Akademii Nauk.

S17: *Elektroniczny słownik języka polskiego XVII i XVIII wieku* [Electronic Dictionary of 17th- and 18th-Century Polish]. Retrieved June 1, 2026, from https://sxvii.pl/

Dictionaries of the Eastern Borderlands:

Cechosz-Felczyk, I. (2004). *Słownictwo gwary Oleszkowiec i Hreczan (Greczan) na Podolu*. Wydawnictwo Libron.

Czarnecka, K., Kowalska, D., Paryl, W., Rieger, J., & Rudolf-Ziółkowska, E. (Eds.). (2007). *Słownictwo gwarowe przesiedleńców z Ukrainy: Słownik porównawczy kilku wsi w Tarnopolskiem*. Wydawnictwo Lexis.

Czarnecka, K., Kowalska, D., Paryl, W., & Rudolf-Ziółkowska, E. (Eds.). (2015). *Słownictwo gwarowe przesiedleńców z Ukrainy: Słownik porównawczy kilku wsi we Lwowskiem*. Wydawnictwo Lexis.

Corpora

KorBa: *Elektroniczny korpus tekstów polskich z XVII i XVIII w.* [The Electronic Corpus of 17th- and 18th-Century Polish Texts]. Retrieved June 1, 2026, from https://korba.edu.pl

Korpus Mowy IJP PAN [IJP PAN Speech Corpus]. Retrieved June 1, 2026, from https://korpusmowy.ijppan.pl/

Contact information

Krzysztof Nowak
Institute of Polish Language (Polish Academy of Sciences)
krzysztof.nowak@ijppan.pl

Dorota Mika
Institute of Polish Language (Polish Academy of Sciences)
dorota.mika@ijppan.pl

Acknowledgements

This research was partially supported by the project Digital Research Infrastructure for the Arts and Humanities DARIAH-PL (DARIAH.lab), implemented under Measure 4.2 of the Operational Programme Smart Growth 2014–2020 (Project No. POIR.04.02.00-00-D006/20).

Author contributions

Krzysztof Nowak devised the overall approach to modelling and processing the lexicographic data; designed and implemented the data model, the encoding and publication workflow, the software infrastructure, and the quantitative analysis of the dataset; and wrote the corresponding parts of the manuscript. Dorota Mika documented and inventoried the source resources to inform the encoding and modelling, prepared and validated the lexicographic data, contributed the dictionary examples used throughout the paper, and drafted the corresponding sections of the manuscript. Both authors revised and approved the final version.

Vasyl Starko & Andriy Rysin

LARGE-SCALE SEMIAUTOMATIC DETECTION OF NEW AND UNREGISTERED WORDS IN MODERN UKRAINIAN

Abstract We have created and implemented a workflow for large-scale semiautomatic detection of new and previously unregistered words in Modern Ukrainian. Crucially, it involves a diverse corpus of Ukrainian (2 billion tokens), a large electronic morphological dictionary of Ukrainian, and a specially crafted NLP toolkit. The key methodological innovation lies in the dual use of the dictionary. On the one hand, it serves as a resource for lemmatizing and morphological tagging the corpus; on the other, it is cyclically enriched with new words detected in the corpus. The corpus is dynamic, as it is regularly expanded by adding a variety of texts from the 19th century to the present day. The dictionary, currently containing 444,000 lemmas, is dynamically reused with the help of a tailor-made NLP toolkit as an exclusion source (Janssen, 2009) against the growing corpus. The paper provides a detailed step-by-step description of the procedure. This semiautomatic pipeline has yielded thousands of new and unregistered Ukrainian words, laying the foundation for the systematic, large-scale exploration of the Modern Ukrainian lexicon.

Keywords Ukrainian dictionary; modern Ukrainian; neologism; neologism detection methodology; Ukrainian corpus; unregistered lexis; Ukrainian lemmatizer; Ukrainian POS tagger

1 Introduction

Identifying words in a language is a fundamental task and the first step in its lexicographic description. While there are multiple initiatives to document new words (Baloh et al., 2010; Belei et al., 2022; Karpilovska, 2013; Neliuba, 2012; Neliuba, 2017; Neliuba & Redko, 2022; Turovska & Vasylkova, 2008) and the general lexical stock of Modern Ukrainian (Busel, 2005; SUO; SUM), most of them do not make substantial use of corpus data.

One reason is that, for a long time, Ukrainian lacked a suitably large and diverse corpus for large-scale lexicographic exploration of its dynamics. By now, the situation has changed dramatically, with several large corpora of Ukrainian available for lexicographic use (ukTenTen22; PAWUK; AU). While valuable for exploring current usage and recent trends, these web corpora lack textual variety and chronological depth, both of which are necessary for a full-fledged lexicographical description of a language.

In the age of large electronic textual collections, a lexicographer needs a suitable computational tool to sift through massive amounts of data and automatically suggest candidate items for further inspection. Various corpus managers, such as SketchEngine, can be used for this purpose, but they may not have the functionality needed for the task at hand. In our case, a specialized toolkit was required to establish which words are new or unregistered in Ukrainian, as explained below.

2 Tools and resources

The history of Modern Ukrainian is complicated, to say the least. It has experienced more than a century of forced Russification, oppressive regimes, wars, multiple spelling reforms, and attempts at restandardization (Vakulenko & Karunyk, 2018). A new official orthography

system was adopted as recently as 2019 (UP19) and revised in 2026 (UP26). Together, these factors create an extremely varied tapestry of the language with intricate patterns of variation. The only existing text collection that does some justice to this complex situation is the General Regionally Annotated Corpus of Ukrainian (GRAC) (Shvedova et al., 2017–2026; Shvedova, 2026), which we have used for our research. While the size of the corpus is fairly modest (2 billion tokens), its strength lies in its diversity: over 800,000 texts by some 35,000 authors spanning two centuries, from all regions of Ukraine and the diaspora. A variety of text types are included—from fiction, journalism, and official documents to transcribed speech, ego-documents, and internet communication. That said, the corpus is not balanced, and its coverage across different time periods is uneven, as illustrated in Figure 1 below. The corpus is dominated by 21st-century texts, especially journalism. Fiction, non-fiction, and academic texts are found in decent proportions in every time period. The sizes of decade-long subcorpora decrease as we go further back in time. Despite this, no other corpus of Ukrainian comes even close to this level of variety.

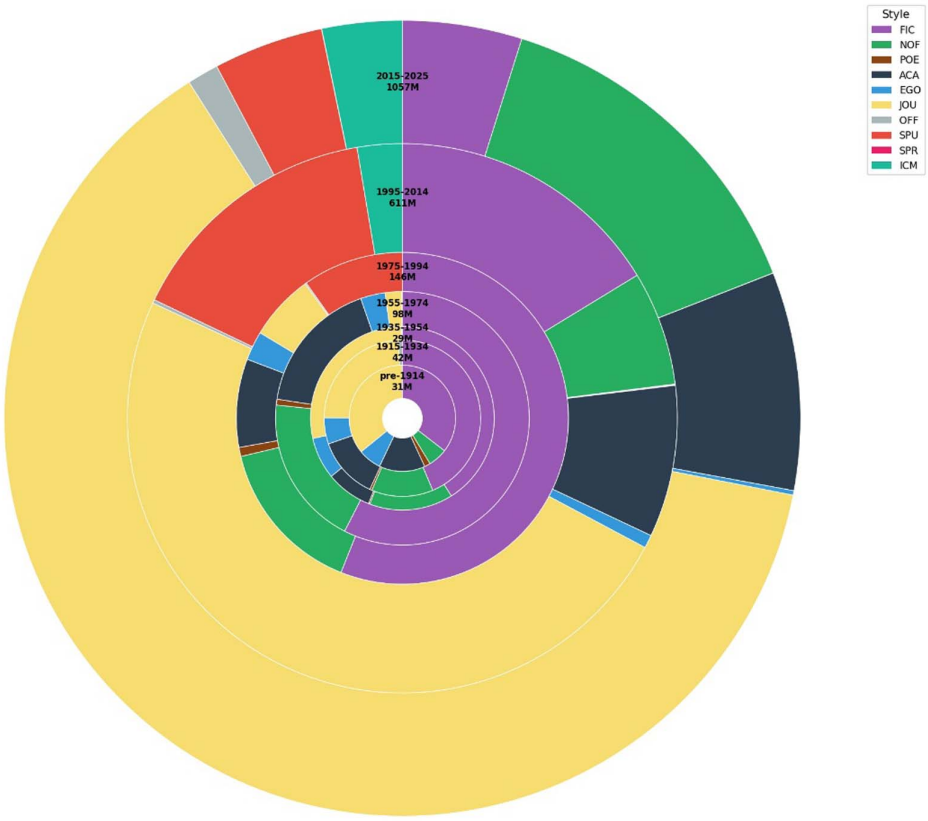


FIGURE 1 Style distribution by period in the corpus, where the ring area represents token count (Shvedova et al., 2017–2026)

An important feature of the corpus is that it is regularly expanded and updated annually. The texts that are added come from different time periods, genres, and geographical regions, reflecting the spelling and word usage characteristic of their respective times and places. Each time the corpus is expanded, it is lemmatized and tagged. After eighteen such cycles, the corpus is currently in version 19.

Lemmatization and morphological tagging are performed using a large morphological dictionary of Ukrainian (VESUM) and an NLP toolkit (NLP-UK). The dictionary provides lemmas,

word forms, morphological tags, and additional tags, e.g., to label slang and archaic words or spellings that do not conform to the current official orthography (Starko & Rysin, 2022). The NLP toolkit includes a tagger that draws on dictionary data and uses a dynamic tagging algorithm to handle various types of items, such as hyphenated words or compound nouns, not covered by the dictionary but decomposable into known components. The tagger has the function of automatically generating a list of unknown words from the entire corpus or from any text collection provided as input. This candidate list is manually verified, as explained below, and suitable lexical items of various kinds are added to the dictionary. Each such iteration of corpus and dictionary expansion leads to progressively better coverage of corpus data and, by extension, any Ukrainian texts. This concerns not only the most recent publications but also texts published 50, 100, or more years ago, as we add archaic, dialectal, and slang words, scientific terms, and various spellings that were common at different stages in the past. Thus, the dictionary is used as an exclusion source (Janssen, 2009) against the corpus.

3 Procedure

The complete procedure for adding new lexical items to the dictionary comprises the following steps:

1) automatic generation of candidate lexemes (OOV vocabulary in a current version of the corpus)

2) manual selection of new words to be included in the dictionary

3) assignment of inflection codes to the selected items

4) updating the dictionary

5) lemmatization of a new, expanded version of the corpus using the updated version of the dictionary

6) repeating the procedure.

Steps 1, 4, 5, and 6 are self-explanatory. Step 3 aims to assign the correct code to a given lemma, thereby automatically generating the full paradigm of correct word forms. This step is more technical in nature and is not covered here. Step 2 involves several actions, of which some combination is undertaken in each case, depending on the item in question and its familiarity to the annotator:

2a) check whether the item is a misspelling; if so, discard

2b) check whether the item is a word form of a lemma already included in the dictionary

2b) check whether it is an alternative spelling of a known lemma

2d) look up the item in the corpus to understand how and when it is or was used

2e) interrogate online dictionary collections and in-house dictionary indices to see if the item is recorded and if so, what information is supplied, e.g., usage labels, part of speech, endings, etc.

2f) search in other corpora and/or perform an online search to estimate the item's currency

2g) decide whether the lexical item merits inclusion in the dictionary

2h) if so, establish whether a restrictive label is necessary.

A list of candidate items automatically generated from the corpus contains a lot of noise, primarily in the form of misspelled words. Apart from a handful of recurring patterns that can be exploited to narrow down this list, it is reviewed manually (point 2a). Point 2b refers to instances in which the candidate item is unrecognized but is actually a form of a known lemma (one included in the dictionary). For example, *intellihentsiiam* is the plural dative of *intelligentsia* 'intelligentsia,' a lemma that previously had only singular forms in VESUM. In cases like this, the inflection code for the given lemma is modified to include plural forms in the dictionary. Alternatively, the candidate may be a substandard, archaic, or rare form of a known lemma. Here, the decision hinges on whether the item is attested at least 5 times in the corpus, with evidence from other textual sources and dictionaries serving as a supplement. Point 2c refers to variable spellings, a situation not uncommon for Ukrainian: if a word has already been included in the dictionary, its alternative spellings are added as well. This step requires knowledge of Ukrainian historical spelling systems.

Points 2d–2f describe the information-gathering stage applicable to candidate items unfamiliar to the annotator. Theoretically, if a word is well-known to the lexicographers, this stage could be skipped, but this occurs extremely rarely in practice at the current stage, when the dictionary has 444,000 lemmas and counting. Information is collected, first of all, from the GRAC corpus itself and then, if necessary, from other corpora and text collections, as well as from lexicographical sources. To this end, we extensively use online dictionary portals (E2U; R2U; SE; SUO), which are especially helpful for archaic or dialectal items unlikely to appear in online searches of modern texts. The goal is to assess the word's currency in the language at the present stage or in the past to make an informed decision (point 2g). The inclusion criteria are discussed in detail below. If a candidate item does not meet them, it is deemed a potential lemma in need of more textual evidence and is thus put on file for later consideration. The final decision (point 2h) is to assign restrictive labeling, for example, for slang, vulgar, archaic, and dialectal words, based on the inspection of empirical data. This is where the chronological depth and rich metalinguistic annotation of the GRAC corpus play a decisive part, enabling the compilers to make the necessary distinctions.

We approach the question of inclusion criteria from the standpoint of the dictionary's overall utility. Our dictionary has multiple applications. First and foremost, it is a practical dictionary designed for NLP purposes, specifically for lemmatization and morphological tagging. It has co-evolved in close tandem with our corpus and aims to ensure maximum lexical coverage of its data and, by extension, of other Ukrainian texts. Second, the dictionary and the accompanying NLP toolkit are the backbone of an advanced Ukrainian text checker, placing a high demand on keeping the dictionary up to date with the latest developments in the language. Third, the dictionary is used for human lookup via a web interface. Thus, its third most important priority is meeting the needs of online users, which, in practice, translates into the imperative to present full paradigms, provide grammatical information, and draw distinctions and label different categories of words. These are the overarching priorities considered in the decision-making process.

Our dictionary is empirically grounded and relies heavily on corpus data. Thus, frequency, dispersion, and duration of use are central to our lexicographical endeavor. At the same time, as Graeme Diamond (2016) demonstrated with the example of the *Oxford English Dictionary*, compilers of large dictionaries use supplementary criteria alongside purely quantitative indicators. A major factor in an empirically based lexicographic enterprise is the availability of text data. We are dealing with unevenly distributed data in a constantly expanding corpus. As illustrated in Figure 1 above, our corpus currently contains fewer than 100 million tokens for the pre-1914 era and each two-decade later period until 1974, reaching 146 million for 1975–1994, 611 million for 1995–2014, and over a billion for 2015–2025. The text acquisition

rate varies by period, with the most recently published data being added at the highest clip. Thus, we must distinguish between modern general vocabulary, for which we can expect ample corpus data, and lexical items restricted along some dimension (temporal, social, or geographical), for which textual evidence may be scarce. For the first group, we expect at least 20 occurrences in our corpus; the most recent coinages are, if needed, also explored in a larger corpus (ukTenTen22) and in online searches. This frequency threshold is lowered for items in the second group. Lexical dispersion is utilized to distinguish between items of common usage and items with high burstiness, such as technical or scientific terms. Given the large size of the dictionary, we are now adding words from low-frequency vocabulary bands, which are poorly attested in corpus data. As a result, the rule of thumb for their inclusion is the 5-5-5 criterion: a minimum of 5 occurrences across 5 independent sources over at least 5 years. This approach (previously used for the OED) has proven its utility in our circumstances, along with additional positive factors for inclusion that are currently taken into account for the *OED*, such as linguistic importance, continuity of usage, and lexicographical significance (Diamond, 2016). We also make use of additional corpora when needed. For example, slang words were explored in the Polish Automatic Web Corpus of Ukrainian (PAWUK), which contains many unfiltered online comments, while beekeeping terminology was assessed in a specially constructed corpus of texts.

In summary, there is no one-size-fits-all answer to the question of inclusion criteria. We rely on quantitative guidelines and draw on supplementary principles and additional resources to increase the dictionary's usefulness in every case.

4 Discussion

What counts as a new or unregistered word? For our purposes, new words are understood as those that have appeared in the language (more exactly, in common usage) relatively recently and have not been recorded in Ukrainian dictionaries. An example is a group of words associated with the COVID-19 epidemic. Before 2019, monolingual Ukrainian dictionaries listed only the technical term *koronavirusy* 'corona viruses' (Busel, 2005). Meanwhile, the COVID-related adjectives *koronavirusnyi* and *kovidnyi* do not appear in major Ukrainian dictionaries despite being amply attested in corpus data. Using our corpus, as well as other corpora, text collections, dictionaries, and online search, we are able to establish with a high degree of certainty whether a word is a new arrival.

In contrast, unregistered, or overlooked, words are those that have been used in the language for quite a long time to be recorded by lexicographers but, for various reasons, are not found in existing dictionaries. Yet, if they meet the inclusion criteria described above, they are added to our dictionary with an appropriate label. For example, the verb *rozpasiiuvatysia* 'to become angry, to fly into a passion' is attested in the corpus (27 occurrences) since the late 1960s but is not included in major dictionaries. The "unregistered" status is established by identifying a lexical item in corpus data and cross-checking it in other dictionaries, dictionary portals, and in-house indices compiled from multiple lexicographic sources.

Notably, there is a third category, viz., words recorded in other large Ukrainian dictionaries but not (yet) included in ours. This is due to differences regarding inclusion criteria and, most importantly, the overall approach to dictionary compilation. Our dictionary is corpus-based, and each entry must be supported by sufficient textual evidence. If evidence is lacking at the present stage, the lexical item is not added, but this does not necessarily mean the word is or was not used in Ukrainian. We constantly consult other Ukrainian dictionaries and lexicons (Baloh et al., 2017) and treat lexicographical records as an additional positive factor for inclusion. As we continue developing our dictionary, this category will shrink, with processed words

being either included in our dictionary or put on file for a future round of revision with a more extensive empirical basis.

Furthermore, new and previously unregistered word forms are added to the lemmas already in the dictionary, provided there is sufficient corpus data. Our dictionary has a practical, rather than normative, focus, so it is much more descriptive than other Ukrainian dictionaries and features rich paradigms. These include not only standard but also substandard, archaic, and rare variants, all added based on corpus data, except for several regular patterns of variation, where items are automatically included in all relevant paradigms via inflection codes. For example, the substandard verbal form *robym* 'we are doing' is listed along with the standard *robymo* for the lemma *robyty* 'to do,' and this applies to all such forms in all verbal paradigms.

Over the years, the repeated application of the procedure described in section 2 has yielded thousands of lexical items that meet the inclusion criteria. To illustrate, here are some recent additions representing distinct productive domains: abstract nouns (*bezavtorytetnist'* 'absence of authority,' *stereotypotvorennia* 'stereotype generation'), Anglicisms (*storytelinh* 'storytelling,' *suburbanizovanyi* 'suburbanized'), feminine personal nouns (*krainoznavytsia* '(female) country studies expert,' *lipidolohynia* '(female) lipidologist'), scientific terminology (*subkontsept* 'subconcept,' *sirtuin* 'sirtuin'), IT terms and slang words (*bekapuvannia* 'backupping', *bahovanyi* 'containing bugs'), colloquialisms (*arkhivchyk* 'little archive,' *bezalkoholka* 'non-alcoholic beverages'), and slang (*diudyk* 'detective story,' *rubylovo* 'intense fighting').

A large and diverse corpus is crucial for a comprehensive exploration and revision of the Ukrainian lexicon. At the same time, the absolutely indispensable components of the pipeline presented here are the dictionary, which serves as an exclusion source, and an NLP tool to automatically generate a list of candidate items. We have applied our dictionary and tagger to multiple other corpora and datasets to extract unknown words for review. The long lists generated this way have been partly processed, and selected lexical items have been added to the dictionary. In principle, any Ukrainian text or collection of texts can be subjected to the procedure described above.

5 Conclusions

The combination of an electronic dictionary used as the exclusion source, a constantly growing corpus, and an NLP toolkit has proven effective in identifying new and unregistered lexical items in Ukrainian. This approach has the potential to significantly improve the lexicographic description of Modern Ukrainian by providing it with a robust empirical foundation.

A major challenge we have encountered is the scarcity, limited variety, and uneven distribution of corpus data, especially for more distant historical periods, as well as a lack of spoken data. These are the high-priority directions for the corpus compilers.

By design, our dictionary does not provide semantic information. Still, it can be used for compiling monolingual and bilingual dictionaries and serve as the foundation for a corpus-based explanatory dictionary of Modern Ukrainian. Our goal is to bring its register to at least 500,000 entries while adhering to the project's empirical criteria.

References

AU: *Araneum Ucrainicum II corpus*. Retrieved July 14, 2026, from http://unesco.uniba.sk/aranea/

Baloh, V. O., Lozova, N. Je., Tymenko, L. O. & Tyscenko, O. M. (2010). *Novi i aktualizovani slova ta znachennia* [New and Reactivated Words and Meanings]. Vydavnychyi dim Dmytra Buraho.

Baloh, V.O., Boiarchuk, O.S., Didun, L.I., et al. (2017). *Ukrainskyi leksykon kintsia XVIII – pochatku XXI st.: slovnyk-indeks: u 3-x t.* [Ukrainian Lexicon from the End of the 18th until the Early 21st Century: An Index Dictionary, in 3 vols.] Vydavnychyi dim Dmytra Buraho.

Belei, L. Zymovets, HV, Honcharenko, A., Kivu, M. & Oleksandruk, I. (2022). *Slovnyk novitnikh anhlizmiv* [A Dictionary of New Anglicisms]. Naukova dumka.

Busel, V. (Ed.). (2005). *Velykyi tlumachnyi slovnyk suchasnoi ukrainskoi movy* [A Large Explanatory Dictionary of Modern Ukrainian]. Perun.

Diamond, G. (2016). Making Decisions about Inclusion and Exclusion. In Ph. Durkin (Ed.), *The Oxford Handbook of Lexicography* (pp. 532–545). Oxford University Press.

E2U: *Anhliisko-ukrainski slovnyky* [English-Ukrainian Dictionaries]. Retrieved June 1, 2026, from https://e2u.org.ua/

Janssen, M. (2009). Detección de Neologismos: una perspectiva computacional. *Debate Terminológico*, 5, 68–75.

Karpilovska, Ye. et al. (2013). *Aktyvni resursy cuchasnoi ukrainskoi nominatsii: ideohrafichnyi slovnyk novoi leksyky* [Active Resources of Contemporary Ukrainian Nomination: An Ideographic Dictionary of New Words]. KMM.

Neliuba, A. (2012). *Slovotvorchist' nezalezhnoi Ukrainy. 1991–2011: Slovnyk* [Word Formation in Independent Ukraine. 1991–2011: A Dictionary]. Kharkivske istoryko-filolohichne tovarystvo.

Neliuba, A. (2017). *Slovotvorchist' nezalezhnoi Ukrainy. 2012–2016: Slovnyk* [Word Formation in Independent Ukraine. 2012–2016: A Dictionary]. Kharkivske istoryko-filolohichne tovarystvo.

Neliuba, A. & Redko, Ye. (2022). *Leksyko-slovotvirni innovatsii (2017–2021): Slovnyk* [Lexical-Derivational Innovations (2017–2021): A Dictionary]. Kharkivske istoryko-filolohichne tovarystvo.

NLP-UK: *NLP-UK toolkit for Ukrainian*. Retrieved June 1, 2026, from https://github.com/brown-uk/nlp_uk

PAWUK: *Polish Automatic Web corpus of the Ukrainian language*. Retrieved June 1, 2026, from https://pawuk.ipipan.waw.pl/

R2U: *Rosiisko-ukrainski slovnyky* [Russian-Ukrainian Dictionaries]. Retrieved June 1, 2026, from https://r2u.org.ua/

SE: *Slovnyky i entsyklopedii* [Dictionaries and Encyclopedia]. Retrieved June 1, 2026, from https://slovnyk.me/

Shvedova, M. (2026). General Regionally Annotated Corpus of Ukrainian: Recent developments and future plans. In P. Bański, D. Knight, M. Kupietz, A. Witt & A. Wróblewska (Eds.), *Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026* (pp. 66–70). European Language Resources Association.

Shvedova, M., von Waldenfels, R., Yarygin, S., Rysin, A., Starko, V., Nikolajenko, T., Lukashevskyi, A., et al. (2017–2026). *GRAC: General Regionally Annotated Corpus of Ukrainian*. Retrieved July 14, 2026, from http://uacorpus.org

Starko, V. & Rysin, A. (2022). VESUM: A Large Morphological Dictionary of Ukrainian As a Dynamic Tool. In V. Lytvyn, N. Sharonova, I. Jonek-Kowalska, A. Kowalska-Styczen, V. Vysotska, Ye. Kupriianov, O. Kanishcheva, O. Cherednichenko, Th. Hamon & N. Grabar (Eds.), *Computational Linguistics and Intelligent Systems. Proc. 6th Int. Conf. COLINS 2022, Gliwice, Poland* (pp. 71–80). Silesian University of Technology.

SUO: *"Slovnyky Ukrainy" online* ["Dictionaries of Ukraine" online]. Retrieved June 1, 2026, from https://lcorp.ulif.org.ua/dictua

SUM: *Slovnyk ukrainskoi movy* [A Dictionary of the Ukrainian Language], vols. 1–16. Retrieved June 1, 2026, from https://sum20ua.com

Turovska, L. & Vasylkova, L. (2008). *Novi slova ta znachennia. Slovnyk* [New Words and Meanings. A Dictionary]. Dovira.

ukTenTen22: *Ukrainian Web corpus 2022*. Retrieved June 1, 2026, from https://www.sketchengine.eu/uktenten-ukrainian-corpus/

UP19: *Ukrainskyi pravopys* [Ukrainian Orthography]. (2019). Naukova dumka.

UP26: *Ukrainskyi pravopys* [Ukrainian Orthography]. (2026). Natsionalna komisiia zi standartiv derzhavnoi movy. Retrieved June 1, 2026, from https://mova.gov.ua/storage/app/sites/19/2026/rishennja-komisiji/01-03/sdm-ukrayinskii-pravopis-vidannia.pdf

Vakulenko, S. & Karunyk, K. (2018). *Ukrainska mova: unormuvannia, rozunormuvannia, perevnormuvannia (1920–2015)* [The Ukrainian Language: Standardization, Destandardization, Restandardization (1920–2015)]. Kharkivske istoryko-filolohichne tovarystvo.

VESUM: Rysin, A. & Starko, V. *Large Electronic Dictionary of Ukrainian (VESUM)*. Web version 6.8.2-SNAPSHOT. Retrieved July 14, 2026, from https://vesum.nlp.net.ua

Contact information

Vasyl Starko
Ukrainian Catholic University
v.starko@ucu.edu.ua

Andriy Rysin
arysin@gmail.com

Pius ten Hacken

THE THEORETICAL FOUNDATION OF DICTIONARY DEFINITIONS

Abstract The place of theory in lexicography has been the matter of intensive debate. There are different views of what a theory is. In line with most approaches in philosophy of science, it is assumed here that a theory has to provide an explanation. It is argued that lexicography should be considered as an applied science, i.e. the same category as medicine. In the analysis of applied science, the identification of problems and the specification of evaluation criteria for solutions are crucial. In lexicography, problems are the reasons why users consult a dictionary. Problem types are distinguished on the basis of the kind of information they need and the background knowledge they have. For certain problem types, definitions can be used as solutions, but the requirements on the definition depend on the type of problem it is meant to solve. Guidelines issued to lexicographers working on a dictionary are not theories, because they do not explain. They only instruct. The theoretical foundations of a theory should identify criteria for distinguishing problem types.

Keywords Applied science; explanation; definition; theory

1 The position of theory in lexicography

The aim of this paper is to set out the theoretical foundations for the decisions involved in formulating dictionary definitions. The role of theory in lexicography has been a persistent issue for several decades. At the end of a detailed overview of the discussion, Rundell (2012, p. 84) concludes that "the widespread lack of enthusiasm among lexicographers for 'lexicographic theory' is not due to ignorance, laziness, or an antitheoretical mindset. It is simply a utilitarian impulse to not waste time on what looks unlikely to be useful". Béjoint (2010, p. 381) expresses a stronger view in a very brief chapter, stating that "I simply do not believe that there exists a theory of lexicography, and I very much doubt that there can be one." By contrast, Piotrowski (2013, p. 308) criticizes authors who "seem to think that if lexicographical theory would not be like those in the sciences then there is no theory at all". He concludes that "there are indeed theories of lexicography" (2013, p. 318).

As indicated in the quotation from Piotrowski (2013, p. 308), a crucial point is the interpretation of *theory*. In order to approach this question, Piotrowski (2013, pp. 305-306) consults a dictionary. Adamska-Sałaciak (2019) uses the corpus of Euralex proceedings as a source. A dictionary will give information about how the word has been used in a wide range of contexts on the basis of a lexicographer's interpretation of corpus data. The corpus used by Adamska-Sałaciak (2019) restricts the range of contexts, but does so in a way that takes usage among the speakers at Euralex congresses as normative.

Theory is sometimes used in a sense that is not founded in philosophy of science, but simply as the opposite of *practice*. An example is Erben's (2017) history of the theory of architecture. The idea of a theory of architecture is to discuss questions of the purpose of buildings, ideal forms and the relationship between buildings and the society they are designed for. A theory of architecture is then a set of considerations pertaining to such questions. It may result in an overview of different building styles and classify individual buildings. This is not the sense of *theory* I will adopt here.

In the philosophy of science, a crucial element of a theory is explanation. Musgrave (2004) gives a systematic overview of the use of *theory* in different approaches to the philosophy

of science and concludes that the most widely shared claim about theories is that they are "descriptions of reality [which] can be used to explain features of it" (2004, p. 101). This is also the sense Béjoint (2010, p. 381) refers to: "A theory is a system of ideas put forward to explain phenomena". Here I will adopt this sense of *theory* and explore how it can be applied to lexicography.

A central question raised by the focus on explanation is which phenomena should be explained. Here, I will discuss this for lexicography in relation to dictionary definitions. As a preliminary, Section 2 will discuss the distinction between empirical and applied science and argue that dictionary definitions are in the scope of the latter. Section 3 analyses the status of definitions in more detail. The practical aspects are then approached from a user's perspective in Section 4 and from a lexicographer's perspective in Section 5. On this basis, Section 6 formulates some general questions to be answered by a theory about dictionary definitions.

2 The nature of applied science

In the philosophy of science, the focus of the discussion about the nature of science is on empirical science. In empirical sciences, theories are invoked as explanations for phenomena that are observed, typically in experiments. In order to explain something, we have to ask a question about it. A mountain is not a question and cannot be explained as such. When we consider how mountains come about, we have a question of geology that can be the subject of a theory. The theory will describe a model of the world that explains the phenomena we observe. In the case of geology, the model will involve tectonic plates as well as physical forces studied also in other sciences.

Ten Hacken (2019) applies this idea of science to linguistics. Language is as little a question as is a mountain. A research programme in linguistics is determined by the selection of a question or a set of connected questions. Ten Hacken (2019) describes the research programme of Chomskyan linguistics as built around the questions of the nature of the knowledge of language in a speaker's mind, its origin in an individual speaker through language acquisition and the origin in the human species in the course of evolution. These are not the only possible questions that can be asked about language, but together they form a coherent basis for the development of theories.

Lexicography is not an empirical science. As Béjoint (2010, p. 381) states, "[l]exicography is above all a craft. … It does not try to explain phenomena." A dictionary is not a theory. As Wiegand (1984, p. 13) states, "[s]cientific activities as a whole are aimed at producing theories, and precisely this is not true of lexicographical activities". However, not all sciences are empirical sciences.

Perhaps the most prominent example of a science that is not an empirical science is medicine. Medicine is traditionally taken to be a science, but in medicine it is not sufficient to develop a theory explaining observations. The aim is to cure patients, not just to explain why and how they suffer from a disease. As described by Bynum (1994), significant progress in terms of science was made in the course of the 19th century. On one hand, there was progress in empirical sciences that produced knowledge that could be used to cure patients, on the other, specific medical theories came up. Among the empirical sciences, anatomy, the science of the human body, was important. In anatomy, it is not sufficient to identify organs and body parts. Theories of anatomy also explain their functions and interactions. Theories of anatomy are not medical theories. In medical theories, it is explained how a disease can be prevented or cured. This implies understanding the cause of the disease and of its progress, and explaining how a cure can interrupt or prevent it.

In the philosophy of science, little attention has been paid to applied science. A handbook such as Humphreys (2016) does not treat it. An essential consideration in identifying applied science is that the application of science is not sufficient. The paint used by a professional painter is produced in a process that was conceived by using knowledge of chemical and physical properties of materials, but this knowledge is not used to explain why the paint is adequate for a particular job. Painters have a practical knowledge of different paints and are not paid any extra for knowing the underlying chemical and physical theories. In applied science, theories are developed to explain how the solution solves the problem. Based on an analysis of medicine and of computational linguistics, ten Hacken (2024, p. 18) proposes four questions that determine a theory of applied science. They are listed in (1).

(1)

a. Problem identification: What are the boundaries of the problem?

b. Evaluation: How can it be determined that the problem has been solved?

c. Problem decomposition: How can the problem be analysed into components?

d. Knowledge selection: Which types of knowledge should be used to solve the component problems?

The starting point is the recognition of a problem, as in (1a). In medicine, the problem is a disease, but what constitutes a disease is often not clear until we understand how it comes about. The evaluation in (1b) is tightly linked to the identification of the problem. In medicine, it requires a set of criteria for measuring to what extent the disease has been halted or cured. Issues (1c) and (1d) pertain to the methodology applied in the solution. A theory of medicine explains why and how the cure works by analysing the natural process of the disease and the way a cure interferes with it.

In lexicography, a crucial question is the status of the dictionary. Ten Hacken (2009) argues that a dictionary should not be seen as a description of the vocabulary of a language, but as a tool for the user to solve lexical problems. In the central position it assigns to the dictionary user, this approach is in line with Tarp (2008). For (1a), we have to identify the problem for which the user consults the dictionary, for (1b) the criteria by which successful consultation of the dictionary can be measured. In (1c) and (1d) we have methodological questions. An example of (1c) is the question of how to trigger the relevant insight in the user. For (1d), a question is how to use corpora in the selection of information for an entry.

The answers to such questions are tightly linked to the qualified practice in solving problems. On this basis, ten Hacken (2024, p. 24) argues against a distinction between lexicography and metalexicography, where the former would be the activity of compiling a dictionary and the latter the research into this activity. There is no parallel distinction between medicine and what might be called *metamedicine*, where the former would involve the contact with patients and the latter the research into diseases. Medicine is a single field which includes the entire span of these two components and the space between them. If lexicography is seen as an applied science, there is scope for a wide range of intermediate activities in the same way as there is in medicine.

3 Definitions as a lexicographic tool

When we consider lexicography as an applied science, it is tempting to see the formulation of definitions as a problem in the sense of (1a). However, the purpose of a dictionary is not to

contain definitions, but to serve users by assisting them in solving lexical problems. A definition is rather a part of the solution. As such, we can see formulating a definition as a component in the sense of (1c).

Hanks (2016) gives a detailed analysis of the development of thoughts about the idea of definition since antiquity. Two main traditions can be distinguished. Ten Hacken (2022) calls their results *logical* and *lexicographic* definitions. In a logical definition, the boundary of the concept is the focus, in a lexicographic definition, the meaning of a word is described.

Logical definitions are used in contexts in which it is necessary to classify instances as belonging to a concept or not. From antiquity, logic has been a prominent example of a domain that requires such decisions. We also find such definitions in terminology. In Wüster's (1974) *Allgemeine Terminologielehre* ('General theory of terminology'), the purpose of specifying concepts is to achieve a sharp boundary between them (1974, p. 67). This requires a definition of the classical Aristotelian type consisting of a hyperonym and a set of necessary conditions that together are sufficient to determine whether something is an instance of the concept that is defined or not. This view is also adopted in ISO 1087:2019.

Lexicographic definitions do not start from a concept, but from a word. As Atkins & Rundell (2008, p. 405) state, "[e]xplaining what words mean is the central function of a monolingual dictionary". This justifies the use of *lexicographic* to identify this type of definition. As argued by Rosch (1973), the meaning of words in natural language is not determined by a set of necessary conditions specifying the boundary of the meaning, but rather by the prototypical cases. Labov (1973) investigated this experimentally for the meaning of *cup* and found that items that are further removed from the prototype show a gradual decrease in acceptance as a cup among informants, without any natural boundary to non-cups.

The choice between logical and lexicographic definitions depends on the purpose. They each have their function and it is according to the intended user that one type or the other is to be preferred.

4 The user's perspective: Problems

When we consider lexicography as an applied science, an analysis of the role of definitions will have to start from practical problems of the type a dictionary user is meant to solve by means of the information the dictionary provides. This somewhat convoluted way of expressing the position of dictionaries in the process of solving lexical problems follows from the argument summarized in Section 2 and presented in more detail in ten Hacken (2009, 2024). As there is no single, general type of dictionary user and each type of user has their own set of problems, we have to choose an example of a user in order to make the role of definitions clear.

A good example of a class of valid problem specifications is that a learner of English is unsure how to understand or use certain words. Some instantiations of actual questions of this type are given in (2).

(2)

a. What does *turtle* mean?

b. What is the difference between the verbs *use* and *utilize*?

In (2a), the solution can involve an image, a translation, or a description in English. In (2b), images are less useful. Whether a translation is useful depends on the languages the user

knows. If they know a language that has a similar distinction, this may be the most useful approach. A description of each verb may help, but a better solution may be a usage note addressing the difference directly.

General dictionaries are often less useful for questions such as (2), because they aim at a different set of users, typically native speakers. Native speakers often consult a dictionary to solve a conflict. Some examples are given in (3).

(3)

a. Can a vehicle intended for transporting people be called a *van*?

b. What is the difference between a *pill* and a *tablet*?

Questions such as (3) can come up in a discussion and different native speakers may have different opinions on them. For (3a), there are two possible answers, for (3b), some speakers may consider the two words synonyms, others may come up with a variety of differences they find crucial. In the case of (3a), a definition may address the question directly. For (3b), the differences may be deduced from a comparison of the two definitions of *pill* and *tablet*, or a separate usage note may target them explicitly.

Native speakers may also come across words they do not know, but these are usually of a different type to the one in (2a). Some examples are given in (4).

(4)

a. What is *osteoarthritis*?

b. What is special about a *lithium battery* compared to other batteries?

Questions such as (4) may arise in the context of a pub quiz, but there may also be more natural environments. They illustrate cases in which non-specialists may be confronted with terms. In (4a), we have an example of a medical term. People often come across medical terms for the first time when they apply to their own health or to the health of someone in their environment. In (4b), the word in question refers to a technical innovation that at some point touches a wider range of people than only domain specialists. In both cases of (4), a definition can be very useful, but it is crucial that the user of the dictionary understands it. Images or translations will probably be less successful, although in some cases they may support the definition.

The problems exemplified in (2) - (4) represent types that trigger the use of definitions. They are not meant as an exhaustive typology of problems for which users may consult a dictionary and look at the definition, but they can serve to illustrate the variety of problems.

5 The lexicographer's perspective: Guidelines

In describing the meaning of words, lexicographers use their knowledge of language. There is a long tradition of supporting the decisions they make by references to sources using the word in the relevant sense. Examples in a dictionary entry have a dual function in the sense that they may serve users and lexicographers. Svensén (2009, pp. 281-288) considers them from the perspective of the information about the word they convey to the user. In this context, he distinguishes authentic and editorial examples. However, as Mugglestone (2015) describes, Samuel Johnson used them mainly to demonstrate that his description of the meaning of a word was in accordance with proper use, i.e. with the use by authorities of the language. Such examples are used by lexicographers to support their description of the meaning in the definition.

The actual procedure of how to address questions of the types illustrated in (2) - (4) in a dictionary can remain implicit as long as the dictionary has only one author. Johnson (1747) does not really address the question of how he arrived at his descriptions of meaning, but only argues why the resulting description is trustworthy. His aim is to justify his decisions to the user in order to convince the user of the dictionary's authority. In this sense, his perspective is user-oriented. Such user-oriented presentations of the motivation for decisions are often found in prefaces to dictionaries. Many decisions in formulating definitions are based on intuitive evaluations of a range of evidence. User-oriented justifications use examples of how decisions were taken in particular cases in order to convince the user that the dictionary can generally be trusted.

Lexicographic guidelines go beyond prefaces. They are used when a team of lexicographers has to collaborate and produce a coherent result. Gilliver (2016) describes how, when James Murray realized that he could not compile the OED alone, he struggled with accepting another editor's definitions. His primary reaction was to rewrite them himself, which was of course not desirable in terms of task management. The combined purposes of justifying his approach (to himself and to his collaborators) and instructing other editors led to the formulation of lexicographic guidelines. Lexicographic guidelines are much more specific than prefaces. Whereas a preface contains instructions for using the dictionary, lexicographic guidelines are instructions for writing entries. They address questions such as (5).

(5)

a. For what types of word should an image be added?

b. How much information should be given about the meaning of a word?

c. In which cases should the difference in meaning between two words be described?

The questions in (5) can be related to the problems in (2) - (4). For designations of animals, as in (2a), we might agree that they should generally be illustrated. However, an image is much more specific than the meaning of a word. Therefore, for words like *mammal*, an illustration is less useful. Users may simply note whether they find an image, but lexicographers have to decide whether to include one in individual cases, so that (5a) should be answered in lexicographic guidelines. In (5b), the amount of detail in answering questions such as (4a) should be specified. It is also relevant in (3a). Whereas for (4a), it depends on how much detail a user wants, for (3a), it depends on which kind of questions about the meaning a user is likely to ask. The guidelines have to take into account the available space, especially in printed dictionaries, and the amount of effort the user has to invest in finding the relevant information. In (5c), it is the crossreferences that are in focus. In a case like (2b), the crossreference may be the subject of a usage note, but in (3b) the question is rather whether to address the difference directly in the definition.

According to Piotrowski (2013, p. 307), "Style Guides are theoretical statements that describe how to write a specific dictionary". He considers them the lowest level of a hierarchy of theories of lexicography. A first observation is that the use of "describe" is misleading. Lexicographic guidelines, or Style Guides, are instructions, not descriptions. They tell lexicographers what to do in response to questions such as (5). In view of the sense of *theory* developed in Sections 1 and 2, we should also note that lexicographic guidelines are not a theory. Their focus is on instruction, not on explanation. Good guidelines give clear answers to (5) that can be interpreted in the same way by all lexicographers on the team, so that the result is a coherent approach in the dictionary. They explain the strategy only in the sense of making clear what the strategy is. A theory of lexicography as an applied science would require explaining also why the strategy works.

6 Properties of lexicography as an applied science

In order to formulate some more specific ideas of what a theory of lexicography that can be applied to definitions should be like, it is worth comparing the situation in lexicography with the one in medicine, as this is probably the most developed applied science. Both medicine and lexicography aim to solve problems for individual people. From an epistemological perspective, we can identify two important differences in the way problems are solved in medicine and in lexicography. One relates to the interaction with the person whose problem needs to be solved, the other with the way problems are distinguished from each other.

In medicine, the problem to be solved is the health condition of a patient. If the patient corresponds to the dictionary user, the lexicographer takes the role of the medical personnel. The solution of the patient's problem consists of a diagnosis and a cure. The first important difference between medicine and lexicography is that in medicine, the diagnosis and the cure are administered in direct interaction between the medical personnel and the patient, whereas in lexicography, the user interacts with the dictionary, but not directly with the lexicographer. As a consequence, dictionary users have a much more active role, especially in the diagnosis of the problem. Conversely, lexicographers have to work on the basis of much less specific information about the actual problems the users want to solve.

The second main difference concerns the distinction between different problem types. In medicine, research focuses on the identification of diseases. Once a set of symptoms has been recognized as belonging together, it becomes a syndrome. The step from syndrome to disease is marked by the identification of the cause. This is a crucial point in finding a cure. The aim is to understand the development path of the disease to such an extent that it can be interrupted. A theory in medicine explains how the disease works and how the cure works against it. Historical overviews of modern research that show examples of this pattern are, for instance, van Doornum & van Helvoort (2020) and Eckerle (2023).

In lexicography, we do not find a concept corresponding to the disease in medicine. A crucial property of diseases is that they are discrete. Once our understanding has reached the point that enables us to recognize a disease, it is possible to determine whether a patient suffers from a particular disease or not. In the case of AIDS, the disease is caused by a particular virus, HIV. On this basis, Pepin (2011) can investigate when and how AIDS originated. As a consequence of the discreteness, we can delimit the problem for which a cure is necessary. This is a significant help in formulating a theory explaining the effectiveness of a cure. In lexicography, the problems do not fall into discrete types.

In the absence of discrete problem types, lexicography will have to work with prototypes in approaching the questions in (1). Instead of a precise delimitation of the problem in (1a), one has to start from a set of prototypical instantiations of the problem. As an adequate solution depends on the user's knowledge, such a prototype should specify the user type. Questions such as (2a) about *turtle* and (4a) about *osteoarthritis* do not belong to the same class, because (2a) is a typical question that may arise for a learner and (4a) a typical question for a native speaker. A crucial difference is that the user is likely to know the concept in (2a), but not in (4a). Such a division of problem types makes sense, because the solutions will reflect this distinction.

Apart from the user's background knowledge, another parameter distinguishing problem types is the user's purpose. In (2a) and (4a), we can assume that the user's purpose is to increase understanding. In such a case, a verbal definition competes with an image. The optimal distribution of tasks between the verbal definition and the image depends on the salience of visual properties of the concept.

In (3a), about *van*, we have a very different situation. Here the point is to settle a dispute. This requires a different type of definition, in which authority plays a significant role. The user assigns some degree of authority to the dictionary, because otherwise they would not consult it. In order to solve the user's problem adequately, the dictionary has to work with this expectation. The definitions of *van* in (6) illustrate how different dictionaries approach this.

(6)

a. a large covered wagon; a light vehicle used in transporting goods; a railway carriage or compartment for luggage, etc.; a caravan

b. a covered motor vehicle used for transporting goods or people ■ Brit. a railway carriage for conveying luggage, mail, etc. ■ Brit. a caravan.

c. **1** short for **caravan** (sense 1) **2** a covered motor vehicle for transporting goods, etc, by road **3** *Brit* a closed railway wagon in which the guard travels, for transporting goods, mail, etc **4** *Brit* See **delivery van**

The definition in (6a) is from Chambers (1998), (6b) is from COED (2011) and (6c) from Collins (2014). The relevant sense for (3a) is described in all three dictionaries using "vehicle" as a hyperonym and mentioning "transporting goods". The relation between these two properties differs, as does the selection of other properties. None of the definitions addresses (3a) explicitly. Only (6b) mentions "transporting … people", but "used for" is not the same as "intended for". Still, in (6b) the answer to (3a) should be taken to be positive. In (6c), the answer depends on whether "goods, etc" includes people. In (6a), the omission of any alternative to "goods" implies rather a negative answer. There seems to be also a certain degree of regional variation. OED (2000-2026 [2021]: *van* n[4], sense I.1.b) marks the use of *van* as a vehicle for carrying passengers as "Chiefly North American". The definitions in (6) are all from one-volume general dictionaries that cannot give the degree of detail that OED gives. The formulation in (6a) may be the most adequate characterization of the prototype, at least in Britain. (6b) places goods and people at the same degree of prominence, which does not match this prototype ideally. (6c) can be thought of as a reasonable compromise, but the vagueness of "goods, etc" is not optimal. The choice between the three is partly a matter of taste, partly a matter of which answer to (3a) one wants the user to arrive at.

Questions such as (3a) can also arise in a different situational context. If legal obligations or sanctions depend on the decision whether a particular vehicle is a van, we need a precise boundary of the concept. As argued by ten Hacken (2025), in such circumstances, a definition of what in Section 3 was called the logical type is necessary. Such definitions may be less than optimal for the purpose of understanding, especially for non-specialists, but they serve as a solution to a different problem.

The comparison of lexicography with medicine shows some differences in the nature of problems and the process of problem solution that should be reflected in a theory of lexicography. In lexicography, the dictionary serves as an intermediary between the person having the problem (the user) and the person solving it (the lexicographer). Moreover, the problems do not fall into discrete cases, but can only be characterized on the basis of prototypes.

7 Conclusion

The central question addressed here is how a theory can be devised that supports the decisions to be taken in formulating definitions in dictionaries. I approached this question from two angles. On one hand, Sections 1 and 2 argued that lexicography can be seen as an applied

science and that applied science requires explanatory theories. The focus on definitions, elaborated in Section 3, orients the discussion of lexicography away from questions such as the representation of phonological and syntactic information. On the other hand, Sections 4 and 5 apply the analysis of applied science summarized in (1) to the actual use and compilation of dictionaries. Section 4 presented some problems from the user's perspective, Section 5 some problems from the lexicographer's perspective. Section 6 elaborates some characteristics of problems in lexicography by comparing them to problems in medicine.

In a view of lexicography as belonging to applied science, definitions are meant to solve problems for which users consult the dictionary. The underlying theory should address the issues in (1). It should identify classes of problems (1a) and criteria for determining successful solutions (1b). In contrast to medicine, problems are not discrete, but should be treated as prototype concepts. Different problems depend on what the user wants to know and what the user knows already. Different prototypes will be determined by the type of solution that is required. As lexical problems can often not be decomposed, (1c) is less prominent in lexicography. An important point of (1) is that the selection of knowledge only appears as (1d), the last issue. A theory in lexicography should not start from the selection of knowledge, but derive the selection of knowledge from the analysis of the problem. Some examples of problem types are listed in (7).

(7)

a. matching an unknown word with a known concept, as in the case of *turtle* for a learner

b. getting an approximation of the meaning of an unknown word, as in the case of *osteoarthritis*

c. settling an informal dispute, as in the case of whether a *van* can be intended for transporting people

d. settling a terminological dispute, as in the case of whether the regulations for vans apply to a particular vehicle

The list in (7) is by no means exhaustive, but it is meant to be suggestive of how problem types at the basis of a theory of dictionary definitions can be distinguished. Each problem type in lexicography corresponds to a disease in medicine in the sense that it requires its own cure. A theory of lexicography will explain for problem types such as (7) how information provided by a dictionary can serve as a solution.

References

Adamska-Sałaciak, A. (2019). Lexicography and Theory: Clearing the Ground. *International Journal of Lexicography*, 32, 1-19.

Atkins, B.T. S. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press.

Béjoint, H. (2010). *The Lexicography of English*. Oxford University Press.

Bynum, W.F. (1994). *Science and the Practice of Medicine in the Nineteenth Century*. Cambridge University Press.

Chambers (1998). *The Chambers Dictionary*. Chambers Harrap.

COED (2011). *Concise Oxford English Dictionary*, 12th edition. Oxford University Press.

Collins (2014). *Collins English Dictionary*, 12th edition. HarperCollins.

van Doornum, G. & van Helvoort, T. (2020). *Corona & Co: Een eeuw onderzoek naar virussen in Nederland*. Walburg.

Eckerle, I. (2023). *Von Viren, Fledermäusern und Menschen: Eine folgenreiche Beziehungsgeschichte*. Droemer.

Erben, D. (2017). *Architekturtheorie: Eine Geschichte von der Antike bis zur Gegenwart*. Beck.

Gilliver, P. (2016). *The Making of the Oxford English Dictionary*. Oxford University Press.

ten Hacken, P. (2009). What is a Dictionary? A View from Chomskyan Linguistics. *International Journal of Lexicography*, 22, 399-421.

ten Hacken, P. (2019). The research programme of Chomskyan linguistics. In A. Kertész, E. Moravcsik & C. Rákosi (Eds.), *Current Approaches to Syntax: A Comparative Handbook* (pp. 549-572). De Gruyter.

ten Hacken, P. (2022). Lexicographic and Logical Definitions for Terms. *Lingue e Linguaggi*, 49, 13-29.

ten Hacken, P. (2024). Theories of the Lexicon and Decisions in Dictionary Making. In T. Margalitadze (Ed.), *Proceedings of the I International Conference Lexicography in the XXI Century* (pp. 13-25). Centre for Lexicography and Language Technologies, Ilia State University.

ten Hacken, P. (2025). Terminology from a lexicographic perspective. In I. Dias, R. H. Gouws, A. Lobenstein-Reichmann & S. J. Schierholz (Eds.), *Regarding the Significance of Lexicography: Europe and Australia/Zur Bedeutsamkeit der Lexikographie: Europa und Australien*, (pp. 115-132). De Gruyter.

Hanks, P. (2016). Definition. In P. Durkin (Ed.), *The Oxford Handbook of Lexicography* (pp. 94-122). Oxford University Press.

Humphreys, P. (Ed.) (2016). *The Oxford Handbook of Philosophy of Science*. Oxford University Press.

ISO 1087:2019. *Terminology work and terminology science – Vocabulary*. Genève: International Organization for Standardization.

Johnson, S. (1747). *The Plan for a Dictionary of the English Language*. Repr. in T. Fontenelle (Ed.) (2008). *Practical Lexicography: A Reader* (pp. 19-30). Oxford University Press.

Labov, W. (1973). The Boundaries of Words and Their Meanings. In C.-J. N. Bailey & R. W. Shuy (Eds.), *New Ways of Analyzing Variation in English* (pp. 340-373). Georgetown University Press.

Mugglestone, L. (2015). *Samuel Johnson & the Journey into Words*. Oxford University Press.

Musgrave, A. E. (2004). Explanation, description and scientific realism. In H. Keuth (Ed.), *Logik der Forschung - Karl Popper* (pp. 83-102). Akademie.

OED (2000-2026). *Oxford English Dictionary*, Third edition, edited by John Simpson and Michael Proffitt. Oxford University Press.

Pepin, J. (2011). *The Origins of AIDS*. Cambridge University Press.

Piotrowski, T. (2013). A theory of lexicography – is there one? In H. Jackson (Ed.), *The Bloomsbury Companion to Lexicography* (pp. 303-320). Bloomsbury.

Rosch, E. H. (1973). On the internal structure of perceptual and semantic categories. In T. E. Moore (Ed.), *Cognitive Development and the Acquisition of Language* (pp. 111-144). Academic Press.

Rundell, M. (2012). It works in practice but will it work in theory?' The uneasy relationship between lexicography and matters theoretical. In R. V. Fjeld & J. M. Torjusen (Eds.). *Proceedings of the 15th EURALEX International Congress* (pp. 47-92). Dept. of Linguistics.

Svensén, B. (2009). *A Handbook of Lexicography: The Theory and Practice of Dictionary-Making*. Cambridge University Press.

Tarp, S. (2008). *Lexicography in the Borderland between Knowledge and Non-Knowledge: General Lexicographical Theory with Particular Focus on Learner's Lexicography*. Niemeyer.

Wiegand, H. E. (1984). On the structure and contents of a general theory of lexicography. In R. R. K Hartmann (Ed.), *LEXeter'83 Proceedings: Papers from the International Conference on Lexicography at Exeter, 9-12 September 1983* (pp. 13-30). Niemeyer.

Wüster, E. (1974). Die Allgemeine Terminologielehre: Ein Grenzgebiet zwischen Sprachwissenschaft, Logik, Ontologie, Informatik und den Sachwissenschaften. *Linguistics*, 119, 61-106.

Contact information

Pius ten Hacken
Institut für Translationswissenschaft, Universität Innsbruck
pius.ten-hacken@uibk.ac.at

Lars Trap-Jensen & Henrik Lorentzen

REVISING THE USAGE-LABELLING SYSTEM OF THE DANISH DICTIONARY[1]

Abstract This article discusses the ongoing revision of the usage-labelling system in *Den Danske Ordbog* (DDO, 'The Danish Dictionary') in response to increasing public attention on discriminatory and otherwise socially marked language. In recent years, the dictionary has received growing criticism concerning the treatment of identity-related terms, revealing limitations in the existing inventory of stylistic and evaluative labels. The article examines the lexicographic challenges involved in balancing descriptive dictionary practice with users' expectations for guidance and contextualisation. To address these issues, DDO is developing a revised system based on a closed inventory of evaluative labels, optional modifiers, and controlled combinations. The new model is supplemented by graphical warning symbols and expandable usage notes inspired by practices in dictionaries such as *Merriam-Webster* and *Duden*. The proposed revision seeks to establish a compromise between systematic description, user guidance, and the scientific legitimacy of descriptive lexicography.

Keywords Usage labels; marked language; discriminatory language; dictionary revision; sociolinguistics; Danish lexicography; identity terms; corpus lexicography

1 Introduction and Background

It has hardly escaped anyone's attention that, in recent years, language has become a site of contestation. Growing social awareness and political focus on the privileges of particular societal groups have led to increased sensitivity to discrimination and perceived injustices. Part of this struggle – and the concomitant polarisation that has attended it – is reflected in language, making it increasingly difficult for language users to navigate without conveying particular signals, whether consciously or unconsciously.

For this reason, the field of marked language and its description has received particular attention among lexicographers in recent years, and in this paper we outline how the inventory of usage labels in *Den Danske Ordbog* ('The Danish Dictionary', DDO), along with the principles governing their application, has been reconsidered, resulting in a revised system scheduled for implementation from autumn 2026.

In particular, words relating to parameters of discrimination – such as gender, ethnicity, sexual orientation, and physical or mental challenges – have prompted renewed discussion about how lexicography should balance the descriptive approach of the dictionary with an increased need for guidance and contextualisation.

The compilation of DDO was initiated in the 1990s and the resulting work was first published as a six-volume printed dictionary between 2003 and 2005. Since 2009 it has been available online and has subsequently been updated on a regular basis. The dictionary is a

[1] This paper is based on a talk given to our Danish colleagues at a meeting organised by LEDA, the association for Danish lexicographers, in November 2025 (Trap-Jensen & Lorentzen, 2026). The content also builds on earlier reflections on the same topic presented in Trap-Jensen (2020).

scholarly, monolingual dictionary whose primary target group consists of native speakers of Danish. Except for orthography, which follows the forms prescribed by the Danish Language Council in the official spelling dictionary *Retskrivningsordbogen* ('The Spelling Dictionary'), the dictionary is descriptive in its approach, and when it first appeared, it was the first corpus-based dictionary for Danish. Like other comparable national dictionaries, the DDO has, over the past 10–15 years, received a growing number of enquiries from users expressing dissatisfaction with the treatment of particular words and expressions in the dictionary and urging the editorial staff to revise their descriptions. Some of these cases have even attracted newspaper coverage and public debate, most recently in June 2026 when gender-neutral pronouns were included in the dictionary. Most of the discussions have concerned controversial terms potentially covered by anti-discrimination legislation, i.e. the parameters mentioned above. This development motivated the editorial team to address the issue in a more systematic manner, and in the present article we examine more specifically the ongoing revision of usage labels.

Although the long-term aim is to achieve a more nuanced and balanced treatment of the vocabulary within this sensitive area of language, the present article does not claim to address the issue in its entirety. A non-discriminatory treatment of words in the dictionary entails both more and less than the topic considered here. It involves more because other components of the macrostructure and microstructure are also involved, most obviously lemma selection, the formulation of dictionary definitions, and the choice of citations and other usage examples included in dictionary entries. At the same time, however, the inventory of usage labels concerns more than discriminatory and racist words and expressions alone. Thus, although debates concerning discrimination and exclusion have provided the starting point and an important motivation for the revision, they are far from identical with the subject matter of this article.

Brewer (2016, pp. 489-490) quotes Richard Bailey (1980), who argues that 'labelling is the least scientific and most artistic part of the lexicographer's task (. . .) Status labels arise from a judgement about usage rather than a judgement about meaning, and hence emerge from the lexicographer's sense of language'. In other words, usage labels are difficult to apply objectively and descriptively because they rely on the lexicographer's linguistic intuition as a native speaker and concern phenomena that are not directly observable in actual language use. We agree to that and would like to add that, even if it were possible to devise a labelling system that could be applied consistently, there is no guarantee that users would interpret and understand the labels in the same way. This probably explains why virtually every dictionary employs its own set of labels.

2 Marked Words: the Revision Project

DDO has always provided comments in the form of usage labels on words that deviate from neutral usage, whether in upward or downward direction, including *slang*, *formal*, *informal*, *jocular*, and other marked categories. In the original model, labels were entered in a dedicated but open field, while the principles governing how the field was to be filled in were regulated by the dictionary's editorial guidelines. According to the guidelines, a label such as *slang* was considered sufficient, the underlying assumption being that slang by definition belongs to a lower register and therefore requires no further qualification. This assumption has increasingly been called into question, and editors have in several cases felt it necessary to supplement a slang term with additional information about its usage.

TABLE 1 Slangwords with additional usage labels in DDO (and their English explanations)

bagbordsindianer	'backside Indian' (for homosexual man)
campinghabit	'camping suit' (for track suit)
fisselette	used for promiscuous woman or effeminate man
forsuttet bolsje	'sucked sweet' (for promiscuous woman)
frikadelle	'meatball' (for a bad actor)
glatnakke	'smooth back of the neck' (for a stupid person)
jyske dollars	'Jutland dollars' (for black money)
jøde (verb)	'to Jew' (for steal)
kran	'crane' (for wife/girlfriend)
købekone	'mail-order bride'
ladeport	'barn door' (for bottom, bum)
lagengymnastik	'sheet gymnastics' (for sex)
mappedyr	'briefcase animal' (for office worker, bureaucrat)
overrislet	'irrigated' (for being tipsy)
paryk	'wig' (for hair)
pølsefest	'sausage fest'
skaffedyr	'providing animal ´ (for a male supporter)
snakkemaskine	'chatterbox'
svagpisser	'faint pisser' (for weakling/wimp)

Table 1 lists the majority of cases where this was done. If we consider just a single example, the word *skaffedyr* illustrates the problem (cf. Figure 1, see also section 4 for another example). Although originally coined as a creative slang formation with a literal meaning 'provide-animal', i.e. '(male) provider', the word is often used as a derogatory term for the father in a traditional family, making the use of the word problematic in contemporary usage. Labelling the word simply as slang was therefore considered insufficient, and the editor added a further style label to describe the term's jocular or pejorative value.



FIGURE 1 The entry *skaffedyr* in DDO

Words relating to identity have proven particularly difficult to describe unambiguously as there is often limited agreement within the language community on their status. Ethnic designations such as *buskmand*, *hottentot*, *indianer*, and *eskimo* ('Bushman', 'Hottentot', '(North American) Indian', 'Eskimo') raise postcolonial issues, while terms relating to gender identity or roles – such as *trofækone* ('trophy wife') and *pigefornærmet* ('easily offended like a girl') – involve complex questions about who is targeted by the derogation. Such examples have made it clear that the original labels in DDO are not sufficient to convey both register and broader sociocultural context.

3 Other Problems with Existing Register Labels

We have also found that users often have problems with the correct interpretation of the label *informal*. In the original DDO system, this label covers a very broad – and arguably too broad – range. Relatively harmless words such as *smuksak* ('beauty') and *murstensroman* ('fat volume' (novel)) are placed in the same category as taboo-related vocabulary concerning sex, excretion and other bodily functions. This conflation makes it difficult for users to assess both the type and the degree of markedness.

Although a stronger label (*very informal*) is available, it is used only sparingly and does not sufficiently separate everyday informal language from strongly offensive or vulgar expressions. From the user feedback we have received, we have learned that the two categories are difficult to decode, with the result that we risk failing to meet users' need for clear guidance, especially in relation to potentially offensive language.

The biggest challenge, however, remains that an increasing number of identity-related terms are becoming controversial as the speech community is split between 'woke' and 'anti-woke' interpretations of the terms and is lacking consensus regarding their status and appropriate use. Language thus becomes entangled in political dispute, and words become the ammunition in the battle.

4 Towards a Closed Inventory of Labels

In response to these challenges, a tendency can be traced among the DDO editors to supplement individual entries with additional explanations when existing labels proved inadequate. An example was *skaffedyr* in Figure 1; another is *hottentot*, shown in Figure 2.



FIGURE 2 The entry *hottentot* in DDO

Figure 2 demonstrates how the need for detailed information may lead to rather lengthy explanations. This is possible because the XML element in which the editors write is a free-text field. Although understandable, the solution is not without problems. In some cases, the information given raises more questions than it answers, e.g. in the case of *hottentot* how and why the term reflects colonial-era attitudes. In other cases, the information may be perceived as taking a stance in a polarised debate. Moreover, lengthy explanations like these tend to attract disproportionate attention, which may be distracting for users who just looked up to check a particular word for pronunciation or etymology.

The editorial aim was therefore to return to shorter, more standardised labels selected from a closed inventory. The goal was to create a system that is both manageable from an editorial perspective and immediately intelligible to users. Compared to the original system, the new inventory is slightly expanded, and principles are introduced for the number of labels and how they can be combined and modified.

As mentioned, the inventory of the original system for style value was specified in the editorial guidelines, comprising the following: *informal, formal, derogatory, jocular, elevated, euphemistic, technical, poetic, considered incorrect, ironic, non-technical usage, disparaging, sarcastic, affectionate, clichéd*, and a few more. As the labels were inserted in an open field element, the complete list consists of no less than 270 different designations, the majority of which, however, is either combinations of the above (*jocular or derogatory, formal or elevated*), modifications (*often poetic, sometimes jocular*), combinations of style and other usage labels (*old-fashioned or formal, jocular or children's language*), or both modifications and combinations of all these. The broader system of usage labels in addition to style contains also Slang, Communicative situation, Frequency (e.g. *infrequent*), Geography (e.g. *especially Eastern Denmark*), Chronology (e.g. *historical*), Medium (e.g. *especially spoken language*), and Genre (e.g. *in biblical texts*), of which in particular the first two are important in this connection. We have already mentioned the problems that may arise when Slang appears as the sole label. The field Communicative situation is used to describe more pragmatic usages such as *used as a term of abuse, used when addressing a person, used as exclamation, used as a mild curse*, and for this reason, this is also the element used by the editors who feel the need to supplement the description with more explicit guidance for controversial terms.

4.1 The Revised Model

If we now turn to look at the new model for style labels, we propose a solution that consists of three elements for style:

1. **A closed set of (revised) evaluative labels**: *derogatory, (may be perceived as) offensive, coarse language, euphemistic, informal, formal, jocular, poetic, elevated, ironic, sarcastic*
2. **A closed set of modifiers**: *sometimes*, *often/especially*, *very*
3. **Rules for combinations**: a maximum of two evaluative labels, optionally modified (e.g. *informal*, *sometimes derogatory*)

The labels are similar to the original set, with the following changes: *offensive* is introduced as a new designation, and *coarse language* is separated from the original broad category *informal* to describe the most extreme cases, especially vulgar and taboo terms. The style label *considered incorrect* was removed from the inventory because the content is strictly speaking not about style but rather about grammar or linguistic correctness. Likewise, very infrequent values were

removed (*affectionate, disparaging, clichéd*) and will be solved by other means in the implementation phase.

It was a matter of discussion whether the terms *jocular*, *ironic*, and *sarcastic* may be too fine-grained for consistent use and user interpretation (cf. Svensén, 2009, p. 328), but as the designations are conceptually clearly discernible, it was in the end decided to maintain the distinction. Another consideration was the introduction of a label such as *value-laden* (positive or negative) to apply to terms that are deliberately used in framed communication, such as *cafépenge* (literally 'café money') for 'student grant' or *fjumreår* (literally 'dither year') for 'gap year'. However, it was decided that it might prove difficult to apply this label consistently and it was not included.

The values of the style labels selected from the closed set will be supplied with mouse-over texts which in more detail explain the designations, e.g. *derogatory*: 'the label *derogatory* indicates that the word used in this sense expresses contempt or disdain towards someone or something' or *informal*: 'the label *informal* indicates that the word used in this sense belongs to the unpretentious register of everyday language'. The assumption that such labels are self-explanatory is probably no longer valid, and thus it is vital to help the user understand what the editors wish to convey, cf. similar explanations on Dictionary.com.[2]

The style labels are the most important usage labels in this connection but not the only ones. In addition, DDO also provides labels for Subject domain, Slang, Figurative language, Chronology (*historical*, *children*, *young people*, *old-fashioned*, *outdated*), Medium (*spoken* or *written language*), Genre (e.g. *biblical*, *journalistic texts*, *official language*), Geography (indicating the origin of regionalisms), Communicative situation (e.g. for swear-words, curses or other pragmatic uses), and Frequency.

In the new model, the majority of these labels will be displayed before the definition of a given sense. In the original model, this was only the case for the domain and slang labels. For the editors, a prerequisite for the new solution was that the usage labels are short and readily understandable, so that they can function as signposts to the user, and this has now been achieved with the closed set of values. As such, they will not be preceded by a special marker but must be clearly identifiable as a different type of information through a distinct typographical mark, e.g. a different font colour.

Consequently, the only labels that remain in postposition are Genre, Communicative situation and the former style value *considered incorrect*, the first two mainly for pragmatic reasons: the content of the elements are too heterogenous to be replaced by a short designation, the last for reasons already explained. As the only ones, these will still appear after the marker SPROGBRUG, 'language usage', used in the original model for most usage labels (cf. Figure 1 and 2) and, at least for the time being, remain open field elements.

5 Graphical Warnings and Text Boxes

A key proposal in the revision is the introduction of a graphical warning for particularly controversial words. This idea is inspired in part by *Merriam-Webster*, where a warning icon (a yellow triangle with an exclamation mark) directs users to an extended usage note. A similar approach can be found in *Duden Online-Wörterbuch*, where a short text written on yellow background leads the user's attention to a usage note (also on yellow background) placed later in

[2] https://www.dictionary.com/articles/dictionary-labels-disparaging-offensive-informal?

the entry. In both cases the graphical warning leads to a text box with more information than what is customary to give in a dictionary entry.

As for the nature of the notes, they can in some cases be explicitly prescriptive and normative as can be seen in the following examples:

> Im modernen Sprachgebrauch sollte das Wort *Mädchen* nur noch in der Bedeutung *Kind weiblichen Geschlechts* verwendet werden. In den weiteren veraltenden oder veralteten Bedeutungen gilt die Bezeichnung *Mädchen* zunehmend als diskriminierend (In modern language use, the word *girl* should only be used in the sense *child of female gender*. In other senses that are becoming obsolete or are outdated, the term *girl* is increasingly considered discriminatory). (Duden: *Mädchen*)

A comparable approach can be observed in SAOL (The Swedish Academy Wordlist) in the entry *eskimå* (Eskimo), where the word is accompanied by a usage label: ⟨kan uppfattas som nedsätt.⟩ (may be perceived as derogatory) and a prescriptive suggestion: Använd hellre *inuit* (rather use *Inuit*).

In *Merriam-Webster*, the usage notes can be more subtle recommendations, close to description. This is for instance the case in the entry *Gypsy*, which has the label 'usually offensive' and a warning sign leading to a note explaining why the term is considered offensive and where a recommendation is given: 'Although *Gypsy* is sometimes used as a neutral or positive self-descriptor, it is recommended that those for whom it is not a self-descriptor use *Roma* or *Romani*/*Romany* instead.'

In the entry *Indian* in *Merriam-Webster* sense 2a ('indigenous American') is supplied with the label 'often offensive', and the usage note specifies that 'sense 2a of Indian is often considered offensive when used by a non-Indigenous person'. Later, the note gives a recommendation: 'the usual recommendation is to use the group name that is preferred by the group or member being identified'.

Inspired by *Merriam-Webster* and *Duden*, a similar solution in DDO would make it possible to combine concise labels with more nuanced explanations presented in dedicated text boxes. We also plan to place a clickable symbol next to certain controversial senses of words. The click leads to a text box placed at the end of the entry. In the first phase of the implementation, the box will be a general default text which states that the word can in certain contexts be inappropriate, offensive, derogatory or transgressive, which is why there is a good reason to pay extra attention when using this word. This statement is followed by a general statement that on the one hand it is not the intention of the dictionary to decide what language users do, and on the other hand it is the duty of the dictionary to inform that certain words can be perceived in a certain way by the recipient – irrespective of the sender's intention. In a second phase, the aim is to replace the general text by a text that is specific for the sense in question. The wording should be precise, not too verbose, and tailored to the specific sense or senses. The text could address diachronic changes in usage and connotation, disagreements about the status of a term, and the social and historical factors that render it problematic. Some text boxes would apply to broader categories of controversial terms, while others would be specific to individual entries. In many cases, the entry-specific texts will contain historical information, suggested alternatives, and links to external resources.

6 Delimiting Controversial Words

In order to determine which words can be considered controversial and thus be candidates for warning signs and usage notes, the editorial team used a method (developed by Sanni Nimb in

collaboration with Agnes Aggergaard Mikkelsen on the basis of *Den Danske Begrebsordbog* (The Danish Thesaurus) and described in Nimb (in press)). We shall not go into detail here but just indicate that it has provided the editorial team with evidence for assessing controversiality, also taking differences related to generation and gender into account. A relevant result, however, is that there is general agreement that two categories of words are particularly likely to be regarded as controversial and therefore potential candidates for warning signs. The most relevant cases include:

- words whose evaluative meaning is contested or has shifted over time (e.g. terms relating to gender, ethnicity, or disability), and
- words that may be perceived as discriminatory on the basis of identity categories.

Other categories – such as unequivocal insults or clearly evaluative terms with obvious alternatives – are not considered to require the same level of marking, as existing labels are deemed sufficient.

7 Discussion

The revision of labels for marked language in DDO raises fundamental questions about the function and societal role of dictionaries. Lexicographers editing descriptive dictionaries generally regard their role as that of faithfully representing language use as it can be observed, for example, in large language corpora, whereas their own attitudes and linguistic preferences are regarded as irrelevant and something that should be excluded from the lexicographical description. This role and professional self-understanding are challenged as it becomes increasingly difficult to maintain a position of neutral description. This is particularly true of identity-related terms that lie at the centre of debates concerning political correctness: members of the speech community disagree among themselves about which terms may appropriately be used about one another, and about what is offensive, derogatory, or even racist.

The difficulty lies in the fact that the perception of a given term depends upon the linguistic intuitions of the individual speaker. What for one language user constitutes neutral and unmarked language may for another be perceived as stereotypical and derogatory. How, then, can the lexicographer communicate this difference in a balanced and objective manner without appearing to promote one interpretation over another? One solution is the use of modifiers such as ‘sometimes’, ‘often’, and ‘especially’, which indicate that the characterisation is not absolute and that alternative perceptions are possible. But this is not always sufficient, and another solution, frequently employed in relation to identity terms, is to distinguish between sender and recipient. This is what is reflected in DDO through the introduction of a label such as *may be perceived as offensive*, possibly supplemented by an expanded usage note.

However, this solution has also been questioned, notably from the perspective of critical discourse studies. Kunkel & Neuhaus (2023, p. 36) cite Lobenstein-Reichmann (2022, p. 32) and Arndt (2022, p. 125) as arguing against it, maintaining that it constitutes a form of relativisation that gives the impression ‘als ob Diskriminierung eine Empfindungssache Einzelner wäre’ (‘as if discrimination were a matter of individual perception’). While it is true that discrimination is not merely a matter of individual perception, this does not, in our view, entail the opposite claim either: that discrimination is entirely an objective fact. Nor does it imply that particular words are discriminatory per se, and certainly not when they are presented out of context, as in dictionary descriptions.

This means that, at least for the time being, we maintain that this is the best way of accounting for the problematic nature of a word without becoming unduly normative. Telling users which words they may and may not use risks undermining the dictionary's scientific status and thereby opening up a discussion of its institutional position, which in itself entails a certain risk when funding depends on public grants. This latter point is mentioned only in passing, but it stands in partial contrast to *Svensk ordbok* ('Swedish Dictionary'), who appears to hold that it is possible to take a more definitive stance without undermining scientific credibility: 'The lexicographers in the project aim to develop a product that is as discrimination free as possible, and to construct a resource that correctly describes the Swedish language' (Petersson & Sköldberg, 2021, p. 385).

8 Conclusion

The revision of the usage-labelling system in *Den Danske Ordbog* reflects how contemporary lexicography is increasingly influenced by broader societal discussions concerning discrimination and linguistic responsibility. Traditional labels such as *informal* and *jocular* often fail to distinguish adequately between harmless colloquial language and terms that may have a stereotyping or excluding effect. At the same time, ad hoc explanatory comments risk producing inconsistent and overly verbose descriptions. The move towards a closed and more systematic inventory of labels is therefore intended to strengthen both editorial consistency and user comprehension.

The proposed system seeks to balance different lexicographic principles. Dictionaries must continue to document language as it is actually used, including controversial vocabulary, while users increasingly expect contextual information about possible social implications. The combination of a close set of evaluative labels and expandable usage notes represents an attempt to accommodate both needs without becoming overly prescriptive.

The method mentioned in section 6 demonstrates that perceptions of offensive or discriminatory language vary across speakers and social groups. Because the status of many identity-related terms remains contested, lexicographers cannot rely on fixed or universally accepted classifications. Formulations such as *may be perceived as offensive* therefore remain a practical compromise between description and social awareness.

Finally, it should be stressed that the revised model should be regarded not as a final solution, but as an adaptable framework. Both language and language attitudes will continue to evolve, and so the lexicographic practice must continue to reflect and adapt in order to navigate between scientific credibility and social relevance.

References

Arndt, S. (2022). *Rassistisches Erbe: Wie wir mit der kolonialen Vergangenheit unserer Sprache umgehen*. Berlin.

Brewer, C. (2016). Labelling and Metalanguage. In P. Durkin (Ed.), *The Oxford Handbook of Lexicography* (pp. 488–500). Oxford University Press.

DDO: *Den Danske Ordbog* (The Danish Dictionary). Retrieved May 18, 2026, from ordnet.dk/ddo

Duden Online-Wörterbuch. Retrieved May 18, 2026, from www.duden.de

Kunkel, M. & Neuhaus, L. (2023). Die Dynamik gesellschaftlicher Diskussionen zu diskriminierungsfreier Sprache und sprachlichem Rassismus und ihre Reflexion im Wörterbuch. In R. Bergmann,

K. Donhauser, N. Filatkina, E. Glaser, C. Moulin & T. Vennemann (Eds.), *Sprachwissenschaft*, Band 48, Heft 1 (pp. 29–48). Universitätsverlag Winter, Heidelberg.

Lobenstein-Reichmann, A. (2022). Sprachliche Ausgrenzung in der deutschsprachigen Lexikographie. *Lexicographica* 38(1), 7–38.

Merriam-Webster. Retrieved May 18, 2026, from www.merriam-webster.com/

Nimb, S. (in press). Mandfolkesnak, pigehvin og indianerhyl – metoder til fremme af ligestilling i Den Danske Ordbog, Den Danske Begrebsordbog og afledte sprogteknologiske ressourcer. *Nordiske Studier I Leksikografi*, 18 (forthcoming).

Nimb, S., Lorentzen, H., Theilgaard, L. & Troelsgaard, T. (2014). *Den Danske Begrebsordbog* (The Danish Thesaurus). Det Danske Sprog- og Litteraturselskab and Syddansk Universitetsforlag.

Petersson, S. & Sköldberg, E. (2021). To discriminate between discrimination and inclusion: a lexicographer's dilemma. In Z. Gavriilidou, M. Mitsiaki & A. Fliatouras (Eds.), *Lexicography for Inclusion: Proceedings of the 19th EURALEX International Congress*, Vol. 1 (pp. 381–386). Retrieved May 18, 2026, from https://euralex.org/elx_proceedings/Euralex2020-2021/EURALEX2020-2021_ProceedingsBook-Vol1.pdf

Retskrivningsordbogen ('The Spelling Dictionary'). Retrieved May 18, 2026, from ro.dsn.dk

SAOL: *Svenska Akademiens ordlista* (The Swedish Academy Wordlist). Retrieved May 18, 2026, from svenska.se

Svensén, B. (2009). *A Handbook of Lexicography: The Theory and Practice of Dictionary-Making*. Cambridge University Press.

SO: *Svensk ordbok* (Swedish Dictionary). Retrieved May 18, 2026, from svenska.se

Trap-Jensen, L. (2020). Inklusion eller mindretalsdiktatur? Om politisk korrekthed, minoritetshensyn og leksikografisk deskriptivisme i Den Danske Ordbog. In S. Berg-Olsen & A. Hannesdóttir (Eds.), *LexicoNordica* 27 (pp. 137–156). https://doi.org/10.7146/ln.vi27.123220

Trap-Jensen, L. & Lorentzen, H. (2026). Vulgært, stødende, negativt? – om markeret sprogbrug i Den Danske Ordbog. *LEDA-Nyt*, 81, 23–27. https://leksikografer.dk/ledanyt

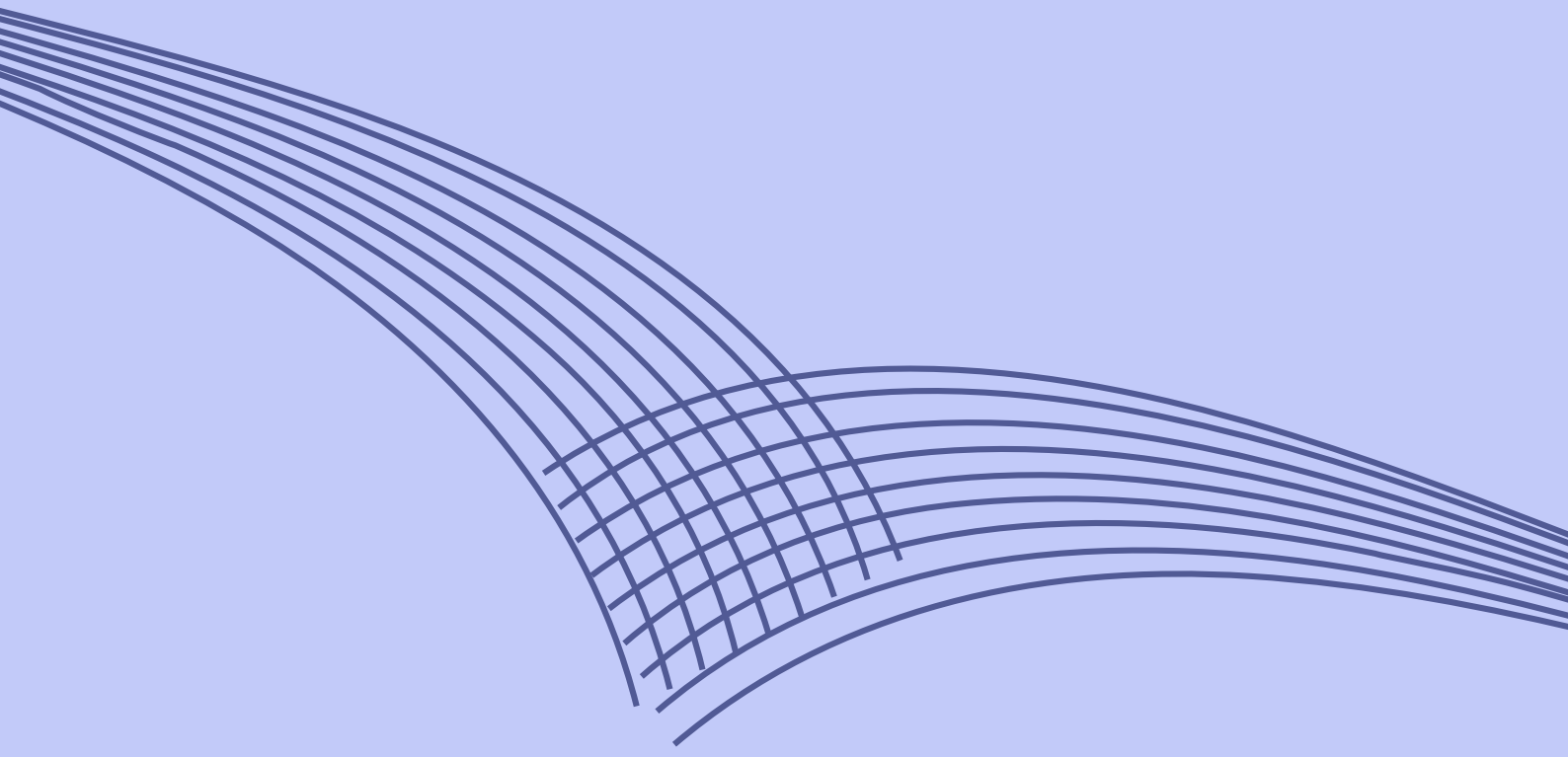
Contact Information

Lars Trap-Jensen
Society for Danish Language and Literature
ltj@dsl.dk

Henrik Lorentzen
Society for Danish Language and Literature
hl@dsl.dk

HISTORICAL & DIACHRONIC LEXICOGRAPHY

EURALEX 2026



Petra Bago & Ivana Eterović

COMPILING A 16TH-CENTURY CROATIAN DICTIONARY IN THE AGE OF AI

Challenges and Perspectives

Abstract This paper presents the theoretical and methodological foundations for an online dictionary of sixteenth-century Croatian. It utilizes a corpus of ten Gospel translations in Glagolitic, Cyrillic, and Latin scripts, encompassing missals, lectionaries, Bible translations, and postils. This work is a primary objective of the project *CroBiLex: Language Variation in Croatian Books of the Early Modern Period (Based on the Gospels Corpus)* (Croatian Science Foundation, IP-2025-02-2567; 2026–2029) at the University of Zagreb's Faculty of Humanities and Social Sciences. The paper outlines a conceptual lexicographic architecture for processing this linguistically heterogeneous, multi-script corpus. It emphasizes a Human-in-the-Loop approach where deterministic processing is complemented by machine-assisted transcription using handwritten text recognition and multimodal models. It also explores the experimental use of large language models (LLMs) for candidate lemma grouping, provisional semantic clustering, and preliminary entry-template preparation. Because highly variable historical texts pose severe risks of historical inaccuracy and hallucination, LLM outputs are treated strictly as non-authoritative proposals requiring expert verification. As CroBiLex is newly launched, this framework is conceptual rather than empirically evaluated. It seeks to balance computational assistance with scholarly accountability, providing a foundation for a sustainable digital model in historical Croatian lexicography.

Keywords Historical lexicography; Croatian language; sixteenth century; Gospels; multiscriptality; language variation; artificial intelligence

1 Introduction

Early 2026 marked the launch of *CroBiLex: Language Variation in Croatian Books of the Early Modern Period (Based on the Gospels Corpus)*. Funded by the Croatian Science Foundation (grant no. IP-2025-02-2567; 2026–2029), it is conducted at the University of Zagreb's Faculty of Humanities and Social Sciences. Bible translations, particularly the core Gospel texts, are central to pre-standard Croatian history due to unparalleled continuity, relative stability, and linguistic conservatism. These textual witnesses offer vital insights into evolving linguistic attitudes toward different varieties and conceptualizations of the Croatian literary language (Barbarić & Eterović, 2022).

The project's first year focuses on establishing theoretical and methodological foundations for an online Early Modern Croatian dictionary, using the Gospels as a representative corpus to develop a suitable lexicographic model. During years two and three, the lexical material of the four Gospels from all corpus texts will undergo comprehensive lexicographic processing. This follows established procedures: collecting and selecting items, normalizing headwords, defining and structuring meanings, selecting illustrative examples, idioms, and collocations, and establishing synonymic relations. All attested items will be recorded, with concordances evidencing usage and meaning. Dictionary entries will feature a headword, grammatical information, terminological labels, syntagmatic data, phraseological units, and etymological data. This groundwork will enable broader philological, textological, and historical-sociolinguistic research.

Since CroBiLex is in its initial design phase, this paper presents a conceptual and technical architecture rather than an implemented or empirically evaluated system. It focuses on

representing multiscript and linguistically variable sources, organizing lexicographic data, and using controlled computational assistance for pre-lexicographic tasks. The practical effectiveness, accuracy, and editorial impact of these components will be evaluated in later stages.

2 The CroBiLex Research Corpus

The research corpus comprises ten sixteenth-century Croatian Gospel translations written in all three Croatian historical scripts (Glagolitic, Cyrillic, and Latin), including four manuscripts and six printed texts. These translations belong to four distinct text types: lectionaries, missals, Bible translations, and postils (cf. Barbarić & Eterović, 2022), according to which they are presented in greater detail below.

In addition to these ten translations, the research corpus will also need to include, indirectly, one lectionary printed at the very end of the fifteenth century in the Latin script and composed in a South Chakavian dialect: *Bernardin's Lectionary* (Venice, 1495), to which the majority of the sources included in the corpus are closely related.

For digital processing, each witness will be represented in a structured source register containing a stable project identifier, bibliographic and repository metadata, script and document type, digitization status, available textual representations, and information on completeness and access conditions.

2.1. Lectionaries

2.1.1. Ranjina's Lectionary (around 1508)

Ranjina's Lectionary is a Shtokavian manuscript written in the Latin script around 1508 by the Dubrovnik nobleman Nikša Ranjina, who may have been only fourteen years old at the time he began writing the manuscript, as can be inferred from his personal notes recorded within the lectionary itself (Barbarić, 2017, p. 91). The original manuscript is preserved in the Archives of the Croatian Academy of Sciences and Arts under the shelf mark IV a 32 (Barbarić, 2021, p. 231). A Latin transcription of the manuscript was published at the end of the nineteenth century (Rešetar, 1894).

2.1.2. *Dubrovnik Dominican Lectionary* (around 1520)

The *Dubrovnik Dominican Lectionary* is a Cyrillic manuscript dating from approximately 1520 and originating from the Dubrovnik region. Previous scholarship has demonstrated that the manuscript represents a copy of the first edition of *Bernardin's Lectionary* (1495) (Rešetar, 1933). The source text was composed in the South Chakavian dialect, which was subsequently linguistically adapted to the Dubrovnik Shtokavian vernacular by the scribe of this lectionary (Kapetanović, 2014; Barbarić, 2017, pp. 96–101). The original manuscript is preserved in the library of the Monastery of St Dominic in Dubrovnik under the shelf mark 39 (Kapetanović, 2014, p. 63). To date, the manuscript has never been published in its entirety in any form.

2.1.3. *Zborovčić's Lectionary* (1543)

Zborovčić's Lectionary is a South Chakavian lectionary printed in Venice in 1543 in the Latin script. Prepared by Benedat (Benedikt) Zborovčić and dedicated to Don Jakov Miličić, this edition effectively constitutes the second edition of *Bernardin's Lectionary*, differing from the first

edition only in minor respects (Barbarić, 2017, p. 102). Several copies of the edition have been preserved and are currently held in the National and University Library in Zagreb (under the shelf marks RIIC–8°–77a and RIIC–8°–77b), the Franciscan Monastery of Our Lady of Mercy on Visovac Island, the British Museum, and collections in Odessa (Barbarić, 2017, pp. 103, 229). To date, the edition has not been published in any form; however, its divergences from the first edition are documented in the footnotes of the recent scholarly critical edition of the original *Bernardin's Lectionary* (Barbarić & Štrkalj Despot, 2020).

2.1.4. *Leipzig Lectionary* (the third quarter of the sixteenth century)

The *Leipzig Lectionary* is a Cyrillic manuscript originating from the Dubrovnik region and dating from the third quarter of the sixteenth century. Previous scholarship has established that the manuscript represents a copy of the second edition of *Bernardin's Lectionary*, namely *Zborovčić's Lectionary* (1543) (Rešetar, 1933). As in the case of the *Dubrovnik Dominican Lectionary*, the text underwent a linguistic adaptation whereby the (South) Chakavian dialect of the source text was adjusted to the Dubrovnik Shtokavian vernacular (Barbarić, 2017, pp. 104–109; Eterović & Barbarić, 2025). The original manuscript is preserved in the Leipzig University Library under the shelf mark Slav. 2 (Barbarić, 2017, p. 104). To date, fewer than 10% of its contents – namely, 35 out of 443 biblical readings – have been published in Latin transliteration (Leskien, 1881; cf. Eterović & Barbarić, 2025, p. 61), although the original manuscript is available in a facsimile edition (Schütz, 1963).

2.1.5. *Ortus anime* (1567)

Ortus anime is a Cyrillic manuscript written in 1567 by Marin Nikolić (also known as Marin Nikolin and Marinus Nicolai), a Dubrovnik merchant, during his stay in Belgrade, for another Dubrovnik merchant, Matej Džore Boždarović. Previous scholarship has demonstrated that the manuscript constitutes a copy of Nikola Dešić's edition *Raj duše*, translated from the Latin *Hortulus animae* and printed in Padua in 1560 in the Latin script. The two texts – *Ortus anime* and *Raj duše* – correspond closely in content, though not entirely on the linguistic level, since the source text underwent a degree of linguistic adaptation toward the Dubrovnik Shtokavian vernacular (Barbarić, 2021). It has further been established that Dešić undoubtedly relied on *Bernardin's Lectionary*, most probably its second edition, namely *Zborovčić's Lectionary* (1543), traces of which may also be identified indirectly in Nikolić's manuscript (Barbarić, 2021, p. 187). The original manuscript is preserved in the Archives of the Croatian Academy of Sciences and Arts under the shelf mark III a 25 (Barbarić, 2021, p. 184). To date, the manuscript has not been published in its entirety in any form, with the exception of a Latin transliteration of its first twenty pages and the comparison of selected biblical passages from *Bernardin's Lectionary*, Nikola Dešić's *Raj duše*, and Marin Nikolić's *Ortus anime* published in recent years (Barbarić, 2020; Barbarić, 2022).

2.2. Missals

2.2.1. Šimun Kožičić Benja's *Misal hrvacki* (1531)

Misal hrvacki [Croatian Missal] is one of the six editions printed in the Glagolitic script at the Rijeka Glagolitic printing house, which operated between 1530 and 1531 after being established by Šimun Kožičić Benja, Bishop of Modruš and Krbava. The Missal constituted the central and most significant publication of this printing enterprise. Comparative analysis with three other Croatian missals written and published in the Glagolitic script during approximately the same period – the Croatian *Editio Princeps* (1483), the *Senj Missal* (1494), and *Pavao Modrušanin's*

Missal (1528) – has demonstrated that its text diverges considerably from the earlier Croatian Church Slavonic tradition. These differences are the result of Kožičić's extensive editorial revision of the text in accordance with the Vulgate, making it the first Croatian missal to be systematically redacted on the basis of the Vulgate tradition (Žagar, 2015a, pp. 7, 14–21). Today, the edition is available in three forms: a Latin transliteration (Žagar, 2016; also available online[1], a reprint edition (Žagar, 2015b), and a digitized version on the website of the National and University Library in Zagreb, where the original is kept under the shelf mark RIIA–8°–8[2].

2.2.2. *Dubrovnik Missal* (sixteenth century)

The *Dubrovnik Missal* is a manuscript written in the Latin script, originating from the Dubrovnik region and dating from the late sixteenth century. The original manuscript is preserved in the Vatican Apostolic Library under the shelf mark *Neofiti 55*. In the scholarship to date, the abundance and complexity of its source models, as well as their interconnectedness with all three historical Croatian scripts, have been thoroughly examined (Grčević, 2021). Today, the manuscript is available in three forms: a Latin transcription (Giannelli & Graciotti, 2003), a facsimile edition (Grčević, 2021), and a digitized version accessible through the website of the Vatican Library[3].

2.3. Bible translations

2.3.1. *New Testament* (1562/1563)

The *New Testament* represents the central, most significant, and most extensive publication produced by the Croatian Protestant printing house in Urach, which operated between 1561 and 1564. It is the earliest complete Croatian translation of the *New Testament* known to date and was published in two scripts: Glagolitic (1562/1563) and Cyrillic (1563). Although neither the accompanying preface by Stipan Konzul and Antun Dalmatin nor Primož Trubar's preface to the edition explicitly identify the source text used by the Croatian Protestants in preparing their translation, subsequent scholarship has established that one of their principal sources was undoubtedly *Zborovčić's Lectionary*. The translators sought, to varying degrees, to adapt its southern Chakavian linguistic features to the northern Chakavian varieties characteristic of the regions from which most of them originated (cf. Barbarić & Eterović, 2020). Today, the edition is available in three forms: a Latin transliteration (Matak, 2013; Matak 2015), a reprint edition (Jembrih, 2007), and a digitized version accessible through different websites – for example, the first part of the *New Testament* through the website of the Digital Library of Slovenia[4] and the second part through the website of the Munich DigitiZation Center[5].

2.4. Postils

2.4.1. Croatian Protestant *Postila* in the Latin script (1568)

The Croatian Protestant *Postila* in the Latin script represents one of the three extant versions of the same text, published respectively in the Glagolitic (1562), Cyrillic (1563), and Latin (1568) scripts. The first two editions were printed at the Croatian Protestant press in Urach, whereas

1 https://glagoljica.hr/?eizvori=i&id=553915
2 https://digitalna.nsk.hr/?pr=i&id=15235
3 https://digi.vatlib.it/view/MSS_Neofiti.55
4 https://www.dlib.si/details/URN:NBN:SI:DOC-W303QB0B?&language=eng
5 https://www.digitale-sammlungen.de/en/view/bsb00078763?page=4,5

the last edition was published in Regensburg. Previous scholarship has demonstrated that the biblical passages in the Latin-script *Postila*, as in the two earlier editions, are based on the Croatian Protestant translation of the *New Testament*; however, they also exhibit a degree of editorial revision. For instance, in the 1568 *Postila*, the grapheme representing the sound *jat* or its reflex *e* were systematically replaced by the Ikavian reflex, while additional modifications have been identified at other linguistic levels as well (cf. Kuštović & Žagar, 2020, pp. 268–277, 303–316). Today, the edition is available both in a reprint edition (Jembrih, 1993) and in a digitized version accessible as an on-demand e-book through the website of the University Library of Regensburg.

2.4.2. *Postila* by Antun Vramec (1586)

In the sixteenth century, the earliest Croatian Gospel texts written in the Kajkavian variety which are presently known were edited by Antun Vramec in his *Postila* and printed in the Latin script in Varaždin in 1586. Previous scholarship has demonstrated that Vramec undoubtedly had access to *Bernardin's Lectionary* or to one of its later derivatives, most probably the Croatian Protestant *Postila* in the Latin script. The source text was adapted at various linguistic levels in order to better accommodate the needs of the Kajkavian-speaking community, primarily at the phonological and lexical levels, often through the replacement of words perceived as less comprehensible or entirely unfamiliar (Jagić, 1913; Murko, 1925). Today, the edition is available only in a reprint edition (Jembrih, 1990).

The CroBiLex sources vary significantly in the availability of digital and scholarly representations. Table 1 distinguishes digitized images (digital surrogates of originals/printed copies) from facsimile and reprint editions, which offer varying access to documentary features. It separately tracks diplomatic transcriptions, transliterations, and linguistic transcriptions, noting if coverage is complete or "partial" (limited to selected pages, passages, or readings).

As Table 1 shows, the corpus lacks a uniform technical starting point. Some witnesses offer digitized originals and complete scholarly transcriptions or transliterations, while others are limited to facsimiles, reprints, partial editions, or unpublished manuscripts. Therefore, the planned digital architecture must accommodate incomplete, incrementally developed textual representations instead of assuming uniform processing levels. When transcriptions are absent, producing an initial machine-readable text is a lexicographic prerequisite; the potential role of assisted transcription here is discussed in Section 4.2.

While Glagolitic and Cyrillic texts can be prepared in diplomatic transcription, Croatian philological practice traditionally favors Latin transliteration and, less frequently, Latin transcription. Conversely, pre-standard Croatian texts in non-standardized Latin script are typically published in diplomatic and Latin transcription. This creates a disparity in presenting sixteenth-century texts: Glagolitic and Cyrillic utilize diplomatic transcription, transliteration, and linguistic transcription, whereas Latin-script texts use only diplomatic transcription and linguistic transcription.

Accordingly, the principal technical challenge during dictionary preparation is organizing multiple distinct, mutually aligned representations rather than producing a single uniform corpus. Depending on the script and editorial work, layers may include a digitized source image, diplomatic transcription, transliteration, linguistic transcription, and a normalized representation for retrieval and analysis. Because each layer serves unique scholarly and computational purposes, they must remain separately identifiable, ensuring normalized forms supplement rather than replace historically attested forms.

TABLE 1 Current availability of the CroBiLex corpus sources and their textual representations

Source	Script	Medium	Digitized source images	Facsimile edition	Reprint edition	Diplomatic transcription	Transliteration	Linguistic transcription	Coverage
Ranjina's Lectionary	Latin	manuscript	No	No	N/A	No	No	Yes	Complete
Dubrovnik Dominican Lectionary	Cyrillic	manuscript	No	No	N/A	No	No	No	Complete
Zborovčić's Lectionary	Latin	printed book	No	N/A	No	No	No	Yes	Partial
Leipzig Lectionary	Cyrillic	manuscript	No	Yes	N/A	No	Yes	No	Partial
Ortus anime	Cyrillic	manuscript	No	No	N/A	No	Yes	No	Partial
Misal hrvacki	Glagolitic	printed book	Yes	N/A	Yes	No	Yes	No	Complete
Dubrovnik Missal	Latin	manuscript	Yes	Yes	N/A	No	No	Yes	Complete
New Testament	Glagolitic	printed book	Yes	N/A	Yes	No	Yes	No	Complete
Croatian Protestant Postila in the Latin script	Latin	printed book	Yes	N/A	Yes	No	No	No	Complete
Postila by Antun Vramec	Latin	printed book	No	N/A	Yes	No	No	No	Complete

Note. Digitized source images are digital surrogates of original witnesses or printed copies. Facsimile and reprint editions are recorded separately because they represent distinct publication forms offering varying access to documentary features. Diplomatic transcription, transliteration, and linguistic transcription are treated as separate textual representations, while "partial" indicates that only selected pages, passages, or readings are available.

The planned data model must link every derived textual unit to its corresponding source witness and, ideally, to a specific page, image region, passage, or verse. Every source requires a stable project identifier and structured metadata covering its script, document type, repository, shelf mark, digitization status, textual representations, completeness, and access conditions. Given the varying degrees of digitization and transcription, the architecture must support incrementally adding and revising textual layers without overwriting earlier representations.

Furthermore, eight of the ten corpus sources share a demonstrably close relationship with *Bernardin's Lectionary: Ranjina's Lectionary*, the *Dubrovnik Dominican Lectionary, Zborovčić's Lectionary*, the *Leipzig Lectionary, Ortus anime*, the *New Testament*, the Croatian Protestant *Postila* in the Latin script, and the *Postila* by Antun Vramec. Because these editions differ in script and linguistic variety, the digital resource must support parallel consultation of corresponding biblical passages. This requires passage- or verse-level alignment (e.g., John 1:1–10) allowing users to compare textual realizations while retaining direct access to each form in its source-specific representation and context.

3 The CroBiLex Dictionary Entry Architecture

Modeling the CroBiLex digital dictionary requires an entry architecture capable of representing substantial graphic and morphological variation without collapsing historically significant distinctions. The project corpus unites sixteenth-century Croatian Gospel texts from four text types (missals, lectionaries, Bible translations, and postils) written in three historical scripts (Barbarić & Eterović, 2022). Representing these heterogeneous sources in a single electronic resource creates structural challenges for organizing textual representations, lemma identification, grammatical analysis, and cross-witness retrieval (Eterović, 2017).

3.1. Script-Agnostic Lemmatization and Multi-Script Harmonization

A primary challenge in designing the CroBiLex architecture is the sixteenth-century Gospel corpus's substantial orthographic variation and multiscript distribution. The Glagolitic, Cyrillic, and Latin witnesses exhibit graphic, phonological, morphological, syntactic, and lexical variations limiting direct computational lemmatization reliability (Piotrowski, 2012). Instead of imposing an early standardized token layer, the proposed architecture maintains distinct but linked representations. An abstract, script-agnostic lemma entity links to diplomatic, transliterated, transcribed, and normalized forms, while every attestation retains its source-specific spelling, script, witness, location, and editorial status. This script-agnosticity supports cross-script retrieval without suppressing individual witnesses' documentary characteristics.

The practical complexity of this mapping process is illustrated by the range of surface orthographic doublets and triplets across the project's test corpus. For instance, the immutable ritual particle *amen* appears in diplomatic transcriptions under complex graphic representations: *am(e)n, amen, amen', am(e)nь, am(en)ь, amenь, and amênь*. Similarly, morphologically variable lexical items, like the frequent verb *vidêti*, introduce extensive morphosyntactic, thematic, and phonetic paradigms across missals, lectionaries, and postils. For example, the *Gospel of Matthew* from the *New Testament* (1562) attests these forms of *vidêti* 'to see': infinitive *videti*; present tense, 3rd person plural *vide*; first sigmatic aorist, 3rd person singular *vide*; present active participle, nominative singular masculine *videći*; present active participle, nominative singular neuter *videći*; present active participle, nominative plural masculine *videći, videć*; second preterite active participle, nominative plural masculine *videli*; pluperfect, 3rd person plural *bihu videli*; periphrastic future, 2nd person singular *budeš videti*; and future perfect, 2nd person plural *budete videli*. As these examples demonstrate, the same form may be morphologically

ambiguous (*videći*), and the same grammatical category may be realized by multiple inflectional variants (as in the present active participle, nominative plural masculine). Standardizing these variations into a single rigid token layer too early in the processing pipeline would introduce a risk of representational bias and obscure the underlying textual transmission patterns (Bollmann, 2019).

Table 2 illustrates how a single attested form could be represented in the proposed architecture. The example describes intended data relations rather than the output of an implemented system.

TABLE 2 Illustrative processing of an attested form in the proposed CroBiLex architecture

Processing stage	Illustrative content	Status/authority
Source evidence	Source image, witness, passage, attested form *videći*	Attested
Textual representation	Diplomatic/transliterated form and token identifier	Human-validated representation
Conventional processing	Candidate set generated through script mapping, string comparison, and morphological heuristics	Machine-assisted
LLM proposal	Possible link to *vidêti* and proposed grammatical analysis, with supporting token/context identifiers	Unvalidated proposal
Editorial decision	Accept, modify, split, or reject, with corrected analysis	Human-validated
Dictionary result	Validated lemma link, grammatical information, citation, and source reference	Publishable content

This example illustrates one lexical item's intended progression, not the output of an implemented or evaluated CroBiLex system. The architecture treats the headword layer as an independent, script-agnostic entity dynamically linked to multi-tiered surface realizations. Each entry separates the abstract lemma from graphic variants, grammatical annotations, usage or variety labels, and source-specific attestations. This accommodates Early Modern Croatian orthographic variation without hindering cross-resource search or indexing (Eterović, 2017). This data model supports downstream applications (like automated candidate grouping and concordance extraction) making them less dependent on physical manuscript or print features, thereby preserving empirical transparency.

3.2. Parallel Modern-Text Alignment Layer

To support comparative consultation of heterogeneous historical witnesses, the CroBiLex architecture includes a modern Croatian parallel-text layer aligned at biblical passage and verse levels. Its navigational purpose provides a shared reference to locate and compare corresponding passages across witnesses differing in script, orthography, linguistic variety, syntax, and lexical choice.

Primary alignment is at the passage or verse level. While reliable granular phrase or token correspondences could be recorded, they should not be presumed complete due to differences in translation strategy, transmission, word order, and lexical selection. The modern layer aids retrieval and comparison, not acting as evidence for a historical attestation's lemma identity, grammatical analysis, meaning, or sense structure. Interpretive decisions rely strictly on historical witnesses, concordances, and expert philological analysis.

This layer digitally serves two functions. First, it enables cross-witness retrieval from a modern passage to historical sources. For example, consulting John 1:1–10 retrieves aligned missals, lectionaries, Bible translations, and postils, while retaining access to each witness's diplomatic, transliterated, transcribed, or normalized representations. This provides shared navigation without replacing source-specific forms.

Second, it facilitates separately encoded links between validated historical lemmas and contemporary resources (e.g., Hrvatski jezični portal[6] or phraseological databases) when editors establish defensible modern equivalents. These support diachronic comparison, discovery, and digital lexicographic integration (Ahačič, Ledinek & Perdih, 2015). They must remain distinct from the historical lemma and entry evidence, since not all historical items have direct modern equivalents.

Separating passage alignment, historical evidence, and contemporary links supports comparative research while preserving corpus autonomy. This layered approach facilitates future interoperability and data exchange, provided historical-modern data relationships are explicitly documented (Salgado et al., 2019).

3.3. Internal Editing Model and TEI Lex-0 Export

Designing the CroBiLex data model balances detailed representation of historical and grammatical data during lexicographic editing with interoperability for publication and exchange (Salgado et al., 2019). The internal model must accommodate sixteenth-century Croatian morphological complexity and preserve links between abstract structures and source attestations. Nouns require representing seven cases across singular, dual, and plural forms. Verbs require recording historical tenses, aspect, and gender-number-case distinctions in participial paradigms. The accuracy and scope of the grammatical categories will be determined through philological analysis.

Representing these structures directly in deeply nested Text Encoding Initiative (TEI) markup would create substantial editorial and technical overhead. Thus, the project separates the internal editing model, editorial interface, and publication format. The internal model structures project-specific grammatical categories, variant forms, attestations, citations, and analyses for historical Croatian lexicography. The interface allows lexicographers to manage this data without interacting with underlying serialization syntax. Precise database technology and serialization formats remain future implementation decisions.

For publication, validated content maps to TEI Lex-0 compatible structures using elements like `<entry>`, `<form>`, `<gramGrp>`, `<sense>`, and `<cit>`, supporting established lexicographic encoding practices (Salgado et al., 2019). This internal-to-TEI mapping requires explicit documentation and testing. Project-specific distinctions lacking TEI Lex-0 equivalents may need documented extensions, feature structures, stand-off annotations, or linked supplementary data. Therefore, the export process cannot be assumed entirely lossless until evaluated against representative dictionary entries.

Automated transformation and schema validation can verify structural conformity to the encoding profile, but cannot establish linguistic accuracy, completeness, or interpretative adequacy. These remain subject to expert editorial review. Separating active editing from standardized dissemination supports detailed philological work while improving data exchange, external

[6] https://hjp.znanje.hr/

linking, long-term maintenance, and reuse in broader digital lexicographic environments (Ahačič, Ledinek & Perdih, 2015).

4 LLM-Assisted Pre-Lexicographic Work: Scope, Uses, and Editorial Control

This section outlines potential LLM uses, not an implemented pipeline. Deterministic rules remain standard for script conversion, concordance generation, identifier management, and schema validation. LLMs are optional proposal generators for tasks requiring supplementation, like difficult page-image transcription or context-sensitive lemma grouping. Their suitability, accuracy, and editorial impact await evaluation (Lew, 2024).

4.1. Replaceable LLM Components and Structured Output

To avoid dependency on a single AI platform, CroBiLex uses a replaceable interface. This permits testing various commercial or open-source models without altering the dictionary data model or editorial interface. Since replaceability cannot guarantee reproducibility across different configurations, any model's identity and version must be recorded whenever its output is retained (Lew, 2024).

The interface supplies models with controlled, structured input (e.g., token identifiers, transliterated forms, metadata, passages, concordances, and validated annotations) instead of unrestricted corpus access. Models must return proposals in predefined machine-readable structures, such as JSON or XML schemas. Required fields include affected attestation identifiers, the proposed grouping or analysis, and an explicit indication of their unvalidated status.

Schema validation determines whether returned data contain required fields, identifiers, and value types. Non-conforming outputs are rejected or returned for correction. Valid responses convert to a common project representation, ensuring technical consistency and facilitating inspection, comparison, and replacement within a Human-in-the-Loop workflow (Wu et al., 2022).

Crucially, structural conformity does not ensure linguistic or historical validity. Schemas verify a field's presence but cannot confirm the correctness of a lemma, grammatical analysis, usage grouping, sense distinction, or definition based on historical evidence. Structured output aids traceability but cannot prevent hallucinations or unsupported inferences; therefore, all AI-generated content remains provisional and requires expert philological review (Lew, 2024).

4.2. Machine-Assisted Transcription of Multiscript Witnesses

As Table 1 indicates, CroBiLex sources vary significantly in preparation; for those existing solely as images or manuscripts, generating a machine-readable text is a prerequisite for lexicographic work. Generic OCR systems may prove unsuitable and must be evaluated separately by script, document type, and condition. The corpus combines Glagolitic and Cyrillic materials with non-standardized Latin, and existing models—especially for sixteenth-century Croatian Glagolitic—lack sufficient training data (Piotrowski, 2012).

Handwritten text recognition (HTR) and multimodal language models offer potential preliminary transcriptions for unedited texts. Within the architecture, these tools act as replaceable,

task-restricted components returning structured output (witness, page/folio, and line/region). This creates a distinct, explicitly labeled representation layer, separate from editor-prepared transcriptions and transliterations.

Two conditions govern this automated assistance. First, machine transcriptions are treated as drafts requiring line-by-line verification against the accessible source image before serving as dictionary evidence. Second, risks are highly script- and text-specific. Generative models are prone to silently regularizing features that carry vital philological information, such as abbreviation and suspension marks, superscript letters, ligatures, the *jer* sign, and damaged passages. Consequently, segments with uncertain recognition quality must be explicitly flagged.

Testing on selected passages will establish whether this assistance genuinely expedites transcription and for which scripts it is viable. Should recognition quality prove insufficient, sources will be prepared via conventional editorial means, as the proposed architecture does not depend on the success of automated transcription.

4.3. Candidate Lemma Grouping and Provisional Usage Clustering

Candidate lemma grouping and usage clustering combine computational procedures with optional LLM proposals. LLMs do not replace deterministic or rule-based analysis; forms are initially collected via token aggregation, script conversion, morphological heuristics, and concordance generation. LLMs then serve as secondary components to rank links, identify similarities, or flag ambiguous forms.

Models receive controlled evidence packages (identifiers, forms, metadata, contexts, annotations) instead of unrestricted tokens. Using this data, models propose lemma realizations (e.g., variants of *vidêti*), generating non-authoritative hypotheses rather than final lemmatization decisions.

Proposed groups must preserve links to supporting attestations, identifying grouped instances, candidate lemmas, and incompatible forms. Editors accept, modify, split, merge, or reject these proposals. This organizes material computationally while reserving the final philological interpretation of lemma identity for expert editors.

For provisional concordance clustering, LLMs identify recurring contextual patterns in retrieved attestations. This helps editors navigate occurrences and compare usage across witnesses. These clusters represent provisional evidentiary arrangements linked to source lines, not established dictionary senses.

Generated labels and descriptions must rely exclusively on supplied evidence, without introducing unattested examples or unsupported historical meanings. Outlier lines must remain visible to the editor, as they may indicate incorrect groupings, distinct uses, textual peculiarities, or processing errors.

LLMs may also populate provisional entry templates with source-linked forms and fields. This serves strictly as a workspace; headwords, grammar, senses, and definitions demand explicit expert validation, treating the model as an organizational aid, not an autonomous lexicographer (Beliga & Filipović Petrović, 2024).

The goal is structuring evidence for expert review, not automatic compilation. Whether LLM assistance improves consistency or efficiency over manual procedures remains an empirical question to be examined (Lew, 2024).

4.4 *Separating Machine Proposals from Validated Dictionary Content*

Scholarly accountability requires distinguishing attested source evidence, provisional machine-generated proposals, and human-validated editorial content within the data model. Machine proposals must not auto-publish or overwrite source data. They must maintain links to original attestations and carry explicit statuses (e.g., proposed, accepted, modified, rejected, superseded). Published content is strictly human-validated, while original evidence remains independently accessible.

Retained machine proposals require technical metadata, including proposal and token identifiers, model identity, generation date, and editorial status, alongside optional model versions and input configurations. This metadata supports future review and output comparison but does not guarantee reliability or reproducibility.

For interpretively complex cases, like uncertain readings or competing lemma links, the system should record paradata explaining editorial decisions (Dillen, 2024). While routine corrections need only a status change, complex interventions require detailed rationales.

These records could map to provenance frameworks like W3C PROV for data exchange and research auditing (Groth & Moreau, 2013). While not all internal actions must be public, user-facing resources must distinguish source attestations from editorial information, enabling users to inspect the supporting evidence.

This separation does not prove computational assistance improves dictionary compilation speed or accuracy. Rather, it prevents unvalidated suggestions from becoming scholarly data, ensuring expert editors retain full responsibility and LLM outputs remain inspectable aids rather than independent lexical authorities (Lew, 2024).

5 Conclusion

This paper outlines theoretical and methodological foundations for the CroBiLex online dictionary of sixteenth-century Croatian, utilizing a multi-script, linguistically diverse Gospel corpus. The proposed architecture addresses orthographic and linguistic variation via a script-agnostic lemma layer dynamically linked to source-specific attestations. Abstract lemmas remain strictly separate from diplomatic, transliterated, transcribed, and normalized forms. A parallel modern-text layer facilitates passage-level navigation and comparison without substituting the historical evidence required for lemma identity or grammatical analysis.

The data model intentionally separates active editing from publication and data exchange. An internal editing model accommodates complex historical grammar, variant forms, and editorial analysis, while validated content maps to TEI Lex-0 structures for interoperability. Project-specific distinctions lacking direct TEI Lex-0 equivalents require documented extensions, feature structures, stand-off annotations, or linked supplementary data, demanding evaluation against representative entries prior to publication.

Additionally, the architecture accommodates replaceable AI components within a Human-in-the-Loop environment. Deterministic and rule-based procedures remain primary, while LLMs act as optional proposal generators for unedited source transcription, candidate lemma grouping, usage clustering, and entry-template preparation. Structured input and output aid integration, but since schema validation cannot guarantee linguistic or historical correctness, machine-generated content must remain clearly distinguishable from human-validated editorial content and attested source evidence.

This framework is currently a conceptual design, not an implemented operational model. No claims are made regarding LLM accuracy, alignment with expert analysis, or editorial impact. Future stages must evaluate LLM-assisted procedures against rule-based, corpus-based, and manual alternatives, assessing both output accuracy and necessary expert correction. Subject to evaluation, this combination of philological methods, layered architecture, and controlled AI assistance may serve as a valuable reference for related historical-lexicographic projects handling heterogeneous, under-resourced texts.

References

Ahačič, K., Ledinek, N. & Perdih, A. (2015). Fran: The next generation Slovenian dictionary portal. In *Proceedings of the 8th International Natural Language Processing, Corpus Linguistics, Lexicography Conference* (pp. 9–16).

Barbarić, V.-T. (2017). *Nastajanje i jezično oblikovanje hrvatskih lekcionara*. Institut za hrvatski jezik i jezikoslovlje.

Barbarić, V.-T. (2020). *Ortus anime* (1567.). Retrieved May 17, 2026, from https://www.researchgate.net/publication/344826590_Ortus_anime_1567

Barbarić, V.-T. (2021). Nekoliko riječi o ćiriličnome rukopisu Ortus anime iz 1567. godine, prijepisu Dešićeva Raja duše: što o njemu znamo i kako dalje. *Croatica: časopis za hrvatski jezik, književnost i kulturu*, 45(65), 183–195. Retrieved May 17, 2026, from https://hrcak.srce.hr/en/file/395068

Barbarić, V.-T. (2022). *Paralelni tekst/korpus biblijskoga teksta u Bernardinovu lekcionaru (1495.), Raju duše (1560.) i Ortus anime (1567.)*. Retrieved May 17, 2026, from https://www.researchgate.net/publication/358769807_Paralelni_tekstkorpus_biblijskoga_teksta_u_Bernardinovu_lekcionaru_1495_Raju_duse_1560_i_Ortus_anime_1567

Barbarić, V.-T. & Eterović, I. (2020). O utjecaju lekcionara na hrvatski protestantski prijevod Novoga zavjeta. In T. Barščevski, Z. Jelaska & N. Babić (Eds.), *Bibliana – Hrvatski prijevodi Biblije* (pp. 177–196). Hrvatsko filološko društvo, Katolički bogoslovni fakultet.

Barbarić, V.-T. & Eterović, I. (2022). Croatian Biblical Texts in the Early Modern Period: A Historical-Sociolinguistic Approach to Language Change. In K. Bennett & A. Cattaneo (Eds.), *Language Dynamics in the Early Modern Period* (pp. 40–52). Routledge.

Barbarić, V.-T. & Štrkalj Despot, K. (Eds.) (2020). *Bernardinov lekcionar 1495*. Institut za hrvatski jezik i jezikoslovlje.

Beliga, S. & Filipović Petrović, I. (2024). Large Language Models Supporting Lexicography: Conceptual Organization of Croatian Idioms. In Š. Arhar Holdt & T. Erjavec (Eds.), *Proceedings of the Conference on Language Technologies and Digital Humanities* (pp. 23–46).

Bollmann, M. (2019). A large-scale comparison of historical text normalization systems. In J. Burstein, C. Doran & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 3885–3898).

Dillen, W. (2024). Paradata for Digitization Processes and Digital Scholarly Editions. In I. Huvila, L. Andersson & O. Sköld (Eds.), *Perspectives on Paradata: Research and Practice of Documenting Process Knowledge* (pp. 147–163). Springer.

Eterović, I. (2017). O pojedinim problemima izrade rječnika hrvatskoga jezika 16. stoljeća. *Filologija*, 69, 23–41. https://doi.org/10.21857/y6zolb8j3m

Eterović, I. & Barbarić,V.-T. (2025). Early Modern Croatian Cyrillic Lectionaries as Mirrors of Dialect Perception: The Example of the Leipzig Lectionary. In A. Granberg, G. R. Parpulov & A. Radošević (Eds.), *Cyrillic Manuscripts: Script and Language, Scribes and Collections* (pp. 55–78). Peter Lang.

Giannelli, C. & Graciotti, S. (2003). *Il Messale croato-raguseo (Neofiti 55) della Biblioteca Apostolica Vaticana*. Biblioteca Apostolica Vaticana.

Grčević, M. (2021). *Das kroatische volkssprachliche Missale Romanum des XVI. Jahrhunderts: Philologisch-linguistische Untersuchung mit photomechanischer Reproduktion der Handschrift "Neofiti 55". Vol. 1–2*. Studia Croatica – Universitas studiorum Zagrabiensis.

Groth, P., & Moreau, L. (Eds). (2013). *PROV-overview. An overview of the PROV family of documents*. Retrieved from https://www.w3.org/TR/prov-overview/

Jagić, V. (1913). Die serbokroatischen Übersetzungen der Bibel im Ganzen oder einzelner Teile derselben. *Archiv für slavische Philologie*, 34, 497–532.

Jembrih, A. (Eds.) (1990). *Antun Vramec, Postilla, Varaždin 1586*. Pretisak. Zavod za znanstveni rad JAZU u Varaždinu, Kršćanska sadašnjost.

Jembrih, A. (Ed.) (1993). *Antun Dalmatin i Stipan Konzul, Postila 1568*. Pretisak. Istarsko književno društvo "Juraj Dobrila", Katedra Čakavskog sabora Buzet, Hrvatsko kulturno društvo u Gradišću, Hrvatski akademski klub.

Jembrih, A. (Ed.) (2007). *Novi testament 1562./1563*. Pretisak. Teološki fakultet "Matija Vlačić Ilirik".

Kapetanović, A. (2014). Jezična uporaba u ćiriličnom Dubrovačkom dominikanskom lekcionaru. *Filologija*, 62, 61–68. Retrieved May 17, 2026, from https://hrcak.srce.hr/138391

Kuštović, T. & Žagar, M. (Eds.) (2020*). Stumačeno pravo i razumno: Studije o jeziku knjiga hrvatskih protestanata 16. stoljeća*. Hrvatska sveučilišna naklada, Adventističko teološko visoko učilište.

Leskien, A. (1881). Das dalmatinisch-serbische cyrillische Missale Romanum der Leipziger Stadtbibliothek. *Berichte über die Verhandlungen der Königlich-Sächsischen Gesellschaft der Wissenschaften zu Leipzig: Philologisch-historische Klasse*, 33 (I–II), 199–250.

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, 11(1), 1–8.

Matak, D. (Ed.) (2013). *Anton Dalmatin – Stipan Konzul, Novi testament, 1. dio (1562.): Latinički prijepis glagoljskog izvornika*. Adventističko teološko visoko učilište, Filozofski fakultet u Zagrebu, Školska knjiga.

Matak, D. (Ed.) (2015). *Anton Dalmatin – Stipan Konzul, Novi testament, 2. dio (1563.): Latinički prijepis glagoljskog izvornika*. Adventističko teološko visoko učilište, Filozofski fakultet u Zagrebu, Školska knjiga.

Misal hrvacki: *Digitized version of Šimun Kožičić Benja's Misal hrvacki*. Retrieved May 19, 2026, from https://digi.vatlib.it/view/MSS_Neofiti.55https://digitalna.nsk.hr/?pr=i&id=15235

Murko, M. (1925). Nekoliko reči o jeziku srpsko-hrvatskih protestantskih knjiga. *Daničićev zbornik*, 72–106.

Neofiti 55: *Digitized version of the manuscript Neofiti 55*. Retrieved May 19, 2026, from https://digi.vatlib.it/view/MSS_Neofiti.55

New Testament (1562): the first part. Retrieved May 25, 2026, from https://www.dlib.si/details/URN:NBN:SI:DOC-W303QB0B?&language=eng

New Testament (1563): the second part. Retrieved May 25, 2026, from https://www.digitale-sammlungen.de/en/view/bsb00078763?page=4,5

Piotrowski, M. (2012). *Natural language processing for historical texts*. Morgan & Claypool Publishers.

Rešetar, M. (1894). *Zadarski i Rańinin lekcionar*. JAZU.

Rešetar, M. (1933). *Bernardinov lekcionar i njegovi dubrovački prepisi*. Srpska kraljevska akademija.

Salgado, A., Costa, R., Tasovac, T. & Simões, A. (2019). TEI Lex-0 In Action: Improving the Encoding of the Dictionary of the Academia das Ciências de Lisboa. In I. Kosem & S. Krek (Eds.), *Electronic Lexicography in the 21st Century: Smart Lexicography. Proceedings of the eLex 2019 Conference* (pp. 417–433).

Schütz, J. (1963). *Das handschriftliche Missale illyricum cyrillicum Lipsiense. Zweiter Band: Phototechnische Reproduktion des Kodex*. Otto Harrassowitz.

Wu, X., Xiao, L., Sun, Y., Zhang, J., Ma, T. & He, L. (2022). A survey of human-in-the-loop for machine learning. *Future Generation Computer Systems*, 135, 364–381.

Žagar, M. (Ed.) (2015a). *Jezik Misala hruackoga: Studije o jeziku Misala Šimuna Kožičića Benje (1531.)*. Hrvatska sveučilišna naklada.

Žagar, M. (Ed.). (2015b). *Misal hruacki Šimuna Kožičića Benje, Rijeka 1531.: Faksimilno izdanje*. Nacionalna i sveučilišna knjižnica u Zagrebu, Sveučilišna knjižnica Rijeka.

Žagar, M. (Ed.) (2016). *Misal hruacki Šimuna Kožičića Benje: Latinička transliteracija glagoljskog izvornika s usporedbom jezičnih oblika prema drugim glagoljskim tiskanim misalima XV. i XVI. stoljeća*. Nacionalna i sveučilišna knjižnica u Zagrebu, Sveučilišna knjižnica Rijeka. (Also available online; retrieved May 19, 2026, from https://glagoljica.hr/?eizvori=i&id=553915)

Contact information

Petra Bago
University of Zagreb Faculty of Humanities and Social Sciences
pbago@ffzg.hr

Ivana Eterović
University of Zagreb Faculty of Humanities and Social Sciences
isankovi@ffzg.hr

Acknowledgements

This paper was produced with support from the Croatian Science Foundation under the project *Language Variation in Croatian Books of the Early Modern Period (Based on the Gospels Corpus)*, grant number IP-2025-02-2567 (2026–2029).

Harald Bichlmeier

VOL. IX OF THE *ETYMOLOGICAL DICTIONARY OF OLD HIGH GERMAN [ETYMOLOGISCHES WÖRTERBUCH DES ALTHOCHDEUTSCHEN (EWA)]*

History Structure, Contents

Abstract The paper presents vol. IX of the *Etymological Dictionary of Old High German [Etymologisches Wörterbuch des Althochdeutschen]*, the most encompassing etymological dictionary of any (Old) Germanic language. The volume will have roughly 800 pages, contain a total of 2,950 entries, of which 234 are main entries (the number is only about half the number of the main entries of vol. VIII) coming with a presentation of the full etymology and history of the word. Of these 234 main entries over 40 turn out to be loanwords, mostly attained from Medieval Latin (though originally sometimes from elsewhere). About 25 % of all entries (about 60 % of the main entries and almost 22 % of the fillers) are continued in modern Standard German or in German dialects. With these percentages the volume takes a middle position among the last volumes.

Keywords Etymological dictionary; old high German; etymology; Germanic; Proto-Indo-European

1 Introduction

By mid-2026, vol. IX of the *Etymological Dictionary of Old High German* will (hopefully) have gone to the press. It contains the lemmas from *tâen* 'to breast-feed' till *ûzwurft* f. 'the act of throwing something or someone over bord'. The present article is a supplement to and continuation of the analyses of older volumes of the dictionary presented by Kozianka (2010a [vol. IV], 2010b), Lühr (2012), and Bichlmeier (2016a [vol. V], 2020a, 2025a [vol. VIII]). For some preliminary notes on vol. IX see Bichlmeier (2025b, 2026).

Here again, first a short overview over the history and the future of this dictionary will be offered. I shall not offer an insight into the general structure of the dictionary – for this the reader is kindly referred to the literature just mentioned. The second paragraph of the paper will present the contents of the volume and will mainly contain several statistical data as e.g. the number of the subgroups of lemmas, the percentage of loanwords and the percentage of the Old High German (henceforward: OHG) words, which still live on in (Modern) Standard New High German (henceforward: NHG) or in the NHG dialects. The (once prominent) question of whether a certain portion of the Germanic and thus German lexicon can be said to be of substrate origin, i.e. of pre-Germanic, but non-Celtic, potentially thus even non-Indo-European origin, and how big this portion is according to the data accumulated in the volumes of the *EWA* need not be of concern here (cf. for that point Bichlmeier, 2015, 2016a, 2016b, 2020b): The current volume contains roughly a dozen main entries for which other scholars had proposed a substrate origin, but in the end we could offer viable Indo-European etymologies for all of them.

2 History

The *EWA* is the amplest and most detailed etymological dictionary of any Germanic language. Work on the dictionary started in 1978 in the United States at the University of Pennsylvania by

Albert Larry Lloyd (†2025) and Otto Springer (†1991). The first volume of *EWA*, containing the words from *-a* to *bezzisto*, was published ten years later in 1988 (for the other volumes see the bibliography). Since 1987 Prof. Dr. Rosemarie Lühr figured as a collaborator, later on as co-editor and starting with vol. V finally as the single editor of the dictionary. In 2004 the dictionary was taken over into the programme of the Saxonian Academy of Sciences at Leipzig and has been written since then mainly at the chair of Indo-European Linguistics at the Friedrich-Schiller-University at Jena. At present the dictionary is produced primarily by a group of three scholars (indoeuropeanists throughout), PD Dr. Dagmar S. Wodtko (head of the team, full position), PD Dr. Harald Bichlmeier (75 %), Dr. Roland Schuhman (50 %), while Dr. Maria Kozianka (the former head of the research group), is still cooperating and writing entries on a part-time basis. Most lemmas have been written by the just mentioned persons in the last roughly two decades. In the last volumes a few articles have also always been written by senior students doing one-month practical stays with the project (in this case by Hannes Trost and Gleb Bogdanov).

The original work plan presented over 20 years ago envisioned four full positions, we have been working most of the time with slightly over 2 full positions, currently we are at almost 2.5 positions, which explains easily, why we are quite visibly behind schedule. The lag has grown a bit with every volume. Just to give an example: the volume *q- – sk-* should have been finished in 2019, *sl-* to *t-* in 2022, but as can be seen from Table 1, *s-* could be brought to an end only in 2023 etc. Moreover, we will now publish one volume more, that is 10 instead of 9 volumes, because things (i.e. lemmas and secondary literature) turned out to become longer and more.

TABLE 1 Plan and reality in dictionary-making.

plan		reality	
VII: *qu- – sk-*	2019	VII: *qu- – ska-*	2021
VIII: *sl- – t-*	2022	VIII: *ske- – swu-*	2024
IX: *u- – z-*	2025	IX: *t- – u-*	2026
X: *Index*	2025	X: *w- – z-*	2028
		XI: *Index*	2028/29

Originally, funding was granted until the end of 2025. We applied for another three years of funding in the summer of 2023, which was finally granted at the end of November 2025. Thus, we are now indeed able to finish the dictionary and even to write the index volume which had been promised under the original plan. Without the extension of financing, the dictionary would have remained without the words beginning with *w-* and *z-* and without the index.

Another important step was done in June 2023. The first five volumes of our dictionary finally went online – again with a delay two years – on woerterbuchnetz.de, hosted at the University of Trier. Thus, these volumes are now accessible for free. By the end of the current year all volumes up to the volume VIII will be online. Vol. IX will follow probably in 2028.

3 Contents and analysis

3.1 The lemmas of the dictionary and where they come from

Vol. IX will again consist of roughly 800 pages and will contain 2,950 lemmas (about 2 % less than did vol.s VII and VIII, which contained slightly over 3,000 entries). Of these only some 234 entries will be main entries (which is a number clearly below the numbers for the preceding volumes, vol. VIII, e.g., had 477) which offer full data on the etymology (back to

Proto-Indo-European) and the word history (up to the Standard NHG or the NHG dialects – if the word did not disappear before that) of each respective entry. The other lemmas only give data on the attestation of the words and – if they exist – correspondences in other (Old) Germanic languages, else for the etymology the reader is referred to the respective main entry.

The lemmas of the current volume were written by the following persons (cf. Table 2):

TABLE 2 Lemmas by collaborators

collaborator	main entries	fillers
Wodtko	2	2375
Bichlmeier[a]	137	0
Kozianka	1	153
Schuhmann	93	122
Trost	1	8
Bogdanov/Wodtko	0	29
Kozianka/Wodtko	0	3
Kozianka/Trost	0	18
Schuhmann/Wodtko	0	8
total:	**234**	**2716**

[a]This seemingly low number is due to the fact that the author already started working on vol. X and had produced about 120 main entries beginning with *wa-, we-, wi-* by the end of 2025, when the rest of the text of vol. IX finally approached completion.

The aim of our dictionary is to etymologize all OHG words, that is all German words attested between 750 CE and the 12th century. The borderline between OHG and Middle High German (henceforward: MHG) is traditionally put at 1050 CE (e.g. in most handbooks for students) but is often not taken into regard very strictly. In this we follow the way the data have been collected by the *Old High German Dictionary* [*Althochdeutsches Wörterbuch (AWb)*]. There the upper border is actually 1150 – but first entries of words taken into that dictionary may even be younger, if the colleagues regard them to be part of 'the Old High German tradition'. Thus, there are regularly words with first attestations in the 13th, 14th, and rarely even in the 15th centuries in that dictionary. These are then etymologized by our team, as the corpus of the Leipzig dictionary is our corpus in the end as well (see below Table 4 and Table 5).

The *AWb* is also compiled by the Saxonian Academy of Sciences at Leipzig. It is running for about eight decades now, the first fascicle appeared in 1952, and the colleagues will finish it in 2035. Up to the beginning of the lemmas beginning with *s-* we could conveniently use the data of that dictionary as the basis for our work, but as we overtook our colleagues in the first quarter of the *s*-part, we have to control now ourselves philological questions, spellings etc. concerning the words to be etymologized. This obviously slowed down our work to a certain degree.

The OHG lexicon comprises roughly 30,000 lexical units in the appellative lexicon (Bulitta, 2010a, p. 271). The up to now most complete dictionary of OHG, the one by Splett (1993), contains slightly over 29,400 lexemes. This number is a bit exaggerated in the end, as adverbs productively formed from adjectives (by adding an adverbial ending OHG *-o*) are counted as separate entries. So, taking these adverbs as forms of the respective adjectives would lower the number of lexemes certainly by a few hundred – thus, 28,500 to 29,000 lexical units would

probably a more accurate guess. Moreover, to my knowledge, nobody has ever counted, how many more OHG appellatives are attested in names (be it anthroponyms or toponyms) only.

3.2 Do OHG words continue their lives today?

Of these roughly 30,000 lexemes about 20,000, that is two thirds, are only attested in glosses to Latin texts. Many of these glosses clearly betray their momentary character, the writer often giving a seemingly more or less invented word as translation of a Latin word. These words seem to have been formations coined by the writers at the moment, when the gloss was written down – more than a few of them look like (and probably are) mere morpheme-by-morpheme translations. Especially such formations were prone to get lost again rapidly, as most of them hardly ever entered the spoken language. This is also reflected in the often rather different continuation-rates of main entries (words almost always also attested in actual texts) and fillers/secondary lemmas (words predominantly and often exclusively attested in glosses), which can be up to around 60 % for main entries (cf. the data for vol.s VIII and IX below) and sometimes below 20 % for fillers (cf. the data for vol. IX below) in certain sections of the dictionary.

E.g., in the current volume 24.7 % of the lexemes show continuation into any kind of NHG (over 19 % to [Standard] NHG – about a quarter of them marked as obsolete or antiquated, however –, over 5 % to NHG dialects). Thus, the percentage of OHG words of this volume which have lived on up to the 21st cent. will hardly exceed 20 % (cf. Table 3). This low percentage is certainly due to the high number (ca. 1,200) of privative compounds with the prefix/preverb *un-*, of which many seem to be simply translations of Latin words created in the glossing process. The volume's numbers thus lie between the numbers we find for vols. IV and V. Vol. VIII, however, appears as a clear upward exception: almost 40 % of the words treated in that volume exist still in Standard NHG or NHG dialects.

The current volume shows about the same continuation ratio which Maria Kozianka (2010a) found for vol. IV, in which some 900 of about 4,000 lexemes given in that volume have survived into Standard German, some 200 more have survived in dialects. I.e., in that part of the lexicon, comprising the words beginning with OHG *g-* and *h-*, only roughly 23 % of the words existing in OHG times were still in existence in (Standard) NHG (an average of some 15 % being marked as obsolete or antiquated), and another 5 % were still attested in German dialects in the 19th/20th century, when most dictionaries of German dialects were compiled. Thus roughly 28 % of these OHG lexemes made it into some kind of NHG.

TABLE 3 Continuation rates of OHG words in Standard New High German and in German dialects (ME = main entry; F = filler).

		lemmas	continuation in (Stand.) NHG	continuation in NHG dialects	total
Vol. IV		~4000	~900 (~22,5 %)	~200 (5 %)	**~1,100 (~27.5 %)**
Vol. V	*I-*	~570 (incl. 25 suff., 2 pref.)	~70 (~12 %)		
	J-	~130	~30 (~23 %)		
	K-	~1,270	~255 (~20 %)		
	L-	> 1,400	~235 (~17 %)		
	total:	~3,400	~590 (~17 %)	~100 (~3 %)	**~690 (~20 %)**

TABLE 3 (*CONTINUED*)

		lemmas	continuation in (Stand.) NHG	continuation in NHG dialects	total
Vol. VIII		~3,000	ME: 277 (269 + 8 obs.) = 58.1 % of the ME F: 728 (638 + 90 obs.) = 28.8 % of the F	ME: 62 = 13 % of the ME F: 125 = 4.9 % of the F	
	total:		1005 (33.4 %)	187 (6.2 %)	**1,192 (39.6 %)**
Vol. IX	*t-* *u-*	940 (ME: 176 + F: 764) 2,010 (ME: 58 + F: 1952)	ME 89 (50.1 %) F 182 (23.8 %) ME 20 (34.5 %) F 342 (17.5%)	ME 28 (15.9 %) F 33 (4.3 %) ME 5 (8.6 %) F 30 (1.5 %)	**total:** **ME 142 (60.7 %)** **F 587 (21.6 %)**
	total:	2,950	ME 109 (46.6 %) F 524 (19.3 %)	ME 33 (14.1 %) F 63 (5.2 %)	**729 (24.7 %)**
Vol.s IV + V + VIII + IX	**total:**	**~13,350**	**~3,130 (~23.45 %)**	**~585 (~4.38 %)**	**~3,715 (~27.83 %)**

Before vol. X will be finished (and before proper counts of the volumes I, II, III, VI, VII will have been conducted), it will of course be hard to give final numbers, but the numbers from the four volumes analysed so far, encompassing about 45 % of the attested OHG lexicon, we may guess that a maximum of 30 % of all OHG words have made it into NHG (maybe 25 % into [Standard] NHG + ca. 5 % into the NHG dialects) – but hardly more than a total of 25 % are still alive in the 21st cent.

3.3 Are all the words in EWA really OHG?

Looking through vol. IX and taking as a rule the periodisation of the history of the German language as given in standard (university) handbooks (thus accepting first attestations from the 11th cent. as definitely OHG, but those from the 12th cent. not necessarily so and if we treat dozens of (else in the dictionary not explicitly dated) first attestations from the *Summarium Heinrici* (a kind of dictionary, which was first presumably composed in the 11th cent., but has not come to us in manuscripts older than the 12th cent.) as unanimously OHG we get the following picture:

Among the lemmas treated in vol. IX almost 6 %have their first attestation in the 12th cent. (according to the data in the dictionary about half of them first attested after 1150 – the upper boundary applied in *AWb*), over 2 % in the 13th, almost 0.6 % in the 14th and 0.2 % (six words, one of them a main entry) even only in the 15th cent. (see Table 4). This means that by standard criteria of periodisation almost 9 % (and by the standards of *AWb* still some 7 %) of the words treated in this volume (roughly 1/11 – 1/14 respectively) are rather MHG or even Early NHG. Vol. VIII showed a similar distribution of first attestations; there we found even over 15 % of the lemmas not having first attestations in OHG times according to handbook periodisation (and by the standards of *AWb* still some 10 %; see Table 5).

TABLE 4 OHG words in vol. IX by first attestation (total: main entries 234; fillers 2716).

12th cent.	13th cent.	14th cent.	15th cent.
main entries | fillers	main entries | fillers	main entries | fillers	main entries | fillers
13 | 163	5 | 58	3 | 14	1 | 5
5.56 % | 6.00 %	2.23 % | 2.14 %	1.28 % | 0.52 %	0.43 % | 0.18 %
total:	total:	total:	total:
5.97 %	2.14 %	0.58 %	0.20 %
grand total: 262 (22 + 240) | 8.88 %			

TABLE 5 OHG words in vol. VIII by first attestation (total: main entries 477; fillers 2529).

12th cent.	13th cent.	14th cent.	15th cent.
main entries | fillers	main entries | fillers	main entries | fillers	main entries | fillers
27 | 278	10 | 84	2 | 52	0 | 3
5.56 % | 10.99 %	2.08 % | 3.32 %	0.4 % | 2.06 %	0 % | 0.12 %
total:	total:	total:	total:
305 | 10.11 %	94 | 3.11 %	54 | 1.8 %	3 | 0.1 %
grand total: 456 (39 + 417) | 15.15 %			

3.4 Loanwords in EWA and where they are from

Finally, a few words shall be shed on loanwords: Of course, final answers also on the question, how big the proportion of loanwords is among the main entries, can be given only after the finalisation of vol. X. An exact count has been conducted so far for vol. V (cf. Bichlmeier, 2016a, p. 72, 2016b, p. 330), and vol. VIII (cf. Bichlmeier 2025a, p. 59), while the numbers for vol. III and IV seem less exact. Taking together vol.s V, VIII, IX, for which data achieved in a comparable way is available, we see that about 20 % of the main entries are loanwords. Table 6 offers the data on the frequency of loanwords in vol.s III, IV, V, VIII, and IX:

TABLE 6 Loanwords in the *EWA*.

Vol.		lemmas (total)	main entries	loanwords	percentage
III + IV		~6,100		~100	(1.6 % of all)
V	*I-*	~570 (incl. 25 suff. and 2 pref.)	46 (73 incl. suff. und pref.: 46 + 25 +2)	17	~37 %
	J-	~130	27	9 (8+1)	~33 %
	K-	~1,270	293	103 (96+4+3)	~35 %
	L-	> 1,400	157	41	~26 %
	total:	**~3,400**	**524 (551)**	**170**	**~32 %** (2 % of all)
VIII		**~3,000**	**477**	**38**	**~8 %** (1,3 % of all)
IX		**~2,950**	**234**	**56**	**23.9 %**
total V + VIII + IX		**~9,350**	**1235**	**264**	**~21.4 %** (~2.8 % of all)

Over 95 % of the loanwords in OHG contained in vol.s V, VIII, IX are from Medieval Latin or Old Romance. A clear decision between these two origins is not always possible. For vol.s III and IV the numbers are not given exactly and are not split up according to the provenience of the words. Vol. VIII contained only 38 loanwords, 37 of them from Latin or Old Romance/Old French, one (a very young loan, first attestation in the 13th cent.) from Slavic.

The numbers pertained for vol. IX are clearly above the 8 % found for vol. VIII – in absolute numbers (56) as well as regarding the percentage (ca. 24 %) –, but just as clearly below the 32 % in vol. V. One young loanword (12th cent.) is from Slavic again, four are clearly from Old French, four are from Medieval Latin or Old French, the other 47 are from Medieval Latin or were loaned via Medieval Latin from other original languages. Of these 47 lexemes four are originally from Greek, one from an Old Iranian language and one containing a placename is from Egyptian. Table 7 shows the exact provenience of these loanwords:

TABLE 7 Provenience of the loanwords in vol. IX.

origin	*t-* (176 ME)	*u-* (58 ME)	total 234
(M)Lat. (< Lat.)	31[a]	7 (+ 1)[b]	39
(M)Lat. < Gr.	6[c]	0	6
(M)Lat. < Iran.	1[d]	0	1
(M)Lat. < Egypt.	1[e]	0	1
(M)Lat. total:	**39 | 22.2 %**	**8 | 13.8 %**	**47 | 20.1 %**
(M)Lat./OFr./Early Rom.	4[f]	0	4
OFr.	4[g]	0	4
Slav.	1[h]	0	1
total:	**48 | 27.3 %**	**8 | 13.8 %**	**56 | 23.9 %**

[a]Cf. OHG *tafala, taverna, talenta, tanterôn, tegil, tehmôn, tempal, temparôn, teppih, testîsarn, tinko*, tinkta, titul, torkul, torso, traht, trahtā?ri, trahtôn, traktât, treso, tribuz, trumba, trunk¹, tū?fstein, tunicha¹, tunna, tuom², turnella, turnen, turtella, turtur.*
[b]Cf. OHG *ûla, ulve*, umbrâl, unna, unza, urzeol, usinā?ri*; it is not clear, whether OHG *ulm* is a loan or inherited.
[c]Cf. OHG *timpana, tisc, tiufal, tracho, tragant, trîakil.*
[d]Cf. OHG *tigirtior.*
[e]Cf. OHG *Tanefeld* < placename *Tanis.*
[f]Cf. OHG *talamaska, taska, terzil, trapa.*
[g]Cf. OHG *tragmunt, tunil, turbil, tanzā?ri.*
[h]Cf. OHG *turnizza.*

4 Results and perspectives

Before the tenth and last volume will have been finished some of the questions touched upon in the preceding chapters will necessarily have to remain open. As soon as that volume will be presented, many final counts will be possible – and then it will finally also be possible to compare the situation of Old High German with the situation in parallelly attested languages (as e.g. in Old Slavonic; cf. for a first try on such a comparison Bichlmeier, 2024). Only then parallels and differences in the development of these languages will become visible. 2028 will be the year of reckoning.

Three results, however, already now seem to have a certain probability:

If 1150 CE is taken as the upper limit for OHG, about 10 % of the words treated in *EWA* are not OHG in that sense as they have their first attestations after 1150 CE.

The percentage of loanwords (predominantly from Medieval Latin and/or Old Romance/Old French) among the main entries will hover around 20 %.

The percentage of attested OHG words living on today in (Standard) NHG and NHG dialects combined will hardly exceed 25 % (or 30 %, if obsolete/antiquated words, i.e., NHG words which came out of use in the 18th/19th/20th centuries, are included).

References

AWb = Althochdeutsches Wörterbuch. Auf Grund der von Elias v. Steinmeyer hinterlassenen Sammlungen im Auftrag der Sächsischen Akademie der Wissenschaften zu Leipzig. E. Karg-Gasterstädt, Th. Frings, B. Bulitta R. Große, G. Lerchner & H. U. Schmid (Eds.). Akademie-Verlag & de Gruyter, 1952ff.

Bichlmeier, H. (2015). Zur Frage des nichtindogermanischen Substrats im Germanischen (auch anhand der Materialien des *Etymologischen Wörterbuchs des Althochdeutschen*). *Das Altertum*, 60(1), 1–17.

Bichlmeier, H. (2016a). The Etymological Dictionary of Old High German, Vol. V: *iba – luzzilo* [Etymologisches Wörterbuch des Althochdeutschen, Bd. V: *iba – luzzilo*]. In V. Domínguez-Rodríguez, A. Rodríguez-Álvarez, G. Rodríguez Herrera & V. C. Trujillo-González (Eds.), *Words Across History: Advances in Historical Lexicography and Lexicology* (pp. 67–77). Universidad de Las Palmas de Gran Canaria (CD-ROM; ISBN 978-84-9042-256-4).

Bichlmeier, H. (2016b). Zur Frage des nichtindogermanischen Substrats im Germanischen: Ein kurzer Überblick. *Slovo a Slovesnost*, 77, 316–336.

Bichlmeier, H. (2020a). Das *Etymologische Wörterbuch des Althochdeutschen* als Vorbild für ein etymologisches Wörterbuch des Alttschechischen. In V.Boček (Ed.), *Teorie a praxe české etymologické lexikografie* (pp. 51–78). Nakladatelství Lidové Noviny (NLN) 2019 [2020].

Bichlmeier, H. (2020b). On the question of the non-IE substratum in Germanic. In H. Bichlmeier, O. Šefčík & R. Sukač (Eds.), *Etymologus: Festschrift for Václav Blažek* (pp. 73–92). Baar-Verlag.

Bichlmeier, H. (2024). *Etymologický slovník staroslověnštiny* a *Etymologický slovník starohornoněmčiny* – srovnání dvou fakticky paralelních projektů. In V. Boček (Ed.), *Etymologický slovník jazyka staroslověnského: tradice, kontexty, inspirace* (pp. 69–81 and 143–155 [bibliography]). Nakladatelství Lidové Noviny (NLN) 2023 [2024].

Bichlmeier, H. (2025a). Band VIII (*skebidîg – swummôd*) des Etymologischen Wörterbuchs des Althochdeutschen. In H. Bichlmeier, E. Klotz & F. Steindl (Eds.), *Vorträge des XXXVI. Namenkundlichen Symposiums in Kals am Großglockner, 6.–9. Juni 2024* (pp. 45–64). Praesens-Verlag.

Bichlmeier, H. (2025b). Vol. IX of the *Etymological Dictionary of Old High German*. In A. Salgado, A. Villalva, E. Cardeira & R. Costa (Eds.), *ICHLL 15. 24–27 June 2025, Lisbon, Portugal*. [book of abstracts] (pp. 39–40). ACL Portugal. https://doi.org/10.58164/q9mt-tn09

Bichlmeier, H. (2026). Band IX (*tâen – ûzwurft*) des Etymologischen Wörterbuchs des Althochdeutschen. In H. Bichlmeier, E. Klotz & F. Steindl (Eds.), *Vorträge des XXXVII. Namenkundlichen Symposiums in Kals am Großglockner, 29. Mai – 1. Juni 2025* (pp. 37–44). Praesens-Verlag.

Bulitta, B. (2009). Das Althochdeutsche Wörterbuch. Philologische Grundlagenforschung an der Sächsischen Akademie der Wissenschaften zu Leipzig. *Denkströme. Journal der Sächsischen Akademie der Wissenschaften*, 2, 83–98.

Bulitta, B. (2010a). Stand und Perspektiven der Lexikographie des Althochdeutschen. *Jahrbuch für Germanistische Sprachgeschichte. Perspektiven der germanistischen Sprachgeschichtsforschung*, 1, 269–291.

Bulitta, B. (2010b). Das Althochdeutsche Wörterbuch. Philologische Grundlagenforschung an der Sächsischen Akademie der Wissenschaften zu Leipzig. *Mitteilungen des Deutschen Germanistenverbandes*, 57(4), 363–377.

EWA = Etymologisches Wörterbuch des Althochdeutschen. Vol. I: *-a – bezzisto*. By A. L. Lloyd, O. Springer Vandenhoeck & Ruprecht 1988; Vol. II: *bî – ezzo*. By A. L. Lloyd, R. Lühr, O. Springer † with

collaboration of K. R. Purdy. Vandenhoeck & Ruprecht 1998; Vol. III: *fadum – fûstslag*. By Albert L. Lloyd, R. Lühr with the collaboration of G. Kohlrusch, M. Kozianka, K. R. Purdy, R. Schuhmann. Vandenhoeck & Ruprecht 2007; Vol. IV: *gâba – hylare*. By A. L. Lloyd, R. Lühr with collaboration of G. Kohlrusch, M. Kozianka, K. R. Purdy, R. Schuhmann. Vandenhoeck & Ruprecht 2009; Vol. V: *iba – luzzilo*. Ed. by R. Lühr, written by H. Bichlmeier, M. Kozianka, R. Schuhmann with articles by A. L. Lloyd with collaboration of K. R. Purdy. Vandenhoeck & Ruprecht 2014; Vol. VI: *mâda – pûzza*. Ed. by R. Lühr, written by H. Bichlmeier, M. Kozianka, R. Schuhmann, L. Sturm. Vandenhoeck & Ruprecht 2017; Vol. VII: *quaderna – skazzôn*. Ed. by R Lühr, written by D. S. Wodtko, H. Bichlmeier, M. Kozianka, R. Schuhmann. Vandenhoeck & Ruprecht 2021; Vol. VIII: *skebidîg – swummôd*. Ed. by R. Lühr, written by D. S. Wodtko, H. Bichlmeier, M. Kozianka, R. Schuhmann. Vandenhoeck & Ruprecht 2024; Vol. IX: *tâen – ûzwurft*. Written by D. S. Wodtko, H. Bichlmeier, M. Kozianka, R. Schuhmann. Vandenhoeck & Ruprecht 2026.

Kozianka, M. (2010a): Etymologisches Wörterbuch des Althochdeutschen. Band IV (*gâba – hylare*). *Denkströme. Journal der Sächsischen Akademie der Wissenschaften*, 4, 216–220.

Kozianka, M. (2010b). Loanwords in the "Etymological Dictionary of Old High German". Retrieved October 22, 2014, from http://ora.ox.ac.uk/objects/uuid:35ad4916-bb11-4351-9818-ff5857b84463

Lühr, R. (2012). Zum "Etymologischen Wörterbuch des Althochdeutschen (EWA)". In B. Vykypěl & V. Boček (Eds.), *Methods of Etymological Practice* (pp. 97–103). Nakladatelství Lidové Noviny (NLN).

Splett, J. (1993). *Althochdeutsches Wörterbuch. Analyse der Wortfamilienstrukturen des Althochdeutschen, zugleich Grundlegung einer zukünftigen Strukturgeschichte des deutschen Wortschatzes* (2 vol.s in 3 books). de Gruyter.

Contact information

Harald Bichlmeier (ORCID 0000-0003-0001-1677)
Sächsische Akademie der Wissenschaften zu Leipzig, Arbeitsstelle Jena: Etymologisches Wörterbuch des Althochdeutschen
harald.bichlmeier@uni-jena.de

Martin-Luther-Universität Halle-Wittenberg, Orientalisches Institut, Seminar für Indogermanistik und Allgemeine Sprachwissenschaft
harald.bichlmeier@indogerm.uni-halle.de

Lietuvių Kalbos Institutas
harald.bichlmeier@lki.lt

Luisa Cimander

ENGLISH ADJECTIVES IN GERMAN

Processes of Integration since the 17th Century and Lexicographic Implications

Abstract This paper presents the current state of a PhD project on the integration of English adjectives into German. While English nominal and verbal loanwords have undergone thorough investigation, adjectives were often only covered marginally. Some researchers have approached English adjectives more systematically and have identified integration processes and problems, but they mostly examined small data sets or specific genres (e.g. newspapers) at a certain point in time. For this PhD project, adjectives were collected from 1650 up to 2025 to create a substantial, diachronic basis for analysis. This large time span was further divided into three time periods, based on major turning points in 1870 and 1945, to render diachronic comparisons possible. The sources were scientific works as well as various types of dictionaries (e.g. general, foreign words, Germanising). First results show an increase of adjectives between the time periods, with a total of roughly 1000 adjectives in the collection. Currently, the adjectives are being categorised according to their morphological, phonological and orthographical properties, with the assumption being that adjectives with the same properties behave in the same way in regards to inflection and syntactic usage. These groups will later be verified through further (corpus) analyses.

Keywords English loan adjectives; German adjectives; English loanwords in German; linguistic integration of loanwords; diachronic integration processes; lexicographic registration of loanwords

1 Introduction

Despite the continuous borrowing of English words into German, as well as the steady interest in them for well over a hundred years, few researchers have put dedicated focus on the integration of English adjectives into the German language system. The focus usually lies on nouns, in particular the assignment of a grammatical gender, or on verbs and their conjugational patterns. A reason for the oversight could be their seeming irrelevance with regard to their number, as nouns make up around 75% of all words in the German orthographic dictionary *Duden* (cf. Duden, 2026). There has been some debate in previous research on the word class distribution of English loanwords found in German newspapers, in particular if verbs or adjectives are borrowed more frequently. Yang, for example, claimed in his study that fewer adjectives were borrowed than nouns and verbs (Yang, 1990, pp. 28 f.), while Viereck found adjectives to be borrowed more frequently than verbs (Viereck, 1980, p. 229), and Onysko documented roughly the same frequency (Onysko, 2007, p. 248).

Burmasova approached this debate with a different perspective: She claimed that previous research had fallen victim to a methodical error as it usually focused on absolute numbers that were not put in relation to the natural occurrence of the different parts of speech in the lexicon (Burmasova, 2010, p. 65). In her analysis, she did not compare the number of nouns, adjectives and verbs with each other, but the increase of items between 1994 and 2004 within the three-word classes. She concluded that the increase rate of loanwords in all three classes was roughly the same (Burmasova, 2010, p. 223). These results show that there is no reason to disregard English adjectival loanwords in analyses as they are as much a part of the German language as nominal or verbal loanwords. This PhD project thus tries to answer the question of how English adjectives are being integrated into German, while the focus lies not on individual adjectives,

but on processes regarding the whole word class in a diachronic approach from the 17th century until today.

2 Research on English Loan Adjectives in German

The first researcher to mention a systematic integration process was Ganz, who wrote that in the 17th and 18th century, most English adjectives were integrated by adding the German suffix *-isch*, as in German *anglikanisch* compared to English *Anglican* (1957, p. 6). Later analyses mainly mentioned two groups of English adjectives: The ones that can be inflected like native adjectives and thus can be used attributively, e.g. *der clevere Geschäftsmann* 'the clever businessman' (Zindler, 1959, p. 15), and the ones that are only used in a predicative position as they cannot be inflected – or rather, to actively avoid inflectional problems, e.g. *Er ist down, groggy* 'He is down, groggy' (Carstensen, 1965, p. 68).[1]

Eisenberg was the first to distinguish between seven groups of English adjectives that differ in the way they are inflected and used in German, which are: 1) One-syllable words with foreign orthography or phonology that inflect like native adjectives (*cool*) (Eisenberg, 2018[2], p. 238). 2) Two-syllable words containing a syllabic sonorant that could lead to forms with and without a schwa in inflection (*cleveres*/**clevres*). 3) Adjectives usually found in post-nominal position due to their inability to inflect (*light*). 4) Lexemes ending in a plosive (*fit*) that do not inflect (in written language) as the consonant would then have to be doubled due to orthographical reasons (**fittes*). 5) The colour adjective *pink* that is, like other colour adjectives, rarely inflected (ibid.). 6) Complex adjectives ending in *-y* (*trendy*), which cannot inflect, even in attributive usage (e.g. *ein trendy Buch* 'a trendy book'), as the second syllable ends in an unstressed full vowel (Eisenberg, 2018, p. 239). 7) English participles ending in *-ed* (*relaxed*) that are integrated with a *t* in inflection (e.g. *relaxtes Verhalten* 'relaxed behaviour') (ibid.). However, these groups differ in the linguistic levels by which they are categorised and are therefore not easily comparable. Additionally, some of Eisenberg's groupings and conclusions can be challenged or considered not up-to-date anymore. The inflection of adjectives like *fit* as *fittes*, for example, is now standard.

Onysko claimed that generally "the predicative position seems more susceptible to initial borrowing of an adjectival anglicism than the attributive position", which is why a "majority of predicative adjectival anglicisms in German only occur as such" (Onysko, 2007, p. 251). He also covered some of the adjective groups identified by Eisenberg and added that the defiance of "attributive inflection due to phonological reasons" with adjectives ending in *-y* is sometimes "circumvented by means of denominal derivation" with *-ig* (*trendig*) (Onysko, 2007, p. 251). He listed borrowed present participles (*boomend*) in addition to past participles (Onysko, 2007, p. 257) and also mentioned a range of suffixes like *-isch* and *-ig* that "combine with a variety of mostly nominal bases to form adjectival anglicisms" (Onysko, 2007, p. 259). Finally, he found that, when compared, English adjectives "regularly inflect with comparative *-er* and superlative *-est*" (Onysko, 2007, p. 252).

Burmasova suggested another reason why specific adjectives are rarely inflected and thus not used attributively: They compete with the adjectives' usage in compounds (e.g. **superes Angebot*, but *Superangebot* 'super offer'), which stems from encountering inflectional difficulties (Burmasova, 2010, p. 103). Pittner, on the other hand, focused on the emergence of adjectives

[1] Cf. also Viereck, 1980, p. 233, and Yang, 1990, p. 162.

[2] This research was first published in Eisenberg (2004, pp. 128–131) and was only slightly modified for the book, which was first released in 2011. The chapter about adjectives was not changed for the third edition of the book.

like *top*, *super* and *mega* through the process of debonding (Pittner, 2022, pp. 156 ff.). Both analyses indicate that these adjectives' usage has to be examined in relation to their word formation processes.

3 Data Collection and Sources

This PhD project aims to ground its analysis of English adjectives on a much larger data basis than previous research: Adjectives were collected from the 17th century until today, as the year 1640 has been described convincingly as a turning point in German-English language relations. The increasing interest in English society and culture after the English Civil War and the following execution of King Charles I resulted in an increase in English borrowings (Ganz, 1957, p. 13). Some English words were used in German before 1640 as well, but the only adjective documented is *englisch*, which appeared around 1220 (Palmer, 1950, p. 5). It was decided, for this study, to further divide this long time span into three time periods based on two major historical turning points. After the first period from 1640 onwards, the second time period is set to start in 1870 due to the foundation of the German Empire in 1871, and the World Wars soon after, which led to increasingly nationalistic tendencies and strong linguistic purism (cf. Stiven, 1936, pp. 101 f.). Simultaneously, the influence of the English language increased through the USA's rise to a world power, the influence of the British Empire, and the growing trade connections (cf. Stiven, 1936, p. 82). The third period then begins in 1945, since the influence of English skyrocketed after World War II due to technical and cultural developments as well as the increased internationalisation and globalisation, which continues to this day (cf. Carstensen & Busse, 1993, p. 27).

Scientific analyses as well as different types of dictionaries (e.g. general, foreign words) were used for the diachronical collection of adjectives.[3] However, the data bases for each time period look very different: No major scientific compilation of English loanwords could be found that had been composed in the time up until 1870, although researches like Stiven, Ganz and Palmer analysed the language usage in retrospect, finding a high number of loanwords. Kinderling, for example, only documented 21 English words at his time, and none of them were adjectives (Kinderling, 1795, p. 109). After 1870, however, the scientific examination of the English influence on German became very popular (e.g. Meyerfeld, 1903; Ziegelschmid, 1935). A prominent figure at the time was Dunger, a linguistic purist and advocate against the usage of English loanwords. His *Wörterbuch von Verdeutschungen* ('Germanising dictionary') (1882) as well as many Germanising dictionaries published by the purist *Allgemeiner Deutscher Sprachverein* ('General German Language Association') (cf. Bernsmeier, 1980) are great resources for the English language in use at the time. The scientific interest only increased after World War II, with the existence of large digital corpora resulting in many analyses of English loanwords in newspapers (e.g. Zindler, 1959; Yang, 1990). This interest also resulted in a major work, the *Anglizismen-Wörterbuch* ('Dictionary of Anglicisms') by Carstensen & Busse (1993–1996).

The collection currently contains approximately 50 words documented for the first time period, around 150 for the second, and about 900 for the third, showing a significant rise in the usage of borrowed English adjectives across the three-time spans, particularly in the last 80 years.[4] Some adjectives were documented in more than one time period, while others were exclusive to only one or two, according to the works consulted. The collection should contain a great part of the English adjectives that were relatively frequent, and thus have to have been in some part integrated into German in the last 400 years. It, therefore, provides

[3] Only selected works can be mentioned here.

[4] The numbers used in this paper are preliminary and will likely be subject to change.

an empirical and historical foundation for further analyses, while the division into three time periods can be a way of identifying differing integration processes at different times.

4 Categorisation of the Adjectives

With the collection of adjectives finalised, they are now to be analysed thoroughly and systematically. As a first step, the adjectives will be categorised according to different linguistic features, beginning with their morphological properties. This approach is based on the assumption that adjectives with a comparable structure will present similar patterns in inflection and comparison, their syntactical usage, as well as the problems that arise when they are used in German. Therefore, the focus during the categorisation will initially lie on the adjectives' suffixes, but phonological and orthographical properties will also be considered when they become relevant. While previous analyses turned to the adjectives' form only when encountering inflectional problems as an attempt to explain them, this categorisation will use it as its starting point.

A first distinction can be made between adjectives that are borrowed in the original English form and those that are morphologically changed when entering the German language.[5] Both English and German have a wide set of adjectival suffixes and, since the languages are related, in many cases an equivalent to a suffix exists in the other language (e.g. Engl. *-able* vs. Ger. *-abel*, Engl. *-ish* vs. Ger. *-isch*). These suffixal equivalents can allow a smooth change from the English to the German form, usually with a transitional period in which both forms are used.[6] With the change to the German suffix, inflectional problems are then mostly eliminated. Semi-suffixes, whose "independent status leads to borderline cases between suffixation and compounding" (Onysko, 2007, p. 260), also often have equivalents in German (e.g. Engl. *-made* vs. Ger. *-gemacht*). In these cases, however, a substitution of the semi-suffix seems unlikely (e.g. *homemade* to **homegemacht*). The words are either used in the English form, or rather translated fully into German or exchanged for an existing word (*hausgemacht*).

The categorisation of adjectives that come into the German language unchanged in form is more complicated, and more factors than just suffixes have to be considered. Adjectives ending in the suffix *-al* present a special group: Since this suffix exists in both languages, the adjective's form does not have to be adapted. However, the stress switches to the last syllable in German (e.g. Engl. *environméntal* vs. Ger. *environmentál*), thus allowing for easier inflection.[7] Another prominent case that was mentioned in Chapter 2 is the word ending *-y* that is sometimes changed to the suffix *-ig*. However, this only seems to be the case for adjectives that are derivations of nouns (e.g. *bluesy*). Other adjectives like *happy* that do not have a nominal base do not seem prone to this suffixal change. Adjectives like *fly* might behave differently altogether because they are monosyllabic. Finally, many English adjectives just do not have a distinguishable suffix (e.g. *weird*, *random*). In cases like these, other factors like syllable count and structure will have to be taken into account as well.

5 Further Steps and Lexicographic Impact

The categorisation of the adjectives will function as a basis for further (corpus) analyses, in which it will be examined how the adjectives are inflected and used syntactically. This

[5] This refers only to the base form and not to possible changes through inflection and comparison.

[6] In reality, of course, this process is not as simple and sometimes both forms are used next to each other for a long period of time (e.g. *fashionable/fashionabel* since the 19th century) or one of them does not even establish itself.

[7] The before-mentioned processes of suffix change also often go hand in hand with a stress change.

approach allows the validation of the main hypotheses: Are the adjectives that are grouped together really used in the same or a comparable way? It is important that this analysis is carried out on the basis of representative and appropriate data. Previous analyses often used newspapers, and in particular only one specific newspaper, like *Der Spiegel* or *Die Welt*, to analyse the usage of English loanwords. However, many loanwords are used in more specialised contexts or in specific group settings (e.g. youth language), and especially non-standardised inflections are often found in spoken language (e.g. *am *easyesten* 'the easiest'). It will be important to select the right data, or even collect new language and metalinguistic data through an empirical study, in order to access current processes in adjectival loanword usage and integration.

Dictionaries are being used in this PhD project as one of the sources for the diachronic overview of English adjectives loaned into German. However, the findings of this project can also contribute to the documentation of English adjectival loanwords in dictionaries. Language users often expect prescriptive guidance for authentic language usage from dictionaries, while these mostly restrict themselves to describing the (standard) usage of a word. Grouping new-found English adjectives into previously determined categories, for example according to morphological properties as well as early non-standard inflectional forms, would allow a lexicographer to form predictions about how their inflectional and syntactical usage might develop in German over different adaption stages and with potentially variable forms. These predictions could help speakers with not only using the adjectives, but also with understanding the linguistic process of borrowing and integrating loanwords into German. However, this approach would require innovative lexicographic practices that focus on diachronical integration processes across different linguistic levels beyond the scope of individual words.

References

Bernsmeier, H. (1980). Der Allgemeine Deutsche Sprachverein in der Zeit von 1912 bis 1932. *Muttersprache*, 90, 117–140.

Burmasova, S. (2010). *Empirische Untersuchung der Anglizismen im Deutschen am Material der Zeitung Die WELT (Jahrgänge 1994 und 2004)*. University of Bamberg Press.

Carstensen, B. (1965). *Englische Einflüsse auf die deutsche Sprache nach 1945*. Carl Winter Universitätsverlag.

Carstensen, B. & Busse, U. (Eds.) (1993). *Anglizismen-Wörterbuch. Der Einfluß des Englischen auf den deutschen Wortschatz nach 1945*. Band 1. A–E. De Gruyter.

Duden. Die Verteilung der Wortarten im Rechtschreibduden. Retrieved May 13, 2026, from https://www.duden.de/sprachwissen/sprachratgeber/Die-Verteilung-der-Wortarten-im-Rechtschreibduden

Dunger, H. (1882). *Wörterbuch von Verdeutschungen entbehrlicher Fremdwörter mit besonderer Berücksichtigung der von dem Großen Generalstabe, im Postwesen und in der Reichsgesetzgebung angenommenen Verdeutschungen. Mit einer einleitenden Abhandlung über Fremdwörter und Sprachreinigung*. Teubner.

Eisenberg, P. (2004). German as an Endangered Language? In A. Gardt & B. Hüppauf (Eds.), *Globalization and the Future of German. With a Select Bibliography* (pp. 121–138). De Gruyter.

Eisenberg, P. (2018[3]). *Das Fremdwort im Deutschen*. De Gruyter.

Ganz, P. F. (1957). *Der Einfluß des Englischen auf den deutschen Wortschatz 1640–1815*. Erich Schmidt Verlag.

Kinderling, J. F. A. (1795). *Über die Reinigkeit der Deutschen Sprache, und die Beförderungsmittel derselben, mit einer Musterung der fremden Wörter und andern Wörterverzeichnissen*. Friedrich Maurer.

Meyerfeld, M. (1903). *Von Sprach' und Art der Deutschen und Engländer. Kritische Worte und Wortkritik*. Mayer & Müller.

Onysko, A. (2007). *Anglicisms in German: borrowing, lexical productivity, and written codeswitching*. De Gruyter.

Palmer, P. M. (1950). The Influence of English on the German Vocabulary to 1700. *University of California Publications in Linguistics*, 7(1), 1–38.

Pittner, K. (2022). Vom prädikativen zum attributiven Adjektiv: Die Integration von englischen Adjektiven ins Deutsche. In L. Auteri, N. Barrale, A. Di Bella & S. Hoffmann (Eds.), *Jahrbuch für Internationale Germanistik. Wege der Germanistik in transkultureller Perspektive* (pp. 151–166). Peter Lang.

Stiven, A. B. (1936). *Englands Einfluß auf den deutschen Wortschatz*. Bernhard Sporn.

Viereck, K. (1980). *Englisches Wortgut, seine Häufigkeit und Integration in der österreichischen und bundesdeutschen Pressesprache*. Peter D. Lang.

Yang, W. (1990). *Anglizismen im Deutschen. Am Beispiel des Nachrichtenmagazins DER SPIEGEL*. Max Niemeyer Verlag.

Ziegelschmid, A. J. F. (1935). Englisch-Amerikanischer Einfluss auf den Wortschatz der deutschen Sprache der Nachkriegzeit. *The Journal of English and Germanic Philology*, 34(1), 24–33.

Zindler, H. (1959). *Anglizismen in der deutschen Presse nach 1945* [Doctoral dissertation, Universität Kiel]. IDS OPUS.

Contact information

Luisa Cimander
Leibniz Institute for the German Language, Mannheim
cimander@ids-mannheim.de

Margarita Freixas, Esther Hernández & Marija Žarković Eriksson

E-*DRAE* 1884

Toward a Digital and Hypertextual Critical Edition of a Nineteenth-Century Dictionary

Abstract This paper presents the project "Model for a Digital and Hypertextual Edition of the *DRAE* 1884 (e-*DRAE* 1884)", carried out at the Autonomous University of Barcelona and the Spanish National Research Council (CSIC). Its main goal is to produce an online critical edition of the *Diccionario de la lengua castellana* published by the Real Academia Española in 1884 (hereafter, *DRAE* 1884). The 1884 edition of the *DRAE* was chosen because of its innovative character within the lexicographical tradition of the Real Academia Española and because it was the first edition to take into account the lexical proposals of the newly established American academies. The digital and hypertextual edition offers the full text in an interoperable format, https://edrae1884.uab.cat/edicion/, enriched with structured annotations that provide the information needed for the historiographical, lexicographical, and lexicological interpretation of the dictionary's contents.

Keywords Spanish lexicography; digital humanities; dictionary of the Real Academia Española; nineteenth century

1 Introduction

Dictionaries are products of their time: they reflect the historical and social context in which they are produced and, in turn, can help shape a community's cultural identity. They constitute a dense network of information about language in general and about particular linguistic traditions (Durkin, 2016, p. 3).

In 2026, the Real Academia Española celebrates three centuries of its lexicographical tradition, which spans from the *Diccionario de autoridades* (1726–1739) to the latest edition of the *Diccionario de la lengua española* (2014; electronic update 2025, version 23.8.1). The twenty-three editions of this dictionary, commonly referred to by the acronym *DRAE*, together with the *Diccionario de autoridades*, form an excellent corpus for studying the evolution of the language from different perspectives, and they are an essential part of the history of Spanish lexicography as well as a normative reference for Spanish speakers. In the absence of a complete historical dictionary of Spanish—beyond the resource offered by the *Diccionario histórico de la lengua española* (*DHLE*)—it can be stated that "el diccionario de la Academia es un diccionario con historia. Una obra en la que opera constantemente el peso de su propio pasado" (Seco, 1988, p. 85) ["The Academy's dictionary is a dictionary with a history, a work in which the weight of its own past is constantly at play"]. Consequently, its value lies fundamentally both in its record of the lexical changes that have taken place in Spanish and in the possibilities for metalexicographical study afforded by the successive editions of the dictionary of the Real Academia Española. The twelfth edition, published in 1884, marked a significant turning point in the evolution of the Academy's lexicographical tradition, and for several reasons it has been selected for a digital critical edition.

First, the twelfth edition of the *DRAE* appeared at a decisive historical moment: after a period of political turmoil (1868–1874) and a decade of relative social and political stability (1875-1884). In addition, from the 1860s onward, and in the new context opened by American independence, a process of institutional rapprochement took shape whose most significant manifestation was the creation of the corresponding American academies (Pérez, 2026), with which a

fruitful collaboration began (Clavería & Hernández, 2021). At the same time, between the eighteenth and late nineteenth centuries, scientific and technical vocabulary underwent extensive modernization. The *DRAE* 1884 registered these changes through the incorporation of new terms and, above all, through the revision of numerous definitions of specialized vocabulary in fields such as printing, natural history, maritime terminology, and medicine. Finally, the first "scientific" dictionaries began to emerge in Europe as a result of nineteenth-century comparative philology. The introduction of etymology in the *DRAE* 1884 reflects the lexicographical advances achieved through the new historical methods. These circumstances are clearly reflected in the entries of this edition of the dictionary, whose critical analysis contributes to a deeper understanding of the linguistic, social, and historical dynamics of the second half of the nineteenth century.

2 Toward a New Dictionary Model: Structural Aspects of the DRAE 1884

From the perspective of lexicographical practice, the *DRAE* 1884 represents a decisive milestone in the history of the dictionary of the Real Academia Española, as it marks a qualitative shift in the development of a work widely regarded as a fundamental part of the linguistic heritage of the Spanish-speaking world. As regards the macrostructure, it is, by a wide margin, the nineteenth-century edition that incorporates the largest number of lemmas and lexical forms (Clavería & Terrón, 2022) and that shows a particular interest in the inclusion of Americanisms, specialized vocabulary, and neologisms, thanks not only to the knowledge of the members of the Academy but also to the contributions of institutions such as the Academy of Sciences and the Academy of Medicine, as well as numerous outside collaborators. It is also the edition in which the systematic distinction of homographic or *univocal* entries is introduced in separate entries, following grammatical and semantic criteria (Terrón, 2026). With regard to its microstructure, the dictionary reflects a thorough revision of its lexicographical technique. The application of a "general plan" (*Advertencia*), established in the *Reglas para la corrección y aumento del diccionario vulgar* (1869 and 1870), introduces a new onomasiological method for revising entries by "special subject areas", which entails a comprehensive vision of the dictionary, in which the process of revision is applied to broad domains: linguistics, mathematics, hunting and fishing, natural history, the physical sciences, medicine, philosophy, rhetoric, literature, history, jurisprudence, political economy, theological disciplines, the fine arts, the military arts, maritime terminology, the mechanical arts, games and pastimes, and various other fields of study. The revisions of the entries deeply affect the various parts of the lexicographical article, transforming both the typographical aspects of the dictionary and its content. Etymological data are incorporated into the entries. Grammatical information is revised and made more precise in the *DRAE* 1884 (Hernando García-Cervigón, 2026; Raab, 2026). Diachronic labels are simplified, whereas geographic and technical labels are revised. Senses are reformulated for precision and clarity, as are the definition patterns used for entries in the same lexical category and for lemmas belonging to related lexical series (Freixas, 2026). Usage examples are updated, and the system of cross-references to lemmas and senses is substantially improved. All of this constitutes a new model of lexicographical work, the study and digital critical edition of which is the focus of the coordinated project "Model for a Digital and Hypertextual Edition of the *DRAE* 1884 (e-*DRAE* 1884)" (see https://edrae1884.uab.cat/).

3 Retrodigitization and Digital Humanities

The critical digital edition e-*DRAE* 1884 is conceived from the perspective of retrodigitization, a digital humanities procedure aimed at converting printed dictionaries from before the computer era into structured, machine-readable digital resources.

A substantial body of scholarship has addressed the digitization of historical dictionaries (Kosem et al. 2013; Considine, 2017). In the case of the dictionaries of the Spanish Academy, resources such as the *Nuevo Tesoro Lexicográfico de la Lengua Española*, the *Diccionario de autoridades*, and the *Mapa de diccionarios* are already available. None of these resources, however, has been conceived from the perspective of a critical edition.

This project therefore aims to be a model not only for textual scholarship and linguistic historiography, but also a methodological model for the digital humanities. European initiatives such as DARIAH and ELEXIS have promoted retrodigitization through projects that transform historical repertoires—such as the *Latin-Bulgarian Dictionary*, the *Foclóir Gaeilge-Béarla*, the *Diccionario da Língua Portugueza*, the *Orotariko Euskal Hiztegia*, or *TELEI*, the *Tesoro Lexicográfico Español-Italiano*—into interoperable data by means of extraction techniques, TEI encoding, and semantic enrichment. Their aim is to preserve, normalize, and make reusable the lexicographical heritage, ensuring its integration into contemporary research ecosystems (Nalesso, 2024, p. 139).

4 Text Processing and the Critical Edition

The technical challenge of the e-*DRAE* 1884 project lay in transferring a text conceived for print, comprising 1,118 pages, 57,084 headwords, and 115,423 senses, into a digital environment capable of exploiting the hypertextuality, non-linear reading practices, and structured querying associated with the digital texts. The aim was not simply to provide a navigable version of the dictionary, but to develop a query system that would allow lexicographical information to be filtered, cross-referenced, and retrieved in ways unavailable in the printed format. It is thus an edition that brings together, within a single tool, features characteristic of different editorial models: facsimile, interpretive, critical, and annotated (cf. Clavería, Freixas, Hernández, & Torruella, 2026).

4.1 From Source Data to a Relational Database

The starting point consisted of the PDF and XML files of the original text provided by the Real Academia Española. Three successive operations were performed on this material: the re-encoding of the initial XML markup in accordance with the TEI standard, specifically the dictionaries module; the segmentation of the text into hierarchically organized units, including entry, headword, etymon, sense, and sub-entry; and, finally, the automatic export of the resulting data to a relational database. The use of TEI is justified by its status as an international standard for the exchange of linguistic data and by the guarantees it provides in terms of durability, interoperability, and reusability, in line with the FAIR principles.

The application has been developed entirely with open-source technologies: PHP, using the CodeIgniter framework for the application logic, and MySQL as the relational database management system. This combination facilitates long-term maintenance and avoids dependence on commercial vendors, in keeping with the portal's commitment to open access. The basic textual unit of the model is the entry, identified by its headword. The main table, *drae_ENTRADAS*, records the lemma and its normalized form for sorting, together with three features specific to the 1884 edition: the marker for new lemmas, obtained through comparison with the *Lemateca del DRAE*; homograph status, one of the innovations introduced in this edition; and inclusion in the supplement. Senses are managed in a separate table, *drae_ACEPCIONES*, linked to the main table and provided with an order field that preserves the original editorial sequence. Etymological information, another distinctive feature of the *DRAE* 1884, has been divided into two tables: one for source languages, *drae_LENGUAS*, and another for the word-formation

processes, *drae_PROCESOS*, that may appear within the etymological parenthesis. This dual structure allows for precise etymological queries, such as all headwords with an Arabic etymon or all those formed by derivation. Auxiliary tables for grammatical category and lexicographical labels are currently under development. The model is completed by a paratexts table, *drae_PARATEXTOS*, which covers the fourteen preliminary and final texts of the work, and by tables devoted to users, comments, and bibliographical references, which support the critical dimension of the edition.

Each entry also preserves its topographical reference to the original, namely page and column, thereby making it possible to generate a link to the JPG image of the corresponding facsimile page. This bidirectional link between encoded text and image is central to the edition, since it guarantees the philological verification of all the information recorded in the database (cf. Clavería, Freixas, Hernández & Torruella, 2026).

4.2 The Search Interface and Search Methods

The public interface has been conceived as a single point of access that combines the logic of the traditional dictionary (entering through the headword, reading the entry) with the advanced search and filtering capabilities made possible by the underlying database. The main menu is organized into three blocks: search options, paratexts, and bibliographical references. Table 1 summarizes the four search modes available.

TABLE 1 Search methods of the e-*DRAE* 1884 query interface

Search method	Queried field	Match type	Available filters
lemma search	entry lemma	exact, initial-string, final-string, partial	new lemma, homograph, supplement
etymology search	etymon and word-formation process	exact and partial	source language, word-formation process
full-text search	body of the entries	substring, with automatic highlighting	—
comment search	critical apparatus	controlled list + free text	Americanism, Arabism, addition of sense, etc.

Thanks to the relational structure described above, this combination makes it possible to address specific questions in the history of lexicography: for example, which lemmas were first incorporated in 1884 and were also treated as homographs. In “Etymology search”, filters for grammatical category and lexicographical labels will be added in later phases. “Comment search” provides access to the critical apparatus currently being developed by the team and makes it possible to retrieve, for example, all lemmas with a comment containing a certain word, or all entries annotated as Americanisms in which a given bibliographical source is cited.

The paratexts section provides the full list of the fourteen non-lexicographical texts included in the edition, such as prologues, preliminary remarks, rules, and the table of abbreviations, together with a full-text search box that retrieves the relevant paratext and highlights the search term in context. The bibliographical references section presents an alphabetical list of the works cited in the comments and allows open-ended searches.

This modular architecture, consisting of TEI-encoded data, a normalized relational database, and an independent query layer, allows the interface to expand without substantial rewriting as new tables, such as grammatical category and labels, and new types of comment are added. It thus ensures that e-*DRAE* 1884 functions both as a readable edition and as an environment for exploiting data in the historical study of the lexicon.

4.3 Annotations

The study of the external history of the *DRAE* 1884 and the interpretation of its content are embodied in the critical edition through annotations linked to a selection of dictionary entries considered to be of historiographical, philological, and linguistic interest. These annotations offer observations, explanations, and analyses that contribute to the interpretation of the dictionary text. The archival material used in their preparation consists of the minutes of the Real Academia Española (*Actas*), both those of the plenary sessions and those of the committee for the *Diccionario vulgar*, as well as the institution's index of emendations and additions (*Fichero de enmiendas y adiciones*).

Some annotations concern matters beyond the language itself and focus on relevant historical, social, or scientific developments, while others address philological questions. This is the case, for example, with the inclusion of the lemma *cantiga* [medieval Galician-Portuguese lyric] in the *DRAE* 1884 with stress on the penultimate syllable as the preferred form, in contrast to *cántiga*, which had been recorded in Academy's lexicographical work since the *Diccionario de autoridades*. The preference for the penultimate-stressed form can be traced to a contribution by Leopoldo Augusto Cueto (*Actas*, 6/X/1870), the member of the Real Academia Española who had been entrusted with revising entries relating to literature and who was also the author of an edition and study of the cantigas of Alfonso X.

Systematic comparisons are also carried out with earlier versions of the dictionary and with other lexicographical works, together with an examination of how certain items were treated in later lexicography. For example, the entry *cristofué*, the name of the bird *Pitangus sulphuratus*, offers a significant example of how the *DRAE* 1884 incorporated Americanisms and established definitions that, in this case, remain unchanged up to the current edition. Its inclusion reveals a process of lexicographical fossilization that persists throughout the successive editions of the dictionary. At the same time, the later history of the term reveals the normative weight of the *DRAE*: the word, proposed by José Calcaño in 1873 (Clavería & Hernández, 2025, p. 33), circulated back to Spanish America and was reproduced, often literally, in dictionaries from other regions, suggesting the direct influence of the dictionary of the Real Academia Española on Spanish American lexicography.

The digital edition ultimately seeks to facilitate a comprehensive analysis of the dictionary of the Real Academia Española in order to identify the lexical material that served as the basis for the compilation of the 1884 edition.

5 Conclusion

The model developed for the digital edition of the *DRAE* 1884, situated within the field of the Digital Humanities and grounded in the principles of retrodigitization, transforms a key work in the lexicographical history of Spanish into a structured, machine-processable digital object, suitable for integration into analysis environments, advanced querying, and interoperability. The project seeks to foster closer integration between historical, philological, and humanities

research and the methods and tools of the Digital Humanities. Its aim is for this edition to serve as a replicable model for other historical dictionaries.

The digital dictionary makes its philological foundations accessible through annotations attached to lemmas and senses. These annotations provide information on historiographical, lexicographical, and lexicological aspects that are essential to the interpretation of both the dictionary text and its paratexts. The annotations already published, together with those currently being drafted and reviewed, are intended to offer models for interpreting the various processes of revision, addition, and substitution that shaped the information contained in the entries. The analysis of the rich materials preserved in the archives of the Real Academia Española and the American Academies, still under way, is yielding new data of considerable historiographical interest and broadening our understanding of how the lexical norm of Spanish has been shaped.

The e-*DRAE* 1884 application therefore offers users a fully structured dictionary whose internal units and relations can be retrieved computationally. It provides historians, linguists, philologists, and all those interested in language with an open-access digital resource for exploring the linguistic, cultural, and social ideas underlying a work of notable relevance in the history of the Spanish lexicography.

References

Actas = Real Academia Española (1866-1870): *Actas* (Libro 26). Consultadas en la versión microfilmada del Archivo de la Real Academia Española.

Clavería, G. (2021). La lexicografía académica en la segunda mitad de siglo XIX: tradición e innovación (*DRAE* 1869, *DRAE* 1884, *DRAE* 1899). In M.ª A. Blanco Izquierdo & G. Clavería Nadal (Eds.), *El diccionario académico en la segunda mitad del siglo XIX: evolución y revolución. DRAE* 1869, 1884 y 1899 (pp. 15–56). Peter Lang.

Clavería, G., & Hernández, E. (2021). América en el diccionario académico (*DRAE* 1869, *DRAE* 1884, *DRAE* 1899): primera aproximación. In G. Clavería Nadal & M.ª A. Blanco Izquierdo (Eds.), *El diccionario académico en la segunda mitad del siglo XIX: evolución y revolución. DRAE 1869, 1884 y 1899* (pp. 401–438). Peter Lang.

Clavería, G., & Terrón, N. (2022). La macroestructura de la duodécima edición del diccionario académico (*DRAE* 1884) a la luz de *Lemateca del DRAE*. In Laura Mariottini y Mónica Palmerini (Eds.), *Estudios de lingüística hispánica. Teorías, Datos, Contextos y Aplicaciones* (pp. 1006-1027). Dykinson.

Clavería, G., & Hernández, E. (2025). La norma lexicográfica en el *DRAE* 1884 y las primeras aportaciones de correspondientes americanos. *Boletín de Filología* 60(1), 21-47.

Clavería, G., Freixas, M., Hernández, E., & Torruella, J. (2026). Fundamentos para la edición crítica, hipertextual y digital del *DRAE* 1884 (12.ª edición). *Círculo de Lingüística Aplicada a la Comunicación*, 106(2), 1-12.

Considine, J. (2017). *Historical Dictionaries in Their Digital Futures*. Cambridge University Press.

Diccionario de autoridades = Real Academia Española (1726-1739). Diccionario de autoridades. https://apps2.rae.es/DA.html

DRAE 1884 = Real Academia Española (1884). *Diccionario de la lengua castellana*. Imprenta Gregorio Hernando.

e-*DRAE* 1884 = e-*DRAE* 1884. https://edrae1884.uab.cat/

Durkin, P. (Ed.) (2016). *The Oxford Handbook of Lexicography*. Oxford University Press.

Freixas, M. (2026). Microestructura del *DRAE* 1884: una nueva concepción del artículo lexicográfico. Círculo de Lingüística Aplicada a la Comunicación 105(2026), 43-55. https://www.doi.org/10.5209/clac.108487

Hernando García-Cervigón, A. (2026). Intrahistoria de la configuración del discurso gramatical en el *DRAE* 1884. *Círculo de Lingüística Aplicada a la Comunicación* 106, 57-70. https://www.doi.org/10.5209/clac.108488

Kosem, I.; Lew, R.; Müller-Spitzer, C. (Eds.) (2013). *Electronic Lexicography in the 21st Century*. De Gruyter.

Lemateca = Lemateca del *DRAE*, https://www.draesxix.com/lemateca

Mapa de diccionarios académicos = Instituto de Investigación Rafael Lapesa de La Real Academia Española. 2013. Mapa de diccionarios académicos. https://apps2.rae.es/ntllet/SrvltGUI-LoginNtlletPub

Nalesso, G. (2024). Teoría y práctica de la retrodigitalización de diccionarios: el caso del *Vocabulario de las dos lenguas toscana y castellana*. *Artifara* 24(1), 135-153.

NTLLE = Real Academia Española. Nuevo tesoro lexicográfico de la lengua española. http://ntlle.rae.es/ntlle/SrvltGUILoginNtlle

Pérez, F.J. (2026). El movimiento lexicográfico hispanoamericano del siglo XIX y la creación de las academias correspondientes, *Círculo de Lingüística Aplicada a la Comunicación*, 106(2), 71-78.

Raab, M. (2026). El etiquetaje de las abreviaturas gramaticales en la edición digital del *DRAE* 1884. *Círculo de Lingüística Aplicada a la Comunicación* 106, 27-42. https://www.doi.org/10.5209/clac.1084861

Seco, M. (1988). El léxico hispanoamericano en los diccionarios de la Academia española. *Boletín de la Real Academia Española* 68(243), 85-98.

TEI Consortium. (2022). TEI P5: Guidelines for Electronic Text Encoding and Interchange.

Terrón, N. (2026). La nueva organización de la macroestructura del *DRAE* 1884: los homógrafos. *Círculo de Lingüística Aplicada a la Comunicación* 106(2026), 13-26. https://www.doi.org/10.5209/clac.108485

Contact Information

Margarita Freixas Alás
Universitat Autònoma de Barcelona
Margarita.Freixas@uab.cat

Esther Hernández
Consejo Superior de Investigaciones Científicas
Esther.hernandez@cchs.csic.es

Marija Žarković Eriksson
Universitat Autònoma de Barcelona
Marija.Zarkovic@autonoma.cat

Acknowledgements

Project “Modelo de una edición hipertextual del DRAE 1884. Bases teóricas para la transferencia digital de un diccionario” (PID2022-136666NB-C21), Project “Modelo de una edición hipertextual del DRAE 1884. La presencia de América en el diccionario” (PID2022-136666NB-C22), and the Grupo de Lexicografía, Diacronía y ELE Research Group (AGAUR, 2021 SGR 00157).

Ellert Thor Johannsson, Jóhannes B. Sigtryggsson &
Thordis Ulfarsdottir

UNLOCKING A NINETEENTH CENTURY DICTIONARY

A Closer Look at Konráð Gíslason's *Dönsk orðabók* (1851)

Abstract This article discusses an important Icelandic lexicographic work published in the nineteenth century. *Dönsk orðabók með íslenzkum þýðingum* was compiled and edited by Konráð Gíslason (1808-1891), an Icelandic scholar and language purist. The printed dictionary is described in some detail, with particular attention to its structural characteristics and the ways in which purist ideals shaped many of its entries. Loanwords were systematically avoided and explanatory paraphrases were preferred whenever no suitable native Icelandic equivalent was available. The discussion then turns to recent efforts to produce a digital edition of the dictionary. We account for the workflow implemented in order to turn this complicated printed book into a functioning searchable edition where the content has been marked-up and XML-encoded. The process illustrates some of the issues involved and addresses challenges encountered when producing the new digital edition of the dictionary.

Keywords Lexicography; language purism; digitisation; historical lexicography

1 Introduction

Konráð Gíslason's *Dönsk orðabók með íslenzkum þýðingum* 'A Danish Dictionary with Icelandic Translations', published in 1851, occupies a distinctive place in the history of Icelandic lexicography. Emerging at a time in the history of Iceland when modern society was taking form, it sought to strengthen Icelandic linguistic autonomy and to enable Icelanders to engage with Danish administrative, educational, and scholarly discourse. At a time when Danish was central to Icelandic official life and intellectual exchange, the dictionary addressed a pressing need as it enabled Icelanders to understand Danish texts while simultaneously promoting Icelandic vocabulary.

This article describes some of the main characteristics of the dictionary and its idiosyncratic features, as well as the influence of linguistic purism on the work. In addition, it presents an ongoing project to create a digital edition of the dictionary and some of the challenges associated with the process of turning a complex printed book into a functional online resource.

The article is structured as follows: Section 2 examines the background and historical context of the dictionary and situates it within nineteenth-century Icelandic scholarly efforts. In section 3 we describe the formal characteristics of the dictionary and discuss its relationship to language ideology and purism. Section 4 describes the challenges involved in transforming the printed book into a modern digital resource. Finally, there are some concluding remarks.

2 Background and historical context

Konráð Gíslason's (1808–1891) dictionary marks a significant shift in Icelandic lexicography. Earlier lexicographic efforts in the seventeenth and eighteenth centuries had focused primarily on Old Icelandic or Latin, often with antiquarian or scholarly aims. Major works such as Guðmundur Andrésson's *Lexicon Islandicum* (1683) and Jón Árnason's *Nucleus latinitatis* (1738) were largely rooted in learned traditions rather than in the practical demands of modern

communication. By the nineteenth century, however, social and political developments had created new linguistic needs. The publication of Björn Halldórsson's Icelandic-Latin-Danish dictionary in 1814 had already shifted attention towards modern Icelandic vocabulary, but Gíslason's Danish-Icelandic dictionary represented a more direct engagement with contemporary Danish-Icelandic interactions.

The compilation of this dictionary was an extremely ambitious undertaking (cf. Kristjánsson, 2003, pp. 203–210). Konráð Gíslason was a scholar and Icelandic intellectual, mostly active in the community of learned Icelanders in Copenhagen, where they sought higher education and engaged in scholarly activities. Gíslason thought it vital that Icelanders should be able to convey any concept in their own language and was determined to compile a dictionary to facilitate the use of refined and accurate Icelandic parallel with the dominant language of administration and scholarly endeavours, namely Danish (cf. Kristjánsson, 2003, pp. 203–210).

Gíslason was influenced by the contemporary Danish lexicographic tradition and drew heavily on some well-known published works. The Danish wordlist and source language material is greatly inspired by Danish dictionaries, especially C. Molbech's *Dansk Ordbog* ('Danish dictionary') (1833), while also incorporating foreign-word material from L. Meyer's *Fremmedord-Bog* ('Loanword dictionary') (1844) (cf. Kvaran, 2013).

3 The printed dictionary of 1851

Work on the dictionary started in 1840 and after approximately ten years it was ready to be published. The printed book comprises 596 large-format pages printed in a dense three-column layout and contains 52,620 headwords. It was published in 1100 copies. The dictionary is organised alphabetically by the Danish headwords. Typographically, the volume uses both Gothic and Latin typefaces: the Danish words are printed in Gothic script, while Icelandic equivalents, translations and explanations are presented in Latin script. Variations in font size and typeface are also used for meta-information, emphasis, quotations, and structural differentiation within entries. This is displayed by the partial page shown in Figure 1.


118 fladtrind—Fletteværk.

fladtrind (sjaldh.), kringlóttur.
fladtrykt, [sb. trykke], flattur út (eiginl. með því að fergja); flatur: Jorden er noget f. ved Polerne.
Flag, ft. eins, kl., veifa (t. a. m. á siglutrje): Heise (stryge) F. Sb. Vimpel.
Flage, ft. -r, sk., 1) [sb. flaki], a) flaki, fleki. Sb. Isflage, Kolleflage. — b) [sb. flaga], snjóflyksa í skæðadrífu: I Tovcir falder Sneen ned i Flager. Sb. Sneeflage.
Flageled (fornt: Flagreled) flakahlið.
Flagfald það að taka ofan veifur.
flagge (har), h. s. 1, hafa uppi veifu: Alle Slibe i Havnen flaggede.
Flaggermus leðurblaka *vespertilio*. Sb. Aftenbakke, sem þykir danskara.
Flagknap hnúður efst á veifustöngiani.
Flagmager, ft. -e, kk., sá sem býr til veifur.
Flagmand (í tali sjóferðamanna), skip sem hefur uppi veifu.
flagre (har), h. s. 1, [sb. flögra], flögra, flökta, blakta, flaksast: F. om eða omkring. F. med Vingerne bada vængjum.
Flanevæsen flennuæði.
*Flanke, ft. -r, sk., hlið á fylkingu: Falde Fjenden i Flanerne ráðast á hlið fylkingar, fyrrum: koma e-m í opna skjöldu.
*flankere (har), h. s. 1, skjóta með endilöngum hliðum; reisa hliðvígi; slarka; svalla.
flaske, ás. 1, haft í tali í orðatiltæk. f. sig takast, heppnast: Det vil ikke f. sig.
Flaske, ft. -r, sk., flaska; glerpyttla, glas, (því Flaske er yfirgripsmeira, enn flaska, og er hæði haft um þess konar smá gler-ílát, sem vjer köllum glös, og svo um ílát, sem ekki eru úr gleri, og ekki líkjast því, sem Íslendingar kalla flösku): En F. Viin. En F. Øl ein flaska af öli. — b) (fyrrum), eins konar laf á kvennreyju.
Flaskebrev blað sem farmenn rita um ástand sitt og kasta fyrir borð í flösku svo það skuli finnast.
Flaskebug bumba á flösku.
flaskedannet, [sb. danne], lagaður eins og flaska.
Flaskefoder hólfalaupur til að flytja í flöskur.
Fleerstavelsesord ord sem er meir enn eitt atkvæði.
Fleertal meiri hluti: Han havde Fleertallet paa sin Side. — b) fleirtala ords.
fleertydig, ft. -e, e., (t. a. m. um ord), tvíræður eða margræður.
Fleertydighed, sk., það að geta þýðst ýmislega.
Flekke, ft. -r, sk., býr sem er minni enn Stad og meiri enn Landsby.
Fleng ekki haft nema í orðatiltæk. i F. í graut, hvað innan um annað, af handa hófi: Bruge flere Ting i F. Læse i F. Give i F.
Flenge, ft. -r, sk., stór skurður (í hold), svöðusár: En dyb F. — b) flikki: En F. Kjød.
flenge, ás. 1, [flengja], a) hýða, flengja: F. Huden af een afhýða e-n. — b) rista (t. a. m. með hníf).
flense, ás. 1, [flennsa], skera hval.
Flenser, ft. -e, kk., hvalskurðarmaður.
Flensning, sk., hvalskurður.
flere og fleer [= fleiri], miðstig af mange; sjá mange.

FIGURE 1 A partial page from the dictionary showing three columns as well as different typesets and font sizes

The dictionary has several noteworthy characteristics. Gíslason was influenced by the language purism movement and had a strict view of providing Icelandic definitions of the Danish lexical items (Ottósson 1990, pp. 100–103). The dictionary therefore contains only a small number of loanwords. He frequently expanded entries with semantic explanations, stylistic commentary, historical citations, and carefully selected Icelandic neologisms or revived archaic forms. The dictionary thus functioned both as a means to facilitate understanding and as an instrument for shaping modern Icelandic vocabulary. The following are four examples of paraphrased translations of lemmas that do not have a direct Icelandic equivalent. The Icelandic explanations from the dictionary have been translated to English by the authors.

- *barbeerskilt* 'barber sign' brass bowls hung outside barbers' shops to show where barbers live or operate
- *dessert* 'dessert' supplementary food (e.g. fruit, pastries, or the like) eaten last
- *fritænker* a self-reliant person in matters of religion (i.e. one who does not believe he needs to accept what Scripture says)
- *gadestryger* a kind of wastrel (idler) who wanders about the streets and cannot be bothered to do any work

Entries frequently contain various types of additional information. The entry *abekat* 'monkey' (lit. ape-cat) has the following comment "*apaköttur* is not correct Icelandic". Gíslason's puristic observations also extend to Danish vocabulary. For instance, the adjective *aarsagelig* 'causal' has the following note "búið til á seinni tíðum, til að komast hjá hinu útlenzka orði causal" 'coined in later years [in Danish] to avoid the foreign word causal']. In most cases his puristic ideas are not so explicitly stated but shine through in his treatment of many of the headwords.

Many entries contain additional information of various kinds. Certain aspects of Danish culture had no direct counterparts in Icelandic society and therefore did not have any Icelandic equivalents. In such cases Gíslason found it fitting to give additional information. An example of that is the Danish word *adressecalender* 'register of abodes', where he adds for further explanation 'in order to be able to locate people in cities'.

When the Danish word is translated by an Icelandic equivalent Gíslason often prefers to draw on material from medieval Icelandic literature, manuscripts, and earlier dictionaries rather than promoting neologisms. A good example is the headword *krigsgud* 'god of war'. This word is translated with three Icelandic equivalents drawn from three different medieval works that he abbreviates in parenthesis: "*bardagaguð* (Al.), *orustuguð* (Trójum.), *vígaguð* (Sn. E.)", i.e. 'battle god', 'combat god', 'god of slaying'. The first word is taken from the thirteenth-century *Alexanders saga*, the second from the *Trójumanna saga* (dated to the same century), and the third from Snorri Sturluson's *Edda* (the *Younger Edda*, also from the same century).

Overall, the dictionary contains a wealth of information about the contemporary language and prevailing attitudes towards language at the time. Björn M. Ólsen, who later became the first president of the University of Iceland, summarised its significance in an article about Gíslason's life some years later:

> No book that has appeared in this language has done as much to eradicate Danish contamination from our tongue as this one. . . . For 40 years now, it has taught young and old alike not only to understand Danish, but also to translate it into pure Icelandic. (Ólsen, 1891, p. 66)

4 Digital Edition

From the discussion above it is clear that the dictionary of Konráð Gíslason is an important source for nineteenth-century Icelandic vocabulary. Since its publication in 1851 the book has never been reprinted, so today it is rare and relatively little known. It was therefore considered important to make this resource digitally accessible and searchable for researchers and anyone else interested in 19th century Iceland.

Dr. Auður Hauksdóttir, former professor of Danish at the University of Iceland, had previously initiated the work of digitising the text. The results were relatively error-free, OCR text files with each letter of the alphabet processed separately. A dedicated working group at the Árni Magnússon institute was subsequently established in 2024.

Although the OCR process was quite detailed it did not preserve the distinction between Gothic and Latin script or other typographic distinctions, such as variations in font size and expanded type, all of which are used in the dictionary's microstructure. It was therefore necessary to further process the initial OCR files. This included:

- Text normalisation and cleaning
- Merging of all individual letter files
- Ensuring that entries spanning columns/pages were unified
- Creation of an index file
- Conversion of text into simplified XML
- Division of entries into subcomponents
- Making a script that identifies metadata
- Running a script that identifies sources
- Running a script that analyses language (Danish vs. Icelandic vs. Latin)

Figure 2 shows the difference between the initial XML-document and the first cleaned-up version.



FIGURE 2 The OCR text before (left) and after (right) the initial processing

The most challenging part of this process was separating the two languages. In order to fully process the digital text, it was necessary to establish which parts of the text were in Danish and which were in Icelandic. This required reading and matching every word with known Icelandic and Danish forms, using LT tools wherever possible. For the Icelandic part this was relatively straightforward as the information needed is found in the *Beygingarlýsing íslensks nútímamáls*

'The Database of Modern Icelandic Inflection' (DMII, containing inflected forms of 356.000 Icelandic words). It was only necessary to account for certain regular spelling discrepancies, e.g. that *je* = *é* and *z* = *s*, reflecting differences between nineteenth- and twenty-first-century spelling conventions.

The Danish part of the dictionary proved more challenging. An existing lexical resource, ISLEX, a multilingual Icelandic-Nordic dictionary that includes Danish, provided some of the necessary information. All Danish words were extracted from the ISLEX database and supplemented with a Danish word-form generator, which produced all theoretically possible inflectional forms. Some additional rules were defined for language identification, such as classifying words that include the letter *ø* as Danish, and words containing *ð* as Icelandic. Contextual markers, such as Danish *denne* or *dette* and other language exclusive function words were also included in the script.

This methodology was largely successful in separating the two languages. However, it was still not possible to distinguish between certain ambiguous words, i.e. words that are not formally different in Icelandic and Danish, e.g. common words like *mig* 'me', *var* 'was', *af* 'of', etc. Such ambiguous words were marked separately in the XML file for manual revision. It was also decided to mark Latin forms specifically. Even though Latin is neither a source nor target language, it appears frequently in the dictionary, in particular in entries involving plants, animals and diseases.

All the information was marked up in an XML file. Samúel Thórisson, a software developer and member of the working group, designed an editorial interface for further processing of the data. This included a simple way of providing additional markup or changing the markup of the already processed items.

Once the XML file had been established and streamlined it was possible to add some metadata that had been identified. Figure 3 illustrates the editorial interface where different types of markup are displayed in different colours. Icelandic is represented in blue script and Danish in red. Ambiguous words are highlighted in green, while words that do not fit into either language are highlighted in yellow, e.g. words with alternative or erroneous spellings or that could not be classified automatically. Red text in a darker yellow background indicates grammatical information.



FIGURE 3 A section of the text with highlighted words that need manual post-processing

This interface also allows for the markup of some additional linguistic information, such as *fornt* 'medieval', *hlekkur* 'link' (i.e. inserting a link to another headword), *hornklofi* 'brackets' (for extra information given by Gíslason, usually in brackets), *málsháttur* 'proverb', *samber* 'cf.' (for cross-referencing), *sjaldh* 'rare' and *tilvitnun* 'quotation' (cf. Figure 4).

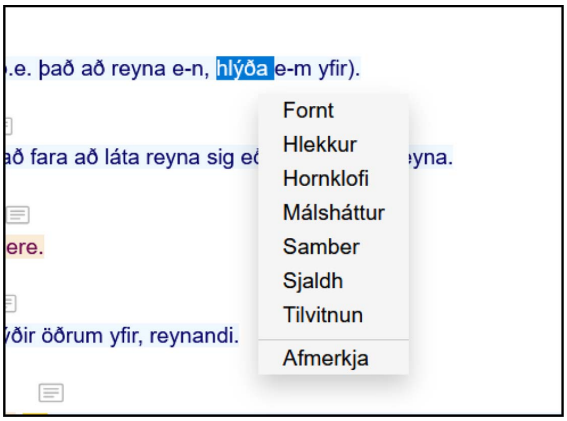


FIGURE 4 The editorial right-click menu for quick markup of selected word or text

Having the option to add such information by right-clicking on the relevant word greatly facilitated the post-processing work.

After the initial automated processing, error rates could be assessed. Estimated mean error rates are based on a sample of 100 entries (852 words), *sude – sukkre*. This section was typical for the text, having both simple and complex entries. In total it contained 40 errors. Five errors were wrong identification of Latin words, whereas the remaining 35 errors were words highlighted green or yellow, (ambiguous Icelandic or Danish) that needed to be processed by a human.

The markup and correction phase turned out to be quite time-consuming, but an anonymous donation to the project allowed us to hire a student part time for several months to assist with the necessary work. At the time of writing, approximately 80% of the dictionary has been corrected and marked up with additional information.

5 Conclusion

Konráð Gíslason's Danish-Icelandic dictionary is an important record of nineteenth-century Icelandic and contains far more information than the basic lexicographic data alone. Its digitisation will greatly increase the accessibility of this material and creates new possibilities for research and analysis. At the same time, the project has demonstrated that careful and systematic annotation is essential in order to exploit the material to its fullest potential. At this stage, considerable work remains to be done. Future steps include defining additional data types, correcting remaining errors, proofreading selected sections, and further processing the Icelandic text, for example through the creation of a lemmatised index. The long-term goal is to make Gíslason's dictionary searchable and accessible through a dedicated website hosted by the Árni Magnússon Institute for Icelandic Studies.

References

Andrésson, G. (1683). *Lexicon Islandicum*. Havniæ.

Árnason, J. (1738). *Nucleus latinitatis*. Hafniæ.

Bjarnadóttir K. & Hlynsdóttir, K. (Eds.). *DMII = Database of Modern Icelandic Inflections/Beygingarlýsing íslensks nútímamál*. Stofnun Árna Magnússonar í íslenskum fræðum. Retrieved May 28, 2026, from https://bin.arnastofnun.is/

Halldórsson, B. (1814). *Lexicon Islandico-Latino-Danicum Biörnonis Haldorsonii. Biørn Haldorsens islandske Lexikon. Ex Manuscriptis Legati Arna-Magnæani cura R. K. Raskii editum*. I–II. Havniæ.

Kristjánsson, A. (2003). *Síðasti Fjölnismaðurinn. Ævi Konráðs Gíslasonar*. Skrudda.

Kvaran, G. (2013). Orðabókarstörf Konráðs Gíslasonar. *Orð og tunga*, 15, 95–119.

Meyer, L. (1844). *Fremmedord-Bog, eller kortfattet Lexikon over fremmede, i det danske Skrift- og Omgangs-sprog forekommende Ord, Konstudtryk og Talemaader; tilligemed de i danske Skrifter mest brugelige, fremmede Ordforkortelser. Anden, forøgede og forbedrede Udgave*. J.H. Schubothes Boghandling.

Molbech, C. (1833). *Dansk Ordbog indeholdende det danske Sprogs Stammeord, tilligemed afledede og sammensatte Ord, efter den nuværende Sprogbrug forklarede i deres forskiellige Betydninger, og ved Talemaader og Exempler oplyste*. Paa den Gyldendalske Boghandlings Forlag.

Ólsen, B. M. (1891). Konráð Gíslason. *Tímarit Hins íslenzka bókmentafélags*, 12, 3–96.

Ottósson, K. G. (1990). *Íslensk málhreinsun. Sögulegt yfirlit*. Íslensk málnefnd 6.

Ulfarsdottir, T. (Ed.). *ISLEX* [an online multilingual dictionary between modern Icelandic and six Scandinavian languages]. Stofnun Árna Magnússonar í íslenskum fræðum. Retrieved May 30, 2026, from https://islex.is/

Contact information

Ellert Thor Johannsson
The Árni Magnússon Institute for Icelandic studies
ellert.thor.johannsson@arnastofnun.is

Jóhannes B. Sigtryggsson
The Árni Magnússon Institute for Icelandic studies
johannes.b.sigtryggsson@arnastofnun.is

Thordis Ulfarsdottir
The Árni Magnússon Institute for Icelandic studies
thordis.ulfarsdottir@arnastofnun.is

Acknowledgements

We would like to thank the following people for their contribution at various stages of the project: IT-specialist Samúel Thórisson, professor emeritus Margrét Jónsdóttir, professor emeritus Auður Hauksdóttir, Björn Þorvarðarson, Sofus Matthias Skjærlund and Kolbrún Ýr Sigfúsdóttir.

Agnes Kim

DIGITAL METALEXICOGRAPHY AS A TOOL FOR HISTORICAL (CONTACT) LINGUISTICS

Abstract This contribution presents a data collection on 'Slavic in Viennese', i.e. (glottonyms for) Slavic languages, (ethnonyms for) their speakers as well as single linguistic phenomena in German varieties typical for Vienna, that comprises 6,406 articles on 268 lemmas from 30 dialect dictionaries published between 1800 and 2020. In the contribution's first part, corpus compilation and the underlying data structure are described and results of existing studies on this data collection are discussed. So far, these have allowed for the identification of reception networks between the dictionaries and revealed the reconstruction of stereotypes regarding speakers of South Slavic languages in them. Inspired by these results, the second part of the contribution presents an approach towards the use of the data collection (and similar ones) for the description of integration, variation and change of (West Slavic) loanwords in 'Viennese'. By means of exploratory cluster analyses, three main clusters of lemma-meaning pairs are identified: one general that includes forms and meanings used throughout the complete observation period, one of historical forms and meanings that were common in the 19th century, and one of innovations from the second half of the 20th century. Building on this and the exemplary comparison of forms and meanings of three Czech loanwords (*Kaluppe, Kolatsche, pomali*), the contribution suggests that most similar loanwords must have been borrowed before the 19th century, thereby demonstrating the usability of the approach for historical (contact) linguistics.

Keywords Viennese; Slavic languages; Czech; metalexicography; language contact; loan words

1 Motivation and goals

Corpora of historical dialects of German are generally too small to capture (semantic) integration, variation and change of lexemes with lower frequency over time. For many of these varieties, however, dialect dictionaries are available and cover most of the 19th and 20th centuries. This is the case for 'Viennese' (for a definition see Section 2) which this contribution focuses on. It presents a systematic metalexicographical approach to the exploitation of historical dialect dictionaries as (additional) sources in historical sociolinguistics that may help to close the gap in data availability and that goes beyond looking up single words. Instead, it relies on the correlation of information given on a segment of the lexicon by all dictionaries of 'Viennese'. This approach allows to trace developments regarding this segment's lexicographical (re-)presentation and – indirectly – also the development of the lexicon and lexemes. The segment of the lexicon selected is 'Slavic in Viennese' (for a definition see Section 2), which is of special interest in the context of Vienna due to the long-term multilingualism in the capital of the Habsburg monarchy. Therefore, Slavic loanwords have often been described as typical for German in Vienna and even Austria in general (see, e.g., Newerkla, 2013).

This contribution pursues two goals: In Section 2, the corpus, i.e., the digital structured collection of dictionary entries referring to 'Slavic in Viennese' that was built for Kim (2022b) is presented. In the context of Kim (2022b), 'Viennese' is understood as the sum of all lects (varieties, registers, codes, etc.) that are described as part of the German diasystem and typical for the city of Vienna in its historically varying borders by linguists or non-linguists. 'Slavic in Viennese' is defined as a cover term designating all (glottonyms for) Slavic languages, (ethnonyms for) their speakers as well as single linguistic phenomena in 'Viennese' that are attributed to language contact with Slavic languages (see Kim, 2022b, pp. 11–12). These definitions are of emic nature and allow to include sources, i.e., 'Viennese' dictionaries, as well as phenomena, i.e., lemmas

representing 'Slavic in Viennese', without (socio-)linguistic and etymological pre-evaluation. This again is a prerequisite for analysing the discursive construction of the research object, which was the main goal pursued by Kim (2022a,b).

Section 3 explores the possibilities for analysing the data collection compiled for Kim (2022b) in historical socio- and contact linguistics. In this context, the focus is on the (semantic) integration, variation and change of West Slavic loanwords in 'Viennese'. This of course requires a linguistically less naive approach and detailed etymological and contact linguistic evaluation, classification and selection of the material under analysis. Hence, the scope of '(West) Slavic in Viennese', as it is described in Section 3, is narrower than in Section 2. Additionally, Section 3 considers only loanwords from West Slavic languages (i.e., Czech, Slovak, Polish, Kashubian, Sorbian). In fact, most of them turn out to be of Czech provenience, which can be historically explained by the long-term German-Czech multilingualism in Vienna that certainly peaked in the second half of the 19th and early 20th centuries, when approximately 25% of all inhabitants of Vienna were born in Moravia and Bohemia (i.e., present-day Czechia), and another 25% were likely descendants of migrants from these crown lands (see, e.g., Glettler, 1972; Newerkla, 2013; Kim, 2021).

2 'Slavic in Viennese' – Corpus description

The corpus presented here was primarily compiled in order to allow the description and analysis of the (re-)construction of 'Slavic in Viennese' in dictionaries of 'Viennese' from the 19th to early 21st centuries in Kim (2022b). It is a digital, structured collection of dictionary entries or lexicological information on 'Slavic in Viennese' (in a broader sense) from 30 printed dictionaries and lexicological studies (e.g., Steinhauser, 1978) of 'Viennese' that were published between 1800 (Sonnleithner, 1800) and 2020 (Brauer, 2020).[1] The resulting corpus consists of 6,406 single dictionary articles or snippets from lexicological studies that contain information on 268 base words (hereinafter: lemmas, e.g., *Jause* 'snack'), or 1,011 derivations, composites or phrasemes formed from/with these root words (hereinafter: forms, e.g., *Jausengegner* 'easy opponent', *jausnen* 'to snack'), and all their 1,566 meanings documented in the processed works (see Kim, 2022b, p. 56).

The first step in corpus creation was to compile a list of lemmas. To this end, four dictionaries of Viennese that pay special attention to etymological information were identified (Jakob, 1929; Schuster & Schikola, 1996; Hornung & Grüner, 2022; Sedlaczek, 2011) and searched for (abbreviations) of glottonyms for Slavic languages (e.g., *tschech**, short for *Tschechisch* 'Czech') or toponyms for (historical) political entities inhabited by speakers of Slavic languages (e.g., *böhm**, short for *Böhmen* 'Bohemia'). This search yielded 624 dictionary articles, which belong to the above-mentioned list of 268 lemmas that served as the starting point for the following steps (see Kim, 2022b, pp. 46–48).

For the compilation of the corpus, 30 dictionaries (including six second editions) published between 1811[2] and 2020 were searched for dictionary articles (n = 6,406) with information on the 268 lemmas or a form derived from or composed with the lemma (n = 1,011). The information was manually entered into a hierarchical XML-database using the browser-based dictionary writing system Lexonomy (Měchura, 2017; Jakubíček et al., 2022),[3] that in a first step represented a collection of information on forms, meanings and etymology belonging to a

[1] The latest dictionary of 'Viennese', Sedlaczek (2023) could not be included.

[2] Sonnleithner (1800) was not included into the data base, since it does only contain lemmas beginning with <A> and <B>. Therefore, the earliest dictionary included is Sonnleithner (1811).

[3] The final data collection is publicly available at https://www.lexonomy.eu/#/SliW2

certain lemma of the lemma list from each source. In a second step, forms, meanings and etymologies from different sources were manually mapped to each other, which resulted in the data scheme depicted in Figure 1. Element names were based on the TEI P5 conventions, but for practical reasons the XML scheme could not fully comply to them (Kim, 2022, pp. 51–52).

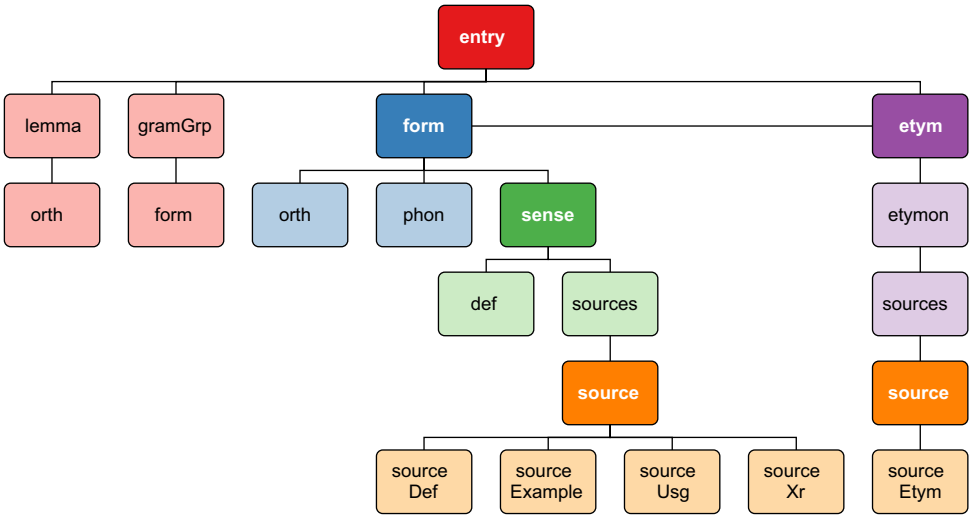


FIGURE 1 Data scheme of the data collection 'Slavic in Viennese' (see Kim, 2022, p. 53)

The colours in Figure 1 indicate the information types that are represented in the XML elements of the data collection. The <source> elements (orange) are those that contain the literal information from the original sources, i.e., meaning definitions <sourceDef>, example sentences <sourceExample>, usage information <sourceUsg>, cross references <sourceXr> and etymological information <sourceEtym>. All semantic and usage information is embedded into <sense> elements (green). Besides a definition <def> these contain all dictionary articles that include information on the respective meaning. Semantic categorization and mapping of meaning definitions onto each other were carried out by the author. The meanings are attributed to forms <form> (blue), and these again to lemmas <entry>. Etymological information <etym> (violet) can refer to either the <entry> (e.g., if the etymology of a lemma such as *Jause* 'snack' is explained) or the <form>(e.g., if the word formation processes of a compound such as *Jausengegner* 'easy opponent' are explained; for closer information see Kim, 2022b, pp. 52–56).

In earlier studies (Kim 2022a,b), which were inspired by forensic dictionary analysis as described by Coleman & Ogilvie (2009), this data collection or parts of it have been analysed with exploratory statistical means (e.g., clustering and network algorithms) in order to identify citation networks between dictionaries or to deconstruct cultural stereotypes conveyed through them. Using these methods, Kim (2022b) described the similarities of dictionaries of 'Viennese' regarding their (re-)presentation of 'Slavic in Viennese' based on the following (quantifiable) measures: (a) number of lexical-semantic or etymological information represented, (b) publication date and/or order, and (c) degree of agreement with other sources. She identified a cluster of 13 dictionaries (including four second editions) that contain a lot of specific information on 'Slavic in Viennese'. Except for Sedlaczek (2011), all of them were published in the 20th century, with the earliest being Gartner (1902a,b,c; 1903a,b; 1904b), a dictionary of loan words in 'Viennese'. All of these dictionaries also share a high density of etymological information, which cannot be found in the other dictionaries (see Kim, 2022, pp. 75–78). Additionally, Kim (2022, pp. 80–89) was able to identify reception networks between these

dictionaries. Regarding lexical-semantic and to a lesser degree also etymological information on 'Slavic in Viennese', Jakob (1929) seems to have had great influence especially on Wehle (1980), which in turn served as an important source for Teuschl (1990). For etymological information, Wehle (1980) also consulted Steinhauser (1978). Hornung & Grüner (2002) included a wider range of (unpublished) sources and, together with Teuschl (1990) became one of the main sources for Sedlaczek (2011).

Kim (2022a) used a subset of the data on 'South Slavic in Viennese' to identify diachronical strata of meanings of loanwords from South Slavic languages or names for these languages and/or their speakers (n = 182 meanings of 100 forms of 24 lemmas). She identified two important layers of lemma-meaning pairs, a historical and a younger one. The lemma-meaning pairs belonging to the first, historical layer already appear in 19th century dictionaries and in Jakob (1929) the latest. In contrast, almost all lemma-meaning pairs in the second layer are described by Hauenstein (1972, 1978) for the first time and thus represent the younger stratum of 'South Slavic in Viennese' (see Kim, 2022a, pp. 297–301). The two layers do not only differ diachronically, but also regarding the information encoded in the meaning definitions. While those on the historical layer focus on rather neutral semantic information, dictionary articles on lemma-meaning pairs from the younger layer reproduce stereotypes regarding the sociodemographic status and other characteristics of speakers of South Slavic languages (see Kim, 2022a, pp. 313–314).

3 Integration, variation and change of West Slavic loanwords in 'Viennese'

In Section 3, a similar approach like in Kim (2022a) is applied to a larger subset of the data, namely all loanwords from West Slavic languages (n = 411 meanings of 286 forms of 103 lemmas). Only lemmas were included, which are (very) likely to have been borrowed from West Slavic languages, whereas others with disputable etymology (e.g., *Bissgurre* 'quarrelsome woman; misgurnus [fish]', which can either be explained as a loan from Czech *piskoř* 'misgurnus' or as related to Middle High German *Gurre* 'inferior horse'; see Kim, 2024) or no Slavic etymology (e.g., *Böhme* 'Bohemian'; see Bauer, 1976) were excluded. Additionally, the number of dictionaries was reduced to 23 by selecting the latest edition of each dictionary in case of more than one available in the data collection. The data set analysed therefore consists of 9,453 data points coded as 1 (information on lemma-meaning pair existing in source) or 0 (information on lemma-meaning pair not existing in source).

The main goal is to explore the use of the data set to gain information on the integration, variation and change of West Slavic loanwords in 'Viennese'. As an auxiliary, 'integration, variation and change' is operationalized as the number of lemma–meaning pairs derived from or formed with a certain loan word and their documentation in the data set over time. For interpretation reasons, the second editions of dictionaries used in the analysis are treated as if they had been published in the years of the first editions (see, e.g., Figure 2).

The first step in the analysis was an exploratory hierarchical cluster analysis performed with hclust (binary distance metrics, ward.D2 clustering method) of the stats library in R. For quality assessment of cluster solutions, the WeightedCluster library (Studer, 2013) was used. Generally, the cluster quality is quite low. Based on the Chsq measure and heuristic comparison of various cluster solutions, an eight-cluster solution was considered the most appropriate. Figure 2 displays the distribution of all lemma-meaning pairs (y-axis) over these clusters and over time, i.e., over chronologically ordered sources (x-axis). The clusters can be characterised as follows:

The two largest clusters 2 and 6 consist of 74 lemma-meaning pairs each and additionally share that their lemma-meaning pairs are also reported by 21st century dictionaries other than Sedlaczek (2011). The meanings attributed to cluster 6 are represented in 8 dictionaries on average (*min* = 1, *max* = 10, *median* = 7) that stem from the complete time period in focus. The meanings in cluster 2 are part of only 2.5 dictionaries on average (*min* = 1, *max* = 6, *median* = 2). The most important dictionaries for meanings in this cluster are Teuschl (2007), Hornung & Grüner (2002) and Sedlaczek (2011). The lemma-meaning pairs in these clusters are likely to be the ones used or at least known in 'Viennese' until the recent days with cluster 6 representing the historical core, while cluster 2 comprises innovations of the second half of the 20th century.

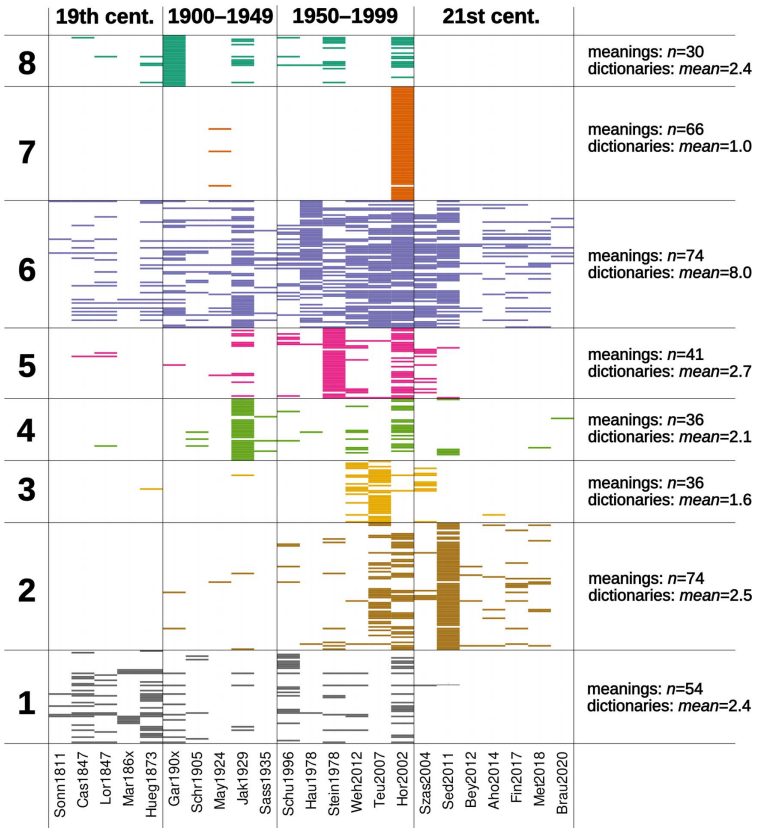


FIGURE 2 Representation of lemma-meaning pairs (y-axis) over chronologically ordered dictionaries (x-axis) and sorted by clusters

The third largest cluster is cluster 7 (*n* = 66). The lemma-meaning pairs in this cluster are, however, mainly reported by Hornung & Grüner (2002) (*mean* = 1, *min* = 1, *max* = 2, *median* = 1) and are thus likely to be of lesser importance in the description of 'West Slavic in Viennese' in general. Of greater interest is certainly cluster 1, which contains 54 lemma-meaning pairs mainly included in dictionaries of the 19th century as well as in Hornung & Grüner (2002) and Schuster & Schikola (1996). The latter dictionary explicitly claims to describe 'Viennese' under the Habsburg rule. The lemma-meaning pairs attributed to cluster 1 can therefore be described as the historical stratum of West Slavic loanwords in 'Viennese'.

The remaining four clusters are smaller and mainly characterised by the importance of single dictionaries. Cluster 5 comprises 41 lemma-meaning pairs that are mainly represented by Steinhauser (1978) and, secondarily, Hornung & Grüner (2002) (*mean* = 2.7, *min* = 1, *max* = 5, *median* = 3). The 36 meanings in cluster 4 were mainly described by Jakob (1929) and later

by Hornung & Grüner (2002) (*mean* = 2.7, *min* = 1, *max* = 5, *median* = 3), whereas the 36 meanings attributed to cluster 3 are astonishingly not included in Hornung & Grüner (2002), but mainly in Teuschl (2007) (*mean* = 1.6, *min* = 1, *max* = 5, *median* = 1). The smallest cluster is cluster 8, which contains only 30 lemma-meaning pairs, which are mainly described by Gartner (1902a,b,c, 1903a,b, 1904b) (*mean* = 2.4, *min* = 1, *max* = 7, *median* = 2).

To sum up, we can identify three clusters that qualify as especially interesting in describing and comparing the integration, variation and change of West Slavic loanwords in 'Viennese': Cluster 1 comprises the historical layer of meanings of loanwords that were likely borrowed before the 19th century, but hardly used in these meanings beyond the 19th century. In contrast, the meanings in cluster 6 also belong to loanwords that were borrowed before the 19th century, but are used or at least still known in the 21st century. The meanings attributed to Cluster 2 are innovations of the second half of the 20th century.

For closer scrutiny, lemmas whose meanings are assigned to at least two of the three central clusters 1, 2 or 6 will be subjected to closer scrutiny. These loanwords have meanings and/or forms that are used in 'Viennese' throughout the complete focus period, others that were mainly used in the 19th century and/or younger ones that developed in the second half of the 20th century. Table 1 provides an overview of the 28 lemmas (of 103 lemmas, i.e., 27% of all 'West Slavic in Viennese' lemmas) that have meanings in at least two of the central clusters.

TABLE 1 Lemmas with meanings in at least two of the central Clusters 1, 2 or 6

Lemmas with meanings in clusters 1+6	Lemmas with meanings in clusters 2+6	Lemmas with meanings in clusters 1+2+6
n = 7	*n* = 13	*n* = 8
Brimsen, kralowat, Pawlatsche, peitschen, petschieren, Powidel, pritsch	*Gatsche, Graupen, Gurke, Lepschi, Mamlas, Muckerl, Nowak, Palatschinke, Polka, Wabe, Wuchtel, Schestak, Schöps*	*Kaluppe, Kolatsche, Kren, pomali, Quargel, Robot, Taschlowitz, Zeiserl*

Table 2 compares the form-meaning pairs of three selected lemmas with meanings in all three central clusters. All of them are well documented Czech loanwords that are also known in dialects outside of Vienna (*Kaluppe* 'shabby hut' < Czech *chalupa* 'hut; weekend house', see Kim, 2025; *Kolatsche* 'filled pastry' < Czech *koláč* 'small, round cake', see Pfeifer, 1993; *pomali* 'slow' < Czech *pomalý* 'slow', see Groschopf, 1983).

It is striking that the form-meaning pairs which are assigned to cluster 6 and thus used (or known) throughout the full 19th, 20th and the early 21st centuries are equal to the lemmas, i.e., the simplex forms that also exist in Czech. Additionally, the meanings are quite close or identical to the meanings in the current form of the donor language. In case of *pomali* and Czech *pomalý*, the meaning is identical (but see below for historical semantics). *Kaluppe* and Czech *chalupa* share their basic reference to a small house but differ in their connotation (negative in 'Viennese' but neutral in contemporary Czech). At the time of borrowing, this difference probably did not exist. Jungmann (1935, p. 791) paraphrases the meaning of *chalupa* as 'inferior hut; farmer's hut/house', which suggests that Czech *chalupa* has undergone semantic amelioration. *Kolatsche* and Czech *koláč* both refer to filled (or topped) pastry, however, the typical form and dough it is made of differs at least nowadays. While the Czech pastries are round (cf. Czech *kolo* 'wheel') and made of yeast dough, the Viennese ones are squares and a kind of puff pastry. However, this type of preparation is certainly a Viennese (or Austrian) innovation, which is also reflected in the dictionaries. Gartner (1902a,b,c, 1903a,b, 1904b) still

describes a *Kolatsche* as a 'common round pastry (filled or topped with plum jam, curd or the like)', which corresponds to what *koláč* refers to in Czech.

Compounds (e.g., *Kolatschenbäcker* 'confectioner'), derived words (e.g., *Kaluppner* 'inhabitant of a shabby hut') or phrasemes (e.g., *Immer/alleweil schön/ganz pomali!* 'Stay calm!') are only found in clusters 1 and 2. In the case of *pomali*, a historical meaning 'calm; quiet; careful' is reported in Castelli (1847) and Hügel (1973), which is no innovation in 'Viennese', but also existed in 19th century Czech (Jungmann, 1837, p. 275). The adjective was therefore probably borrowed with both semantic dimensions and underwent a similar semantic narrowing in both Czech and 'Viennese' in the course of the 19th century. The development of a new meaning with no correspondence in neither historical nor present-day varieties of the donor language can be observed regarding *Kaluppe* 'old woman'.

TABLE 2 Lemmas with meanings in at least two of the central clusters 1, 2 or 6

Lemma	Form-meaning pair attributed to cluster 1 (historical)	Form-meaning pair attributed to cluster 6 (general)	Form-meaning pair attributed to cluster 2 (innovations)
Kaluppe	*Kaluppner* 'inhabitant of a shabby hut' (n = 1)	*Kaluppe* 'shabby hut' (n = 15)	*Kaluppe* 'old woman' (n = 3)
Kolatsche	*Kolatschenbäcker* 'confectioner' (n = 1) *Kolatschengesicht* 'round face typical for Slavic people' (n = 2)	*Kolatsche* 'filled, small pastry' (n = 16)	*Topfenkolatsche* 'pastry with curd filling' (n = 5)
pomali	*pomali* 'calm; quiet; careful' (n = 2) *Er ist in allem ganz pomali und kriegt dabei keinen Husten!* 'In everything he does, he is very "pomali" and doesn't develop a cough' (n = 1)	*pomali* 'slow, leisurely, unhurried' (n = 19)	*Immer/alleweil schön/ganz pomali!* 'Stay calm!' (n = 4)

Throughout the full observation period, new lexemes based on the focused Czech loanwords could and still can be formed both by compounding and derivation. Similarly, phrasemes and secondary meanings developed that are not based on the Czech models. This suggests that the loanwords had already been well integrated into the recipient variety on various linguistic levels in the 19th century and still are in present-day 'Viennese'. It can therefore be concluded, that the lemmas that share meanings in clusters 1, 2 and 6 must have already been borrowed before or in the first decades of the 19th century and not in the heyday of Czech migration to Vienna, which is dated to the second half of the 19th century. This brings long-term Czech-German multilingualism in Vienna and eastern Austria into focus and suggests that (historical) sociolinguistics need to consider the full history of the contact situation in greater detail when describing the influence of Czech and other West Slavic languages on 'Viennese'.

4 Summary and conclusions

The exemplary analysis presented in Section 3 of this contribution does of course not allow for final conclusions regarding integration, variation and change of West Slavic loanwords in 'Viennese' in general. However, it demonstrates the usability of a systematic metalexicographical approach towards (lexical-semantic and/or etymological information in) historical dialect dictionaries. Linking information across various dictionaries does not only allow for insights into the making of dictionaries or the linguistic and extra-linguistic knowledge represented by

them, but also allows for diachronic perspectives on a language's or variety's lexicon. This can be especially useful in the case of historical, primarily spoken varieties such as dialects such as 'Viennese'.

References

Dictionaries and dictionary articles

Bauer, W. (1976). Pêhēim. In *Wörterbuch der bairischen Mundarten in Österreich (WBÖ). Publiziert über das Lexikalische Informationssystem Österreich (LIÖ)*. Kommissionsverlag der Österreichischen Akademie der Wissenschaften. Retrieved June 5, 2026, from https://lioe.dioe.at/articles/lieferung13-3%23Peheim

Brauer, A. (2020). *Wienerisch für Fortgeschrittene: Wean red wi Wean wü. I. E. A.* Amalthea.

Castelli, I. F. (1847). *Wörterbuch der Mundart in Oesterreich unter der Enns: Eine Sammlung der Wörter, Ausdrücke und Redensarten, welche von der hochdeutschen Sprache abweichend, dem niederösterreichischen Dialekte eigenthümlich sind, sammt beigefügter Erklärung und so viel möglich auch ihrer Abstammung und Verwandtschaft, beigegeben grammatische und dialektologische Bemerkungen über diese Mundart überhaupt. Ein Hülfsbuch, um den Oesterreicher über seine Nationalsprache aufzuklären, und Fremden dieselbe verständlich zu machen*. In Commission bei Tendler und Compagnie. Retrieved June 5, 2026, from http://data.onb.ac.at/rep/10568950

Gartner, T. (1902a). Fremdes im Wortschatz der Wiener Mundart. *Zeitschrift Für Hochdeutsche Mundarten*, *3*, 127–151.

Gartner, T. (1902b). Fremdes im Wortschatz der Wiener Mundart: Fortsetzung. *Zeitschrift Für Hochdeutsche Mundarten*, *3*, 184–210.

Gartner, T. (1902c). Fremdes im Wortschatz der Wiener Mundart: Fortsetzung. *Zeitschrift Für Hochdeutsche Mundarten*, *3*, 274–276.

Gartner, T. (1903a). Fremdes im Wortschatz der Wiener Mundart: Fortsetzung. *Zeitschrift Für Hochdeutsche Mundarten*, *4*, 118–121.

Gartner, T. (1903b). Fremdes im Wortschatz der Wiener Mundart: Fortsetzung. *Zeitschrift Für Hochdeutsche Mundarten*, *4*, 252–294.

Gartner, T. (1904). Fremdes im Wortschatz der Wiener Mundart: Schluß statt Fortsetzung. *Zeitschrift Für Hochdeutsche Mundarten*, *5*, 99–184.

Groschopf, E. (1983). Pom 'li. In *Wörterbuch der bairischen Mundarten in Österreich (WBÖ). Publiziert über das Lexikalische Informationssystem Österreich (LIÖ)*. Kommissionsverlag der Österreichischen Akademie der Wissenschaften. Retrieved June 5, 2026, from https://lioe.dioe.at/articles/lieferung18-1%23pomali

Hauenstein, H. (1972). *Wiener Dialekt: Weanerische Drahdiwaberln von A – Z*. Karner.

Hauenstein, H. (1978). *Wiener Dialekt: Weanerische Drahdiwaberln von A – Z* (2. erw. Aufl.). Karner.

Hornung, M. & Grüner, S. (2002). *Wörterbuch der Wiener Mundart* (2., erweiterte und verbesserte Auflage mit mehr als 1000 neuen Stichwörtern und Ergänzungen). öbv & hpt.

Hügel, F. S. (1873). *Der Wiener Dialekt: Lexikon der Wiener Volkssprache: Idioticon Viennense*. Hartleben. Retrieved June 5, 2026, from https://vidi.acdh-dev.oeaw.ac.at/

Jakob, J. (1929). *Wörterbuch des Wiener Dialektes: Mit einer kurzgefaßten Grammatik*. Gerlach & Wiedling.

Jungmann, J. (1835). *Slovník česko-německý: Díl I., A–J*. Knížecí arcibiskupská knihtiskárna. Retrieved June 5, 2026, from https://vokabular.ujc.cas.cz/moduly/slovniky/digitalni-kopie-detail/JgSlov01

Jungmann, J. (1837). *Slovník česko-německý: Díl III., P–R*. Knížecí arcibiskupská knihtiskárna. Retrieved June 5, 2026, from https://vokabular.ujc.cas.cz/moduly/slovniky/digitalni-kopie-detail/JgSlov03

Kim, A. (2024). Gurre. Vollartikel. In *Wörterbuch der bairischen Mundarten in Österreich (WBÖ). Publiziert über das Lexikalische Informationssystem Österreich (LIÖ)*. Retrieved June 5, 2026, from https://lioe.dioe.at/articles/Gurre

Kim, A. (2025). Kaluppe. Vollartikel. In *Wörterbuch der bairischen Mundarten in Österreich (WBÖ). Publiziert über das Lexikalische Informationssystem Österreich (LIÖ)*. Retrieved June 5, 2026, from https://lioe.dioe.at/articles/Kaluppe

Pfeifer, W. (1993). Kolatsch. In *Etymologisches Wörterbuch des Deutschen, digitalisierte und von Wolfgang Pfeifer überarbeitete Version im Digitalen Wörterbuch der deutschen Sprache*. Retrieved June 5, 2026, from https://www.dwds.de/wb/etymwb/Kolatsche

Schuster, M. & Schikola, H. (1996). *Das alte Wienerisch: Ein kulturgeschichtliches Wörterbuch* (Neuausg.). Deuticke.

Sedlaczek, R. (2011). *Wörterbuch des Wienerischen* (Orig.-Ausg., 1. Aufl.). Haymon.

Sedlaczek, R. (2023). *Das große Wörterbuch des Wienerischen. Die Herkunft der Wörter und ihre richtige Aussprache. Mit mehr als 10.000 Stichwörtern und zahlreichen Belegstellen*. Michael Wagner Verlag.

Sonnleithner, J. [anonymous]. (1800). *Mundart der Oesterreicher: Oder Kern aller ächt österreichischen Phrasen und Redensarten*. Auf Kosten des Verfassers. Retrieved June 5, 2026, from http://data.onb.ac.at/rep/10879216

Sonnleithner, J. [anonymous]. (1811). *Mundart der Österreicher: Oder Kern ächt österreichischer Redensarten und Phrasen*. N.N. Retrieved June 5, 2026, from http://data.onb.ac.at/rep/10386F84

Teuschl, W. (1990). *Wiener Dialekt-Lexikon*. Karl Schwarzer.

Teuschl, W. (2007). *Wiener Dialekt-Lexikon* (3., überarb. Aufl.). Residenz.

Wehle, P. (1980). *Sprechen Sie Wienerisch? Von Adaxl bis Zwutschkerl*. Ueberreuter.

Other references

Coleman, J. & Ogilvie, S. (2009). Forensic Dictionary Analysis: Principles and Practice. *International Journal of Lexicography*, 22(1), 1–22. https://doi.org/10.1093/ijl/ecn048

Glettler, M. (1972). *Die Wiener Tschechen um 1900: Strukturanalyse einer nationalen Minderheit in der Großstadt*. Oldenburg.

Jakubíček, M., Kovář, V., Měchura, M. & Rambousek, A. (2022). Using NVH as a Backbone Format in the Lexonomy Dictionary Editor. In A. Horák, P. Rychlý & A. Rambousek (Eds.), *Proceedings of the Sixteenth Workshop on Recent Advances in Slavonic Natural Languages Processing, RASLAN 2022* (pp. 55–61). Tribun EU.

Kim, A. (2021). The Melting Pot Revisited: Historical Sociolinguistic Perspectives on Migration and Language Contact in Vienna. In A. Auer & J. Thorburn (Eds), *Approaches to Migration, Language and Identity* (pp. 11–40). Peter Lang. https://doi.org/10.3726/b18649

Kim, A. (2022a). Das Bild der südslawischen Sprachen und ihrer Sprecher*innen in Wörterbüchern des Wienerischen. In M. Jakiša & K. Tyran (Eds.), *Südslawisches Wien. Zur Sichtbarkeit und Präsenz südslawischer Sprachen und Kulturen im Wien der Gegenwart* (pp. 293–318). Böhlau.

Kim, A. (2022b). *Slawisches im Wienerischen revisited: Die Repräsentation des „Slawischen" in Wörterbüchern des „Wienerischen"* [Doctoral dissertation, Universität Wien].

Měchura, M. (2017). Introducing Lexonomy: An Open-source Dictionary Writing and Publishing System. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek & V. Baisa (Eds.), *Electronic lexicography in the 21st century* (pp. 662–679). Lexical Computing. Retrieved June 5, 2026, from https://elex.link/elex2017/wp-content/uploads/2017/09/paper41.pdf

Newerkla, S. M. (2013). Linguistic Consequences of Slavic Migration to Vienna in the 19th and 20th Centuries. In M. Moser & M. Polinsky (Eds.), *Slavic languages in migration* (pp. 247–260). Lit.

Steinhauser, W. (1978). *Slawisches im Wienerischen* (Zweite vermehrte und verbesserte Auflage). Verband der wissenschaftlichen Gesellschaften Österreichs.

Studer, M. (2013). *WeightedCluster Library Manual: A practical guide to creating typologies of trajectories in the social sciences with R*. LIVES. https://doi.org/10.12682/LIVES.2296-1658.2013.24

Contact information

Agnes Kim
University of Ostrava
agnes.kim@osu.cz

Tibor M. Pintér, Manana Rusieshvili-Cartledge, Dóra Pődör, Katalin P. Márkus & Marine Makhatadze

SUBLEXICAL SYSTEMS OF COLOUR IN BIBLICAL GEORGIAN AND HUNGARIAN

Findings of the Pentateuch

Abstract As a culturally embedded linguistic category, colour varies across translations of the same source text. This paper examines how colour is encoded diachronically in biblical Hungarian and Georgian, addressing two questions: how colour terms are realised across parallel translations, and what cross-linguistic patterns of lexical and syntactic variation emerge. The analysis draws on a parallel corpus of Old and Modern Georgian and Hungarian Bible translations. Colour terms are defined as lexical units denoting chromatic properties, including both basic colour adjectives and material-based or metonymic expressions. The findings reveal a systematic divergence between the two cultural-linguistic traditions. Georgian shows a diachronic shift from materially grounded, periphrastic expressions towards more abstract and generalised colour terminology, often accompanied by increased morphological transparency. Hungarian, by contrast, demonstrates greater lexical continuity, with variation primarily realised through syntactic explicitation and the redistribution of semantic content. In both cultural-linguistic traditions, colour is encoded via material reference, emerging in descriptions of textiles and ritual objects. The findings demonstrate that colour encoding reflects broader processes of semantic restructuring. Hungarian serves as a comparative baseline, highlighting a trajectory in which syntactic rather than lexical change predominates.

Keywords Historical Lexicography; Cognitive Linguistics; Contrastive Colour Semantics; Parallel Corpora; Bible Translation; Digital Lexicography

1 Introduction

Colour is a linguistic phenomenon usually linked to cultures (Wierzbicka, 1990), yet its representation becomes particularly complex in several translations of the Scripture. This raises two interrelated questions: can colour be rendered differently across translations of the same source, and how can corpora support the systematic investigation of such variation?

The main objectives of this research are: (a) to conduct a diachronic investigation of colour terminology in translated biblical texts originating from the same source (the Pentateuch), with special attention to Georgian and Hungarian translations; (b) to identify and compare sublexical systems and material-based encoding strategies; and (c) to demonstrate how AI-assisted parallel corpora can support systematic and reproducible lexicological and lexicographic research on early texts, using AI for text processing, and verse comparisons carried out manually by using regular expressions in XML files.

Recent advances in AI-based language technology offer powerful tools for addressing these questions (e.g., Sketch Engine can be used for both pre-processing and complex analyses, but other web-based solutions can be promising as well). In particular, parallel corpus environments such as the Unified Bible Reader[1] (https://ebo.kre.hu) enable the comparative study of multiple Bible translations across languages, periods, and traditions. In this respect, the

[1] The Unified Bible Reader is a project of the Károli Gáspár University of the Reformed Church in Hungary. It aims to enable the parallel reading and research of Bible translations of different biblical denominations, as well as of Greek and Latin texts. The website is available in three languages: Hungarian, English, and German.

(almost) identical numbering of the books and verses of the Bible across different languages and translations is of great assistance.

AI-assisted tools for corpus construction, text alignment, and semantic search are essential not only for parallelisation but also for the identification, classification, and contextual analysis of colour terms. Automated alignment algorithms, embedding-based similarity measures, and concordance generation allow researchers to track colour lexemes diachronically and cross-linguistically with a level of precision unattainable through manual methods alone. The present research builds on the findings of Rusieshvili-Cartledge and Makhatadze (2025), complemented by recent work on language-technology-based Bible readers and concordances (M. Pintér & P. Márkus, 2024, 2025).

The study focuses on the diachronic investigation of colour terminology in translated texts originating from the same source, with special attention to Georgian and Hungarian Bible translations. After briefly outlining the main bottlenecks in building historical corpora – such as orthographic variation or morphological opacity – we demonstrate how AI-supported materials can facilitate lexicological and lexicographic research on early texts.

2 Theoretical Framework

Colour constitutes a culturally embedded linguistic category par excellence – and specific cultural references within the context depicted in the Bible (Wierzbicka, 1990). Far from being a purely perceptual, universal phenomenon determined solely by human physiology, colour terms are profoundly shaped by the physical environment, material culture, technological practices (especially dyeing and textile production), and the symbolic systems of a given speech community. In biblical texts, this cultural embedding is particularly salient: colour is rarely an abstract hue but is almost always linked to concrete objects, materials, or states of the world. García Ureña et al. (2022, p. 24) describe such terms as "environmental concepts", perceived through prototypical entities rather than abstract chromatic values.

Recent philological research has demonstrated that colour terms in the Bible are rarely abstract designations of hue in the modern Newtonian sense (Newton was not only a pioneering mathematician and physicist but also a deeply devout theologian). Instead, they are fundamentally "embodied," inextricably linked to concrete objects, materials, vegetation, textiles, dyes, metals, and ritual artefacts (García Ureña et al., 2022, pp. 24–26). Cognitive linguistics is the most suitable framework for this study because it recognises that colour language in the Bible frequently encodes a state or salient property of an entity rather than an intrinsic quality (García Ureña et al., 2022, p. 23). The Hebrew Bible tends to be frugal and material-oriented; the Septuagint and Vulgate reinterpret and expand this vocabulary in line with their respective cultural sensibilities. This embodied perspective provides the central theoretical lens for the present comparative analysis of Georgian and Hungarian biblical translations.

Cognitive linguistics offers powerful tools for analysing diachronic and cross-linguistic variation because it views meaning as usage-based and encyclopaedic. As sacred texts move across languages and historical periods, colour expressions undergo processes of generalisation, specialisation, calquing, or syntactic redistribution. García Ureña et al. (2022, pp. 24–25) emphasise that linguistic meaning must be supplemented by "encyclopaedic" knowledge of the speaker's world. The translation of biblical texts thus becomes a privileged site for observing how colour semantics evolves within each language while remaining anchored in the same source text. The present research extends this comparative perspective to two underrepresented traditions: biblical Georgian (Kartvelian) and Hungarian (Uralic).

A central feature of biblical colour language is the frequent use of sublexical and material-based encoding. Rather than relying primarily on basic colour adjectives, translators often employ periphrastic constructions, metonymies, and direct references to source materials (plants, dyes, fabrics, or minerals). Cognitive linguistics offers a range of principles that are particularly useful for lexicographical research (Ibarretxe Antuñano, 2004). Several of these appear especially relevant to the study of colour language in the biblical corpus: the centrality of meaning as the essence of language, the continuum between semantics and pragmatics, polysemy, idiomaticity, and prototype theory. These epistemological principles find expression in various methodological tools. One of the most important is the cognitive mechanism of metonymy, which is essential to the analysis of colour terms in the biblical corpus (metaphor being the other, as the Scripture is filled with metaphors representing knowledge of its era). Conceptual metonymy is not merely a rhetorical figure but a fundamental cognitive process whereby one element within a conceptual domain provides mental access to another element in the same domain. Among the various types of conceptual metonymy, the salient property of entity subtype is particularly relevant. This mechanism is clearly observable in the neuter use of the Greek adjective *χλωρός* in the Septuagint (e.g., *πᾶν χλωρόν* in Gen 2:5, Exod 10:15; *τὸ χλωρόν* in Gen 30:37), where the salient property "green" stands for the entire entity "verdure" or "greenery." The same cognitive pattern underlies the frequent material-based and periphrastic colour expressions found across biblical translation traditions.

3 Methodology

3.1 Bible translations and language technology

The digitisation of biblical texts and their parallel arrangement pose several challenges for researchers. Although many problems arising during the digitisation process can be avoided, parallelising texts stored in databases or XML files often remains problematic. Aligning texts side by side involves several difficulties. The development of parallel Bible readers involves four fundamental issues, of which the last three (b–d) relate to the alignment of already digitised texts: (a) digitisation, (b) data cleaning, (c) importing, and (d) parallelisation (see Király, 2024, for a detailed discussion). Artificial intelligence can be of considerable assistance in several stages of this process: Microsoft Copilot performs optical character recognition (OCR) of Hungarian texts with a low error rate and significantly facilitates the alignment of parallel texts, not only on the basis of locus (verse position) but also through the identification of identical lexical elements. In the Unified Bible Reader, parallel translations can be arranged side by side using the built-in "Search" function to locate verses containing colour-related lexical items. This allows for the easy identification of verses containing colour lexemes across both modern and historical translations. These can then be manually aligned with the corresponding verses in Georgian translations (with the expectation that the software will, in the near future, incorporate the Georgian translations used in this research).

From the books of the Old Testament, two translations for each language were examined. Georgian is represented by the tenth-century *Old Testament*, as in the Mtskheta Manuscript, and by the late twentieth-century *Old Testament: New Translation into Modern Georgian* by Protopresbyter Giorgi Gamrekeli. The Hungarian texts are represented by the early twentieth-century *Revised Translation* and the early twenty-first-century *Revised New Translation*. This selection of texts allows us to examine the semantic framework of colour lexemes both synchronically and diachronically, and to observe potential semantic or grammatical changes.

3.2 Colour(ful) terms in the Scripture

Biblical colours carry a range of hermeneutical and spiritual meanings. The colours found in Scripture do not merely reflect the colours of the physical world — and therefore may be rendered by different lexemes across languages — but possess specific meanings characteristic of either the Old or the New Testament. In general, colours in biblical texts tend to cluster around particular objects. In the Old Testament, such objects include the Tent of Meeting and the Tabernacle, as well as priestly garments (white, blue, purple, red) and elements of the natural world (blue, green). These colours carry both denotative and connotative meanings. It is therefore not surprising that the most frequently attested colours in Old Testament texts include white, red, scarlet, purple, blue, gold, black, green, silver, and bronze.

Cross-linguistic research can reveal how these colour terms are represented semantically and grammatically in the languages under examination, and may also bring to light specific, often archaic formulae characteristic of Bible translations — forms that may even persist in modern usage. Beyond their evident reference (e.g. blue as the colour of the sky, white or purple as the colour of garments), colours also possess symbolic meanings. Thus, they are associated with concepts such as sin (black), sacrifice (red), purity (white), wealth (purple), or the law (blue), forming a distinct symbolic system within biblical discourse.

For this very reason, a recurring tendency was identified through historical-lexicographic analysis: colour in Georgian and Hungarian biblical traditions is frequently encoded by lexemes whose reference extends beyond chromatic value to include material substances (*zöld* növény – *green* plant – თივა *მწვანვილისა*; *vörös* fonal – *scarlet* thread – *მეწამული*). This phenomenon is especially prominent in descriptions of textiles, precious stones, metals, and natural dyes. The linguistic representation of individual colours and the objects associated with them may vary across languages. An attributive structure in one language may appear as a simple noun or as a compound in another, depending on whether the colour is expressed explicitly as a colour term or through a prototypical object representing it. Hungarian translations fluctuate between explicit colour adjectives and references to coloured objects, a variation that can be systematically mapped using AI-based alignment and frequency analysis. For example, in Gen 30:32, the old Hungarian translation renders “the speckled and spotted sheep” as *pettyegetett és tarka bárányt*, whereas the new translation renders it as *minden pettyest és tarkát* [‘all speckled and spotted’].

Another illustrative example appears in Georgian translation in Exodus 10:15, where Old Georgian *თივა მწუანვილისა* (tiva mtsuanilisa [hay of greenery]) corresponds to Modern Georgian *მწვანე მცენარე* (mtsvane mtsenare [green plant/vegetation]). Diachronically, this reflects a shift from a descriptive, periphrastic expression to a more abstract colour-based formulation. The older lexeme conceptualises colour through material substance (hay) and visual experience (greenery), while the modern translation employs a basic colour adjective. The Hungarian translation of this verse is *zöld* (green), which refers not to the colour itself but to the visual experience of plants (their greenness and their symbolism of wealth in nature). Other Hungarian translations demonstrate a shift toward material specificity. In Exodus 26:1, modern Hungarian *bíbor* (‘purple’) functions as a textile designation (‘linen’), whereas older translations restrict its meaning to colour alone, indicating a redistribution of semantic load across lexical and syntactic structures.

Venuti’s notion of domestication is detectable in modern translations through lexical generalisation and semantic economy. A characteristic feature of the Old Georgian biblical lexicon is the frequent overlap between colour and material reference, especially where ancient referents or dyeing practices are no longer culturally accessible. In Exodus 25:4, for instance, lexemes

such as პორფირი (porpiri [porphyry, dark red, purple) and ბისონი (bisoni, dark red)[2] occupy an intermediate position between colour terms and ritual material designations. In Hungarian translations, 'purple' and 'scarlet' occur in the same verse but in different semantic distributions: the old Hungarian translation refers to this as *bíborpiros, és karmazsinszinű fonal* ('purple, and scarlet thread') while the new translation reflects it as *piros bíbort, karmazsin fonalat* where *bíbor* is colour in the older text and 'cloth' in the new one.

Diachronic comparison reveals a seemingly paradoxical development. While many Georgian colour lexemes undergo semantic broadening and simplification (whereas Hungarian tends to preserve lexical continuity), others become more morphologically transparent through calquing. A clear example is Exodus 39:8, where Old Georgian მეწამული (metsamuli [purple or dark red]) is replaced in Modern Georgian by ჭიაფერი (chiaperi [colour of pokeweed, dark red]). This shift reflects a move from a conventionalised, stylistically elevated colour term to a semantically decomposable form that foregrounds the dye source. The modern lexeme emphasises the dye source, restoring the etymological motivation obscured in the older term and thus prioritising semantic precision over stylistic naturalness. A similar tendency can be observed in the Hungarian translations in this verse. However, this tendency toward explicit clarification in modern forms cannot necessarily be regarded as a general linguistic feature (a claim that would require further in-depth research), but rather reflects a clearer — and therefore more explicit — rendering of Old Testament objects and contexts from the perspective of contemporary language use: old Hungarian *bíborpiros, és karmazsinszínű, és sodrott lenből* ('from purple and scarlet of fine woven linen'), new Hungarian *piros bíborból, karmazsin fonálból és sodrott lenfonálból* ('from red textile and scarlet yarn and woven linen').

A similar process of semantic and stylistic neutralisation affects evaluative colour terms. Old Georgian სპეტაკი (spetaki [radiant, brilliant white]) in Genesis 49:12 is rendered as the basic colour adjective თეთრი (tetri [white]) in Modern Georgian. In Hungarian, older translations employ metaphorical references to wine and milk (*Bortól veresek szemei, tejtől fehérek fogai* 'eyes red from wine, teeth white from milk'), whereas modern versions rely on direct comparison (*Szemei bornál ragyogóbbak, fogai tejnél fehérebbek* 'eyes brighter than wine, teeth whiter than milk'). These changes indicate a shift away from expressive markedness towards neutrality and clarity.

Finally, Old Georgian colour vocabulary appears more gradational and perception-based, frequently employing modifiers that encode nuance. Contrasts such as Old Georgian თმაი მოწითლე (tmai motsitle [reddish hair]) versus Modern Georgian შავი თმა (shavi tma [black hair]) illustrate a broader shift toward a discrete system of basic colour terms.

4 Discussion

Overall, AI-supported corpus analysis shows that Georgian biblical colour lexemes evolve from a richly descriptive system toward semantic generalisation, morphological transparency, and a reduced inventory, while Hungarian diachronic variation is primarily realised through syntactic rather than lexical change (in Exodus 38:23 the old translation is *kék, és bíborpiros*, *és*

[2] Besides the colour, both porpiri and bisoni refer to a valuable dark red or purple cloak symbolising the king's authority. http://www.nplg.gov.ge/gwdict/index.php?a=term&d=39&t=649; Their Hungarian equivalent is bíbor (from the lat. purpura and greek porphüra). As the purple robe was worn by bishops, the equivalent of noun cardinal is in Hungarian 'bíboros': noun 'bíbor' + suffix '-os (nouns with this affix are mostly derived from a nominal structure when they are used for the referent without the primary noun).

karmazsinszínű, és lenfonállal, while the new is *kék és piros bíborból, karmazsin fonálból és lenfonálból* where *fonál* 'thread' is a marked explicature with full syntactic form).

The events of the Books of Moses unfold within a distinctive biblical setting, in which colours possess not only aesthetic but also sacral significance. The colours of the Bible belong from the outset to a defined chromatic spectrum, reflecting the Asian–African environment on the one hand, while also indicating the sacral meanings and spiritual qualities characteristic of the given periods.

Colours in the Pentateuch represent only one register within this broader system, most notably reflecting the materials of sacred spaces and elements of the natural world. Differences between Georgian and Hungarian translations may arise either from the specific ways in which colours are represented in the respective languages or from the particular linguistic solutions characteristic of a given translation or historical period. In either case, such differences become clearly observable and describable through a comparative examination of these books and their colour terms.

5 Conclusion

This research states that the encoding of colour terms in Bible translations across the two languages and cultures (Hungarian and Georgian) is not a static rendering or transposition of certain hues, but a dynamic process of cultural and semantic restructuring of colour terminology. As well as this, the contrastive analysis discussed in this paper reveals two distinct strategies of transposition: while Georgian undergoes a lexical modernisation through the adoption of transparent colour adjectives, Hungarian largely maintains a conservative lexicon and relies on syntactic explicitation to resolve semantic ambiguity. The findings also emphasise the need for historical dictionaries to move beyond simple definitions of terms and to incorporate cognitive-stylistic mechanisms, specifically conceptual metonymies and metaphors, as drivers of semantic change. Additionally, the integration of AI-assisted corpus methods demonstrates transformative potential value for lexicography: it provides the precision and scale to trace and follow subtle linguistic shifts across the historical development of translation.

References

Ibarretxe Antuñano, I. (2004). What is cognitive linguistics? A new framework for the study of Basque. *Cahiers de l'Association for French Language Studies,* 10(2), 3–31. https://www.semanticscholar.org/paper/What-is-cognitive-linguistics-A-new-framework-for-Antu%C3%B1ano/01cd0586fe545d807064b7aed8348849358b1da4

Király, L. Z. (2024). Parallelising Bible texts. Developing the database of the Unified Bible Reader online application. *Argumentum*, 20, 438–449. https://doi.org/10.34103/ARGUMENTUM/2024/26

M. Pintér, T. & P. Márkus, K. (2022). The Role of Online Bible Readers in Biblical Concordance Making. *Hungarian Studies Yearbook*, 4(1), 183–196. https://doi.org/10.2478/hsy-2022-0009

M. Pintér, T. & P. Márkus, K. (2024). Possibilities of Language Technology in Lexical Analyses of Canonical Hungarian Bible Translations. *Studia Universitatis Babes-Bolyai Theologia Reformata Transylvanica*, 69(2), 22–43. https://doi.org/10.24193/subbtref.69.2.02

Rusieshvili-Cartledge, M. & Makhatadze, M. (2025). Colourfulness of the lexis: Lexicographic treatment of colour terms in Georgian-English Thematic Dictionary. *Journal of Digital Terminology and Lexicography*, 1(1), 25–37. https://doi.org/10.25430/pupj.jdtl.1749066594

Ureña, L. G., Valeriani, E., Angelini, A., Carreto, C. S. & Gimeno, M. S. (2022). The Language of Colour in the Bible: Embodied Colour Terms Related to Green. de Gruyter.

Wierzbicka, A. (1990). The meaning of colour terms: Semantics, culture, and cognition. *Cognitive Linguistics*, 1(1), 99–150. https://doi.org/10.1515/cogl.1990.1.1.99

Contact information

Tibor M. Pintér
Károli Gáspár University of the Reformed Church in Hungary
m.pinter.tibor@kre.hu

Manana Rusieshvili-Cartledge
Ivane Javakhishvili Tbilisi State University
manana.ruseishvili@tsu.ge

Dóra Pődör
Károli Gáspár University of the Reformed Church in Hungary
podor.dora@kre.hu

Katalin P. Márkus
Károli Gáspár University of the Reformed Church in Hungary
p.markus.kata@kre.hu

Marine Makhatadze
Ivane Javakhishvili Tbilisi State University
marine.makhatadze@tsu.ge

Ana Mihaljević & Josip Mihaljević

ARTIFICIAL INTELLIGENCE AS SUPPORT FOR THE DIGITIZATION AND COMPILATION OF THE *DICTIONARY OF THE CROATIAN REDACTION OF CHURCH SLAVONIC*

Abstract The *Dictionary of the Croatian Redaction of Church Slavonic* (RCJHR) is the first lexicographic description of the Croatian Church Slavonic language. Within the project *Development of the Digital Infrastructure Model of the Old Church Slavonic Institute* (DigiSTIN), the printed RCJHR is being digitized and adapted for online publication, while the RCJHR project focuses on compiling new entries and exploring a fully digital lexicographic workflow. This paper examines the potential of large language models (LLMs) to support the digitization and compilation of a dictionary documenting a low-resource historical language with an exceptionally complex microstructure involving five languages and four scripts. Six leading models (GPT-5.4 Thinking, DeepSeek-V3.2, Gemini 3 Pro, Claude Opus 4.6, Qwen3.6-Plus, and Copilot GPT-5) were evaluated across four tasks: OCR of previously published fascicles, HTR of handwritten index cards, defining entry structure, and compiling new dictionary entries. The results demonstrate that LLMs significantly accelerate digitization and assist in identifying foreign-language equivalents, suggesting sense divisions, and proposing cross-references. However, they remain unreliable in resolving subtle semantic distinctions and recognising the Glagolitic script, confirming that human philological expertise remains indispensable.

Keywords Croatian Church Slavonic; large language models (LLMs); digitization; dictionary compilation; low-resource languages

1 Introduction

The Croatian redaction of Church Slavonic, also known as Croatian Church Slavonic, is the oldest Croatian literary language. It developed from Old Church Slavonic and is distinguished by the use of the Croatian (Angular) Glagolitic script and vernacular Chakavian elements.[1] The *Dictionary of the Croatian Redaction of Church Slavonic* (RCJHR) is the first lexicographic description of that language. The corpus of the RCJHR consists of approximately sixty Glagolitic sources, ranging from the 11th to the 17th centuries.[2] To date, 23 fascicles covering entries from *A* (ⰰ) to *I* (ⰻ) have been published. The RCJHR has a very rich microstructure. Each dictionary entry includes five languages (Croatian Church Slavonic, Latin, Greek, modern Croatian, and English) and four scripts (Angular Glagolitic, Old Cyrillic, Latin, and Greek alphabet). As part of the project *Development of the Digital Infrastructure Model of the Old Church Slavonic Institute* (DigiSTIN)[3] the printed RCJHR is being digitized and adapted for online publication through a specialized web portal[4]. The possibilities of transitioning to a fully digital workflow are also being explored.

1 More about the Croatian Church Slavonic language see Mihaljević, 2014, and about the Glagolitic script Mihaljević, Mihaljević & Šimić, 2025.

2 For more about the RCJHR see e.g. Nazor 2008, 2017; Šimić, 2017, and the webpage of the project https://stin.hr/sadrzaj/rjecnik-crkvenoslavenskoga-jezika-hrvatske-redakcije-rcjhr/ (May 11, 2026)

3 https://stin.hr/sadrzaj/digistin/ (May 11, 2026)

4 More on the development of the portal (https://rjecnik.stin.hr/) see Mihaljević & Mihaljević 2024a, 2024b.

This paper investigates how large language models (LLMs[5]) can be employed in the digitization and compilation of this dictionary, which documents a language with very limited resources and exhibits a highly complex microstructure. This paper presents several tests and potential applications of LLMs.

2 AI in Lexicography: A Review of Previous Research

Since 2022, with the rapid advancement of artificial intelligence based on LLMs, particularly following the launch of ChatGPT, the use of such technologies has become commonplace in everyday activities as well as in professional contexts. (Terzimehić et al., 2025, p. 1). Despite the risks of over-reliance on AI systems and the generation of ungrounded outputs that may appear convincing but are incorrect (Bender et al., 2021, pp. 616–617), many organizations are increasingly integrating AI tools into workplace processes as a form of support, where such systems are designed to support workers by providing real-time assistance (Brynjolfsson et al., 2023, pp. 10–11). There are significant changes in the IT job landscape, particularly affecting lower-level coding roles, as LLMs now generate 41% of all code, reducing demand for entry-level positions focused on writing routine tasks (Pauklin, 2026; Levine, 2026). However, many believe that AI is not here to replace humans; rather, those who know how to use AI will replace those who do not (Lakhani, 2023).

In the field of lexicography, the integration of artificial intelligence has become increasingly prominent, as evidenced by the growing number of scholarly publications addressing AI-related topics (mostly LLMs). In the current era of AI, dictionaries exist not just for human beings, but also for machines, and this shift urges us to deepen the discussion of theoretical lexicography and to incorporate a wider range of data (Utkirova, 2024, p. 71). A search conducted in the Scopus database using the advanced search option identified[6] 21,103 publications related to artificial intelligence and lexicography, spanning the period from 1969 to April 2026. Notably, more than half of these publications (11,754) were published between 2019 and 2026. The topics range from LLMs writing dictionary entries (Evert, Ganslmayer, & Rink, 2024; Lew, 2023), LLMs analysing dictionary entries (or entry parts) (Balenović & Proroković, 2025; Filipović Petrović & Beliga, 2025), and using LLMs to create word examples (Merx, Vylomova, & Kurniawan, 2024; Cai et al., 2025) and translate texts (Lu et al., 2024; Greco et al., 2025). All these papers exhibit a spectrum of perspectives on using AI in lexicography, yet they converge on a central theme: LLMs demonstrate impressive proficiency in generating fluent definitions and structured lexicographic content (Lew, 2023, pp. 8–9), but their reliability diminishes sharply when faced with nuanced semantic and morphologic distinctions (Balenović & Proroković, 2025, pp. 61–62), figurative language (Filipović Petrović & Beliga, 2025, p. 232), or low-resource languages[7] (Merx et al., 2024, p. 67).

[5] LLMs (Large Language Models) are artificial intelligence systems trained on vast text corpora that predict the next token (a word or part of a word) and, based on this, generate or complete text. Examples include GPT, Claude, Gemini, Llama, and DeepSeek (Perak, 2025, p. 309).

[6] Using the Scopus advanced search interface (https://www.scopus.com/search/form.uri?display=advanced), the following query was applied: TITLE-ABS-KEY ((“artificial intelligence” OR AI OR “machine learning” OR “deep learning” OR “neural network*” OR “large language model*” OR LLM* OR “natural language processing” OR NLP) AND (lexicograph* OR lexicology OR dictionary OR “dictionary making” OR lexicon OR “computational lexicography” OR “digital lexicography”)).
The search was conducted across titles, abstracts, and keywords to identify publications addressing the intersection of artificial intelligence and lexicography.

[7] Low-resource languages are languages for which there is limited linguistic data and resources available for NLP tasks (Al-Wesabi, 2023, p. 1).

While AI can significantly accelerate the drafting process (Evert, Ganslmayer, & Rink, 2024, pp. 317–318) and serve as a valuable assistant for conceptual organization (Filipović Petrović & Beliga, 2025, p. 219), the shared view of these works is that human expertise remains indispensable for verification, post-editing, and ensuring the empirical accuracy expected of dictionaries (Balenović & Proroković, 2025, pp. 60–62; Filipović Petrović & Beliga, 2025, p. 236; Greco et al., 2025, p. 716; Merx et al., 2024, pp. 70–71).

The use of AI in lexicography has been widely discussed; however, most studies have focused on major languages – predominantly English. Even when so-called "smaller" languages are considered, these are typically modern languages for which, despite their classification as less widely spoken, a substantial body of attestations and abundant digital resources exist.

Research on the use of AI in lexicographic work involving medieval or low-resource languages remains extremely limited. The same applies to Croatian Church Slavonic and the Croatian Glagolitic script, for which only a limited number of NLP resources are available online. Mihaljević and Mihaljević (2024a; 2024b) mention the process of creating the digital version of the RCJHR, but there is no direct reference to the use of AI. They only mention the use of the AI-based HTR (handwritten text recognition) program Transkribus for machine reading of the cards that form the corpus for the compilation of the RCJHR. However, they note that the results were not usable at the time because the cards contain multiple scripts, such as Old Cyrillic, Latin, and Greek, and approximately twenty different handwritings (Mihaljević & Mihaljević, 2024b, pp. 171–172).

While the usage of AI for the text recognition of medieval Glagolitic and Cyrillic scripts has been explored (e.g., Rabus & Thompson, 2023; Romein et al., 2025), these studies primarily focus on evaluating HTR engines and programs (e.g. Transkribus, eScriptorium) for the automatic transcription of historical documents and printed books and addressing specific paleographic challenges like abbreviations and scriptio continua[8]. Both Rabus & Thompson (2023) and Romein et al. (2025) emphasize that there is no universally optimal solution for HTR of ancient scripts, as performance depends on the script type, language, and data quality (Romein et al., 2025, p. 116; Rabus & Thompson, 2023, p. 12), while the highest accuracy is typically achieved through the time invested in training and refining the model rather than by using a particular program (Romein et al., 2025, p. 130; Rabus & Thompson, 2023, p. 32), even if Rabus considers Transkribus to be the most accessible and easy to use solution for scholars (Rabus & Thompson, 2023, p. 32).

In our case, by contrast, the object of study is a historical literary language, no longer in active use, for which the number of available attestations is severely limited. As a result, the lexicographer is often required to determine the meaning of a given word based on only a few usage examples (sometimes just one). This process relies heavily on extensive philological expertise, including knowledge of Latin and Greek, familiarity with theological concepts, and comparative insight into other Slavic languages. It also involves consulting etymological dictionaries of various languages and dictionaries of Old Church Slavonic to arrive at a well-founded interpretation. Given the highly specific nature of the language under discussion, our expectation is not that AI could replace the human lexicographer. Rather, the aim of this study is to determine in which aspects, and to what extent, AI can function as a supportive tool in the lexicographic process "to speed up dictionary compilation by providing quality draft material which human lexicographers then assess and improve upon" (de Schryver, 2023, p. 377). In addition, we seek to evaluate which model proves to be the most reliable and practically useful in this context.

[8] Scriptura continua (Latin for *continuous script*) in historical Glagolitic texts is defined as a writing style where no spaces exist between words (Romein et al., 2025, p. 119).

In Gilles-Maurice de Schryver (2025), the use of ChatGPT is described in the context of Lusoga, a language that is still very poorly documented. However, a substantial starting point was provided by uploading into the system a corpus of 3.7 million tokens, along with a comprehensive monolingual dictionary in XML format. For Croatian Church Slavonic, no such dictionary or corpus exists. Nevertheless, the available examples often include, alongside the Croatian Church Slavonic sentences, parallel passages in Greek and Latin. These parallels can serve as an important source for drawing conclusions, as well as for identifying the context to which a given example belongs.

3 Selecting Tools

The problem when testing LLMs lies in the fact that their algorithms and software are rapidly changing, resulting in constant evolution and competition between programs such as Google Gemini, OpenAI's ChatGPT, High-Flyer's DeepSeek, and Anthropic's Claude, with each company releasing regular updates that improve AI functions and produce better results than before (Jiang et al., 2025, p. 1). This makes it difficult to determine which LLM tool is superior for certain tasks, as they are still competing and no single model has established clear dominance. As of 2026, comparative analyses of leading LLM models suggests a differentiated performance landscape across multiple areas. Claude is widely regarded as one of the best AI assistants when it comes to coding and programming tasks (SWE-bench, 2026; Local AI Master Research Team, 2026). DeepSeek is often praised for its strong performance in solving complex mathematical problems (Caswell, 2026). Google Gemini is frequently recognized as a strong general-purpose assistant, particularly for fast, well-rounded responses and multimodal capabilities, although its relative performance varies by benchmark and use case (Artificial Analysis, 2026). ChatGPT may have been the leading LLM and remains the most widely used AI tool overall (Similarweb, 2026), but recent benchmark analyses suggest that it is no longer the best at solving complex problems, as Gemini and Claude outperform it on some reasoning tasks (LM Council, 2026). But all of this can change within days, as soon as the next update for one of these applications arrives. For this research, we decided to test these LLMs: GPT-5.4 Thinking, DeepSeek-V3.2 in Expert Mode, Google Gemini 3 Pro, Claude Opus 4.6, Qwen3.6-Plus, and Copilot GPT-5 reasoning model in Microsoft 365. The first four were selected based on their popularity. Copilot GPT-5 was also included because, as members of an academic institution, we are entitled to use the GPT-5 reasoning model within Microsoft 365 free of charge, without the limits that apply to the free version of the same program. Qwen was selected based on recommendations for strong HTR use at the University of Freiburg.

All models were tested through their public chat interfaces between April and May 2026. The evaluation was qualitative: for each task, the outputs were compared with the corresponding material produced by the human editors of the RCJHR. OCR was tested on several pages of the published fascicles and HTR on several handwritten index cards (the results presented here are based on one representative card containing several special characters); entry structuring was tested on a file containing several complete entries, and entry compilation on the *exhaustum* entry *igra*, first with a small set of entries and subsequently with an entire fascicle as input. The compiled entries were assessed against four criteria: headword structure, sense division, foreign-language equivalents, and cross-references.

4 Using LLMs for OCR of the RCJHR and HTR of its Corpus

The web version of the RCJHR is currently available online as a public demo (https://rjecnik.stin.hr/), with new entries being added regularly. A more detailed account of the development

of the portal and the various challenges associated with it is provided in Mihaljević and Mihaljević (2024a, 2024b). During the preparation of the digital edition of the RCJHR, a major difficulty arose from the fact that the earliest fascicles were not available in their original.doc form. Moreover, since the RCJHR contains numerous non-standard characters and multiple scripts and languages, OCR software did not yield satisfactory results. Consequently, texts were initially prepared for training a customized model in Transkribus. The model performed well for the Latin script, which was the most prevalent, but showed poor recognition of Cyrillic and Glagolitic scripts. Despite the inclusion of a substantial number of pages, there was insufficient training material for these scripts, as they occur only occasionally on individual pages. However, further development of a custom model was abandoned after testing various LLMs. In this context, Gemini, Qwen, and Claude Opus proved to be highly effective and useful, demonstrating strong performance in reading Latin, Cyrillic, and Greek scripts. Copilot produced the poorest results, as it incorrectly recognized some Latin letters. It also inconsistently processed a Greek text (often rendering it in the Latin script unless explicitly instructed to preserve the original Greek), and it converted subscript numbers into standard rather than retaining subscript formatting. ChatGPT performed somewhat better in comparison. DeepSeek did not reliably recognize specific Latin characters, such as the yat and reduced vowels. Although Qwen generally produced good results, it tended to confuse the two reduced vowel characters ъ and ь, and it occasionally rendered parts of the text in Cyrillic, including segments that are originally written in the Latin script. None of the tested LLMs, however, successfully recognized the Glagolitic script. The LLMs were tested on a larger number of dictionary pages, and for the purposes of this study we present the results produced by the tested models after being given the instruction to transcribe the page in the Latin script, without further instructions or corrections[9].

The RCJHR corpus consists of paper cards written in various handwriting styles and containing Cyrillic, Latin, and Greek scripts. The original Glagolitic text from the manuscripts was transcribed into Cyrillic on the cards. In recent years, attempts were made to train HTR models to read these cards in Transkribus; however, satisfactory results could not be achieved due to the high variability of handwriting and an insufficient training dataset (Mihaljević & Mihaljević, 2024a, p. 315). With the emergence of more recent versions of LLMs, further efforts to train models in Transkribus were discontinued, as Gemini, Qwen, and Claude Opus produced sufficiently good results. An additional advantage is that the cards are written in the Cyrillic rather than the Glagolitic script, which these LLMs do not recognize. We used LLMs to transcribe a larger number of paper cards; however, we present the results based on the entry *lûtodêê* (one of the imaginary devil's names in *Amulet*), as this entry itself contains several special characters (Figure 1). DeepSeek and Copilot produced completely unusable outputs in which almost nothing corresponds to the original text. ChatGPT recognized Latin script relatively accurately and partially recognized the Cyrillic script; however, the results are still insufficiently reliable for practical use. Qwen yielded somewhat better results. Claude Pro produced very good results and could expand abbreviations, but it did not correctly recognize the target lemma. Gemini provided by far the best results, with only a small number of errors[10].

[9] The outputs produced by the LLMs: https://drive.google.com/file/d/1d2WupesHBN-YyyglfsRh1NGONh05-mC4/view (Retrieved May 13, 2026)

[10] The outputs produced by the LLMs: https://drive.google.com/file/d/1vNQJfurTg5PPMcPRTSF0RpDqoUrdfri0/view (Retrieved May 13, 2026)

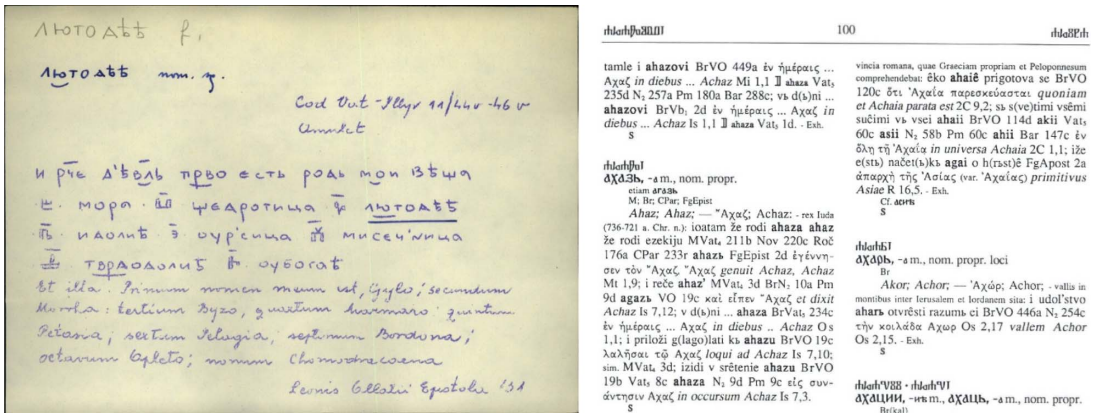


FIGURE 1 Handwritten paper cards from the corpus and a dictionary page used for the tests conducted in this study

Beyond transcription quality, LLMs also offered additional benefits, such as identifying textual sources, locating complete biblical references, and rapidly retrieving Croatian and English parallel passages.

LLMs have proven highly useful both in the OCR processing of the first published fascicles of the RCJHR and in the HTR processing of handwritten index cards on which the RCJHR is based, thereby enabling faster and more efficient digitization.

5 Defining the Entry Structure

Since we are currently transitioning to a new dictionary compilation workflow, from editing entries in Word documents to creating them directly online within a predefined structure, we investigated whether LLMs could assist in developing an appropriate template for the input of new dictionary entries. We wanted to examine whether LLMs could, based on the text of a dictionary entry, create an appropriate entry structure and determine how the available data should be organized. We then compared this proposed structure with our current structure, which was defined entirely by human experts and is implemented in our dictionary system (made in WordPress). The aim of this comparison was to assess whether LLMs can "understand" the content of the dictionary and semantically determine which information belongs in which field. This task is particularly challenging because the entries are written in multiple scripts and languages.

The input prompt (Table 1) was executed on a.docx file that contained entries *idêže* (with attested English equivalents: *where, when, to wit, because of*), idolovъ (*of idol*), and idusъ (*Ides*) (the file also contained examples for the compilation of the entry *igra*)[11].

It should be noted that repeating an identical input in a new chat session can produce variations in the AI output, hereby hindering an objective comparative analysis. However, in this paper, we elected to ground our analysis of this task solely in the first response. For the set of headwords used to test the LLMs, we selected one headword with a complex structure, encompassing multiple meanings and usages, as well as headwords marked with the label *exhaustum*

[11] https://drive.google.com/file/d/1pLz-k6g0llPzwc9VYkeuR0dG01xV1Kxc/view?usp=sharing (Retrieved May 13, 2026)

TABLE 1 The input prompt

Analyse the following historical dictionary material and infer the optimal field structure for an electronic dictionary entry. The material is multilingual and multiscriptal, using Angular Glagolitic, Old Cyrillic, Latin script, and Greek alphabet, and includes Croatian Church Slavonic, Latin, Greek, modern Croatian, and English. Your task is to: 1. identify the fields needed for a digital entry, 2. organize them hierarchically, 3. distinguish entry-level, sense-level, and citation-level information, 4. explain the function of each field, 5. note ambiguities and editorial challenges, 6. avoid inventing information not supported by the material. Then, using the same model, analyse the supplied lexical evidence for one word and: - determine its meaning(s) in modern Croatian, - propose English and Latin equivalents, - identify Greek equivalents if explicitly present, - mark uncertainties clearly. Output should be structured, explicit, and conservative.

(exhaustively analysed), i.e. entries for which all attested examples are listed in the dictionary entry (entries that have less than 8 attested examples). All LLMs, as well as the structure devised by the human editors, share several core components: the structure includes the headword, part of speech, basic morphological classification and grammatical features, variants, cross-references, senses, equivalents in Croatian, English, Greek, and Latin, citations, source sigla, and other text witnesses. Among the evaluated models, Gemini and Claude Opus produced results that most closely aligned with the conceptual framework developed by the human editors. The remaining LLMs tended to generate highly elaborate structures containing several redundant elements, such as repeated specification of script and language despite these already being evident from the entry itself, or multiple repetitions of the headword. Nevertheless, all generated outputs proved valuable to the editors and authors in reconsidering the organization of the RCJHR electronic database, as the results were consistently well-structured and clearly presented[12]. Due to space limitations, the conducted tests will not be presented in detail.

6 Compiling New Entries

As the main task, we instructed the LLMs to produce a complete lexicographic entry of the word *igra* based on the previously mentioned dictionary entries. First, the models were asked to perform this using only a small number of entries used for the previous task. For the headword to be processed independently by the LLMs, we also selected an *exhaustum* entry *igra* so that the LLMs would have access to the same complete set of examples available to the human lexicographer during analysis. The dictionary entry compiled by a human lexicographer distinguishes three separate senses (1. *revelry, carousing*, 2. *dancing, dance*, 3. *jest, joke*[13]) along with one subsense under the first sense (*casting lots, gambling*). The entry also includes an example in

[12] https://drive.google.com/file/d/1M2fYDavu6zhV-sPCNxhBZx0jtTEOPxBU/view?usp=sharing (Retrieved May 13, 2026)

[13] Since the RCJHR also provides English equivalents, those words are used here in the semantic descriptions.

which the word appears as an inappropriate translation of a Latin word; however, this example has been excluded from the present analysis. Deepseek, Gemini, Qwen and Claude Opus gave no response to this part of the task. Both ChatGPT and Copilot divided the meanings into five categories (ChatGPT: 1. *revelry, carousing*; 2. *game, amusement, play*; 3. *dance, dancing*; 4. *mockery, plaything, object of ridicule*; 5. *buffoonery, scurrility*; Copilot: 1. *reveling, carousing*; 2. *play, amusement(s)*; 3. *dance, choral dance*; 4. *mockery, ridicule, (to be) a plaything*; 5. *scurrility/buffoonery? (uncertain)*). Claude Opus divided the meanings into three categories (1. *revelry, carousing*; 2. *game, spectacle*; 3. *dance, round-dance*), and its entry came closest to the original one compiled by a human lexicographer[14].

The same task was subsequently repeated with LLMs; however, instead of providing a small set of headwords, an entire fascicle of the RCJHR (excluding the word *igra*) was uploaded into the systems[15]. For each LLM, a link to the corresponding conversation is provided. A total of four criteria were selected for the analysis of the dictionary entries produced by LLMs.

The first criterion concerns the structure of the headword. Gemini[16] provided all the required elements (grammatical information and sigla – short identifying label or abbreviation used to refer to a manuscript) but gave an incorrect genitive form and omited an alternative orthographic variant. Claude Opus[17] produced a fully accurate headword containing all required elements. DeepSeek[18] included correct grammatical information but omited sigla and alternative orthographic variants. ChatGPT[19] provided correct grammatical information and orthographic variants; however, in the sigla it incorrectly isolated one breviary as a separate category. Qwen[20] supplied correct grammatical information and variants and listed the sigla of all manuscripts in which the word is attested; however, this would require further editorial intervention, since in cases such as this – where the word is attested in three breviaries – only the siglum "Br" should be used, rather than listing individual manuscript sigla as done by Qwen. Copilot[21] provided an incorrect genitive form and included explanatory material when listing variants; such explanations would need to be removed prior to publication, as they are not standard in RCJHR entries. It also listed sigla in the same way as Qwen, i.e., providing more detail than necessary. Within this category, Claude Opus proved to be the most successful and useful.

The second criterion of analysis concerns the number of senses. The original dictionary entry distinguishes three senses. Gemini did not differentiate between senses, instead subsuming them under a single, broad category ("*play, dance, revelry*"). Since the English equivalents themselves are not synonymous, this solution is considered inadequate. All other LLMs – Claude Opus, DeepSeek, ChatGPT, Qwen, and Copilot – identified five senses, exceeding the number established by the human editor. These senses are largely consistent across the models and appear to derive from the Latin equivalents, typically reflecting a *one example – one sense* correspondence. However, it seems plausible that the senses identified by most models as *jest,*

14 https://drive.google.com/file/d/1a8lYXXQNLom4yYdiE0ucTTIhRKU0rCPa/view?usp=sharing (Retrieved May 13, 2026)

15 https://drive.google.com/file/d/1Vd_XWo-jhWvOm1n8lt_85U_Y4PuSxMMW/view?usp=sharing (Retrieved May 13, 2026)

16 https://drive.google.com/file/d/1b6q38DTYdQ0oIhmKXVNgAuYA3m2J464Z/view?usp=sharing (Retrieved April 20, 2025)

17 https://claude.ai/share/8969e7cf-50aa-40ad-b967-3ef0f3a99ac1 (Retrieved April 20, 2025)

18 https://chat.deepseek.com/share/piypu2gg04793w5sy4 (Retrieved April 20, 2025)

19 https://chatgpt.com/share/6a045527-930c-8386-8f01-7fa9684723e0 (Retrieved April 20, 2025)

20 https://chat.qwen.ai/s/t_ebe33b9e-02cb-46ca-8e84-6d2fb7f5ca85?fev=0.2.46 (Retrieved April 20, 2025)

21 https://drive.google.com/file/d/1l3XJKaFoWxz4CjJrL19LDqa1iLUrw6Ni/view?usp=sharing (Retrieved April 20, 2025)

scurrility, and *plaything, laughingstock* could be consolidated into a single sense. In any case, although the models distinguished more senses than the human editor, and although it also appears that certain senses could be merged, we – as lexicographers who also compile the RCJHR – consider the generated output to be a valuable starting point for the development and refinement of sense descriptions.

The third criterion of analysis concerns the equivalents provided in Croatian, English, Latin, and Greek. At this level, artificial intelligence appears highly useful for lexicographic work, as most models identified foreign-language equivalents with a high degree of accuracy. This is particularly beneficial in the case of English equivalents, which can sometimes pose difficulties for editors who are not native speakers of English. Moreover, Latin and Greek equivalents assist compilers without a background in classical philology and facilitate the transcription of Greek terms. The only exception is Copilot, which did not provide Latin and Greek nouns in the nominative case, representing a clear shortcoming. Due to the analysis performed by LLMs, it was also observed that the original human-compiled entry omitted one Latin equivalent (*ludus*).

The final criterion of analysis concerns cross-references to other dictionary entries (synonyms). Claude Opus is the only model that provided separate cross-references at the end of each individual sense, a practice that, although not followed by the human editor in this case, occurs in lexicographic work. The selected synonyms appear to be largely appropriate. In addition, it included general synonyms at the end of the entry that apply to the headword. All LLMs exhibited a recurring error in that they included a cross-reference to the verb *igrati se* ("to play"), whereas the convention in the RCJHR is to refer exclusively to lexemes belonging to the same part of speech. Nevertheless, this also appears to be potentially very useful for the human editor.

At first glance, the models produced similar outputs; however, the comparison with the original entry shows that none of them successfully captured its semantic architecture. The models tend to err in the same direction, namely by excessively splitting senses and overlooking specific subtypes as well as marginal categories. On the other hand, the models still provided outputs that may be useful to editors in the revision process. Some LLMs also explicitly indicated uncertainty in some segments. Fully automated processing of the entry is not feasible without final human editing and lexicographic supervision; however, LLMs can significantly accelerate the initial stage of preparation. Table 2 summarises the comparative performance of the tested models across the four tasks.

TABLE 2 Summary of model performance across the four tasks

Model	OCR (printed fascicles)	HTR (handwritten cards)	Entry structure	Entry compilation
ChatGPT (GPT-5.4 Thinking)	moderate; better than Copilot	Latin script relatively accurate, Cyrillic only partially; not yet practically usable	overly elaborate, redundant elements	correct grammar and variants; error in sigla; oversplit senses (5); good equivalents
DeepSeek-V3.2	unreliable for specific Latin characters (jat, reduced vowels)	unusable	overly elaborate, redundant elements	correct grammar, but omitted sigla and variants; oversplit senses (5); good equivalents

TABLE 2 (*CONTINUED*)

Model	OCR (printed fascicles)	HTR (handwritten cards)	Entry structure	Entry compilation
Gemini 3 Pro	very good (Latin, Cyrillic, Greek)	best results, only minor errors	closest to the editors' framework	incorrect genitive, omitted variant; no sense differentiation; good equivalents
Claude Opus 4.6	very good (Latin, Cyrillic, Greek)	very good; expands abbreviations, but failed to recognise the target lemma	closest to the editors' framework	best overall: fully accurate headword; senses closest to the human entry; only model with per-sense cross-references
Qwen3.6-Plus	generally good; confuses ъ/ь, occasionally renders Latin text in Cyrillic	somewhat better than ChatGPT	overly elaborate, redundant elements	correct information but overly detailed sigla; oversplit senses (5); good equivalents
Copilot GPT-5	poorest results (Latin misreadings, Greek and subscript errors)	unusable	overly elaborate, redundant elements	incorrect genitive, superfluous explanations; equivalents not in the nominative

*None of the models recognised the Glagolitic script.

7 Conclusion

LLMs have proven highly successful in the OCR processing of previously published fascicles of the RCJHR and in the HTR processing of the handwritten index cards on which the RCJHR is based. They have also proven to be a valuable tool in identifying the most effective approaches to the retro-digitization, digitization, and modernization of the RCJHR. The advantages of using LLMs for compilation of the RCJHR include the ability of LLMs to identify Greek and Latin equivalents from parallel texts, suggest ways of distinguishing between senses, and provide English equivalents. In this way, it offers the human lexicographer a potential framework that can help accelerate the process. As a result, lexicographic work becomes increasingly akin to editorial work. Given the complexity of structure of the RCJHR, its highly specialized nature, and the demanding workflow of lexicographic practice, AI software cannot replace the human editor; however, it can provide meaningful support and accelerate the process. LLMs cannot identify parallel passages in other text witnesses or verify Glagolitic variants, as relevant sources are dispersed across various online databases, libraries, and physical collections – such as those held by the Old Church Slavonic Institute, now part of the Institute for the Croatian language – and are often written in the Glagolitic script. Furthermore, the precise determination of word meanings frequently requires consultation of multiple dictionaries, including older editions available only in print. Nevertheless, LLMs can be highly useful in reflecting on semantic distinctions, suggesting possible sense divisions, and expediting the analytical process. Their greatest utility lies in identifying equivalents in other languages, while they also facilitate the retrieval of potential synonyms for cross-referencing. The analysis also shows that, for certain tasks, it is useful to test multiple LLMs, since different models perform best on different tasks and no single model consistently achieves the best results across all tasks.

References

Al-Wesabi, F. N., Alshahrani, H. J., Osman, A. E. & Abd Elhameed, E. S. (2023). Low-resource language processing using improved deep learning with hunter–prey optimization algorithm. *Mathematics*, 11(21), Article 4493. https://doi.org/10.3390/math11214493

Anthropic. (2026). *Conversation with Claude*. Retrieved April 20, 2026, from https://claude.ai/share/8969e7cf-50aa-40ad-b967-3ef0f3a99ac1

Artificial Analysis. (2026). *Models comparison: Intelligence, performance & price analysis*. Retrieved April 15, 2026, from https://artificialanalysis.ai/models

Balenović, K. & Proroković, J. (2025). The lexicographic potential of artificial intelligence: A case study of English loanwords in the Croatian language. *Studia lexicographica*, 19(36), 39–64. https://doi.org/10.33604/sl.19.36.3

Bender, E. M., Gebru, T., McMillan-Major, A. & Mitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In M. C. Elish, W. Isaac & R. S. Zemel (Eds.), *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)* (pp. 610–623). Association for Computing Machinery. https://doi.org/10.1145/3442188.3445922

Brynjolfsson, E., Li, D. & Raymond, L. R. (2023). *Generative AI at work* (NBER Working Paper No. 31161). National Bureau of Economic Research. https://doi.org/10.3386/w31161

Cai, B., Ng, C., Liang, D. & Hotama, S. (2024). Low-cost generation and evaluation of dictionary example sentences. In K. Duh, H. Gomez & S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 3538–3549). Association for Computational Linguistics.

Caswell, A. (2026). I put Claude vs. DeepSeek through 7 real-world tests — and one is clearly better. *Tom's Guide*. Retrieved April 15, 2026, from https://www.tomsguide.com/ai/i-put-claude-vs-deepseek-through-7-real-world-tests-and-one-is-clearly-better

de Schryver, G.-M. (2023). Generative AI and lexicography: The current state of the art using ChatGPT. *International Journal of Lexicography*, 36(3), 355–387. https://doi.org/10.1093/ijl/ecad021

de Schryver, G.-M. (2025). Customised GPTs for lexicography. *Lexikos*, 35(2), 397–422. https://doi.org/10.5788/35-2-2104

DeepSeek. (2026). *Conversation with DeepSeek*. Retrieved April 20, 2026, from https://chat.deepseek.com/share/piypu2gg04793w5sy4

Elsevier. (2026). *Scopus advanced search*. Retrieved April 15, 2026, from https://www.scopus.com/search/form.uri?display=advanced

Evert, S., Ganslmayer, C. & Rink, C. (2024). Multi-level analysis as a systematic approach to evaluating the quality of AI-generated dictionary entries. In K. Š. Despot, A. Ostroški Anić & I. Brač (Eds.), *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress* (pp. 317–335). Institute for the Croatian Language.

Filipović Petrović, I. & Beliga, S. (2025). Can AI understand Croatian idioms? Assessing large language models in lexicographic tasks. *Prispevki za novejšo zgodovino*, 65(3), 219–242. https://doi.org/10.51663/pnz.65.3.09

Google. (2026). *Conversation with Gemini*. Retrieved April 20, 2026, from https://drive.google.com/file/d/1b6q38DTYdQ0oIhmKXVNgAuYA3m2J464Z/view?usp=sharing

Google. (2026). *HTR of one handwritten paper slip from the corpus*. Retrieved May 13, 2026, from https://drive.google.com/file/d/1p2Vw02DgQ2IqXSYDIc15oTLnnN9BLRoW/view?usp=sharing

Google. (2026). *OCR of one Dictionary page*. Retrieved May 13, 2026, from https://drive.google.com/file/d/1p2Vw02DgQ2IqXSYDIc15oTLnnN9BLRoW/view?usp=sharing

Greco, A., Delsanto, M., Di Fabio, A., Mori, L., Onesti, C., Radicioni, D. P. & Scozzaro, C. J. (2025). Compiling bilingual dictionaries: AI-assisted translation of Italian multiword expressions into English and French. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda &

A. Salgado (Eds.), *Proceedings of the Electronic Lexicography in the 21st Century Conference* (eLex 2025) (pp. 702–728). Lexical Computing CZ s.r.o.

Jiang, Q., Gao, Z. & Karniadakis, G. E. (2025). DeepSeek vs. ChatGPT vs. Claude: A comparative study for scientific computing and scientific machine learning tasks. *Theoretical and Applied Mechanics Letters*, 15, 100583. https://doi.org/10.1016/j.taml.2025.100583

Klenovar, M., Vela, J. & Nazor, A. (Eds.). (2021). *23. sveščić Rječnika: i – izgnilstvo*. Staroslavenski institut. Retrieved May 13, 2026, from https://drive.google.com/file/d/1Vd_XWo-jhWvOm1n8lt_85U_Y4PuSxMMW/view?usp=sharing

Lakhani, K. R. (2023). AI won't replace humans — but humans with AI will replace humans without AI. *Harvard Kennedy School, Growth Lab*. Retrieved April 20, 2026, from https://www.hks.harvard.edu/centers/mrcbg/programs/growthpolicy/ai-wont-replace-humans-humans-ai-will-replace-humans-without-ai

Levine, O. (2026). When software engineering stopped being the safe bet. *Greenfield Growth*. Retrieved April 20, 2026, from https://www.greenfield-growth.com/blog-posts/when-software-engineering-stopped-being-the-safe-bet

Lew, R. (2023). ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences Communications*, 10, Article 704. https://doi.org/10.1057/s41599-023-02119-6

LM Council. (2026). *AI model benchmarks Apr 2026: Compare GPT-5, Claude, Gemini, and more*. Retrieved April 15, 2026, from https://lmcouncil.ai/benchmarks

Local AI Master Research Team. (2026). *Best AI coding models 2026: SWE-bench leaderboard (Claude, GPT-5, Gemini)*. Retrieved April 15, 2026, from https://localaimaster.com/models/best-ai-coding-models

Lu, H., Yang, H., Huang, H., Zhang, D., Lam, W. & Wei, F. (2024). Chain-of-dictionary prompting elicits translation in large language models. In Y. Al-Onaizan, M. Bansal & Y.-N. Chen (Eds.), *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing* (pp. 958–976). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.emnlp-main.55

Merx, R., Vylomova, E. & Kurniawan, K. (2024). Generating bilingual example sentences with large language models as lexicography assistants. In T. Baldwin, S. J. Rodríguez Méndez & N. Kuo (Eds.), *Proceedings of the 22nd Annual Workshop of the Australasian Language Technology Association* (pp. 64–74). Association for Computational Linguistics.

Microsoft. (2026). *Conversation with Copilot*. Retrieved April 20, 2026, from https://drive.google.com/file/d/1l3XJKaFoWxz4CjJrL19LDqa1iLUrw6Ni/view?usp=sharing

Mihaljević, A. & Mihaljević, J. (2024a). Digitalizacija i igrifikacija Rječnika crkvenoslavenskoga jezika hrvatske redakcije. *Лексикографски сусрети*, 1, 305–317. https://doi.org/10.18485/lexicog_meet.2024.1.ch26

Mihaljević, A. & Mihaljević, J. (2024b). Mrežna inačica rječnika crkvenoslavenskoga jezika hrvatske redakcije. *Slovo: časopis Staroslavenskoga instituta u Zagrebu*, 74, 169–194. https://doi.org/10.31745/s.74.10

Mihaljević, A., Mihaljević, M. & Šimić, A. (2025). *Glagoljica za znatiželjne*. Staroslavenski institut.

Mihaljević, M. (Ed.). (2014). *Hrvatski crkvenoslavenski jezik*. Hrvatska sveučilišna naklada – Staroslavenski institut.

Nazor, A. (2008). Rječnik crkvenoslavenskoga jezika iniciran na IV. međunarodnom slavističkom kongresu u Moskvi 1958. godine (u povodu 50. obljetnice inicijative). In M. Samardžija (Ed.), *Vidjeti Ohrid. Referati hrvatskih sudionika i sudionika za XIV. međunarodni slavistički kongres* (pp. 65–82). Hrvatsko filološko društvo – Hrvatska sveučilišna naklada.

Nazor, A. (2017). O početcima rada na Rječniku crkvenoslavenskoga jezika hrvatske redakcije. *Filologija*, 68, 117–125.

OpenAI. (2026). *Conversation with ChatGPT*. Retrieved April 20, 2026, from https://chatgpt.com/share/6a045527-930c-8386-8f01-7fa9684723e0

Pauklin, P. (2026). AI-generated code statistics 2025: Can AI replace your developers? *Netcorp Software Development*. Retrieved April 20, 2026, from https://www.netcorpsoftwaredevelopment.com/blog/ai-generated-code-statistics

Perak, B. (2025). *Komunikacija u doba umjetne inteligencije: Razvoj velikih jezičnih modela i komunikacijskih agenata*. Filozofski fakultet, Sveučilište u Rijeci.

Qwen. (2026). *Conversation with Qwen*. Retrieved April 20, 2026, from https://chat.qwen.ai/s/t_ebe33b9e-02cb-46ca-8e84-6d2fb7f5ca85?fev=0.2.46

Rabus, A. & Thompson, W. (2023). Performance of generic HTR models on historical Cyrillic and Glagolitic: Comparison of engines. *Scripta & e-Scripta*, 23, 11–34.

Romein, C. A., Rabus, A., Leifert, G. & Ströbel, P. B. (2025). Assessing advanced handwritten text recognition engines for digitizing historical documents. *International Journal of Digital Humanities*, 7(1), 115–134. https://doi.org/10.1007/s42803-025-00100-0

Šimić, A. (2017). Mikrostruktura i istraživački izazovi obrade rječnika crkvenoslavenskoga jezika hrvatske redakcije. *Filologija*, 69, 99–128.

Similarweb. (2026). *Top AI chatbots and tools websites ranking*. Retrieved April 15, 2026, from https://www.similarweb.com/top-websites/ai-chatbots-and-tools/

Staroslavenski institut. (2026). *Analysis of the structure of the dictionary entry created by AI programs*. Retrieved May 13, 2026, from https://drive.google.com/file/d/1M2fYDavu6zhV-sPCNxhBZx0jtTEOPxBU/view?usp=sharing

Staroslavenski institut. (2026). *Development of the Digital Infrastructure Model of the Old Church Slavonic Institute (DigiSTIN)*. Retrieved April 15, 2026, from https://stin.hr/sadrzaj/digistin/

Staroslavenski institut. (2026). *Dictionary entries from the Dictionary of the Croatian Redaction of Church Slavonic that were used for AI testing*. Retrieved May 13, 2026, from https://drive.google.com/file/d/1pLz-k6g0llPzwc9VYkeuR0dG01xV1Kxc/view?usp=sharing

Staroslavenski institut. (2026). *First Task Results*. Retrieved May 13, 2026, from https://drive.google.com/file/d/1a8lYXXQNLom4yYdiE0ucTTIhRKU0rCPa/view?usp=sharing

Staroslavenski institut. (2026). *HTR outputs produced by LLMs*. Retrieved May 13, 2026, from https://drive.google.com/file/d/1vNQJfurTg5PPMcPRTSF0RpDqoUrdfri0/view

Staroslavenski institut. (2026). *OCR outputs produced by LLMs*. Retrieved May 13, 2026, from https://drive.google.com/file/d/1d2WupesHBN-YyyglfsRh1NGONh05-mC4/view

Staroslavenski institut. (2026). *Rječnik crkvenoslavenskoga jezika hrvatske redakcije*. Retrieved April 15, 2026, from https://rjecnik.stin.hr/

SWE-bench. (2026). *SWE-bench leaderboards*. Retrieved April 15, 2026, from https://www.swebench.com/

Terzimehić, N., Bühler, B. & Kasneci, E. (2025). Conversational AI as a catalyst for informal learning: An empirical large-scale study on LLM use in everyday learning. *arXiv*. https://arxiv.org/abs/2506.11789

Utkirova, S. (2024). Recent developments in lexicography: The case of artificial intelligence. *European Journal of Innovation in Nonformal Education*, 4(8), 71–76.

Contact information

Ana Mihaljević
Institute for the Croatian Language
amihaljevic@stin.hr

Josip Mihaljević
Institute for the Croatian Language
jmihaljevic@stin.hr

Acknowledgements

This work was created within the project *Development of a Digital Infrastructure Model of the Old Church Slavonic Institute – DigiSTIN*, funded by the European Union – NextGenerationEU. The views and opinions expressed are solely those of the authors and do not necessarily reflect the official positions of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

Krzysztof Nowak, Iwona Krawczyk & Jagoda Marszałek

DIALEXPOL

LLM-Assisted Sense Assignment for a Diachronic Dictionary of Latin

Abstract The DiaLexPoL project is building a diachronic lexical database of Neo-Latin (ca. 1550–1800) from the semi-automatically compiled diachronic corpus of Polish Latin (DiaCorPoL). Its principal bottleneck is sense assignment: referring early-modern attestations to an inventory designed for the Dictionary of Medieval Latin in Poland (DMLP). We ask whether deployable LLMs – cheap commercial mid-tier and self-hostable open-weight models, rather than frontier systems a chronically underfunded project cannot reproduce – can draft this step in a human-in-the-loop workflow. Two experiments address it: reproducing DMLP's own sense assignments on its quotations, and assigning Neo-Latin concordance lines to the DMLP inventory against independent human gold-standard annotation. Self-hostable open-weight models give the strongest drafts, outperforming the single commercial mid-tier model at negligible cost, and the best corpus configuration matches human inter-annotator agreement overall, though not on every lemma. Gains over a most-frequent-sense baseline – always assigning a lemma's commonest sense – are modest and concentrate on genuinely polysemous lemmas, where editors most need help, while on dominant-sense lemmas the models over-split, marking where LLM drafting helps.

Keywords Latin; Neo-Latin; LLMs; historical lexicography; post-editing lexicography; word sense disambiguation

1 Introduction

Latin was used in Polish territory from the early 11th century in administration, historiography, science, literature, and religious practice. Although its use began to wane in the 16th century as Polish replaced it in many contexts, it left an extensive body of texts, much of it still in manuscript. The medieval period is now relatively well documented – by the eFontes corpus (https://corpus.scriptores.pl) and the Dictionary of Medieval Latin in Poland (DMLP) – but the language as used between roughly 1550 and 1800, despite its central role in scholarly and literary communication, has yet to receive comparable treatment, for want of basic tools: there is still no electronic corpus and bibliographic coverage is at best patchy. The project we introduce addresses this gap by assembling a semi-automatically built diachronic corpus (DiaCorPoL) and a database of Neo-Latin lexis (DiaLexPoL). In the first stage, DiaLexPoL is planned to include grammatical and semantic information about ca. 1 million word occurrences in the parallel DiaCorPoL corpus. It is conceived as a diachronic complement to the DMLP and, at least initially, will build on its sense inventory as a baseline. But the DMLP's definitions and its deeply nested, print-era entry structure remain to be reconceptualised for corpus-based use, and whether to build on that model or develop a new one is not yet settled. Either way, assigning occurrences to existing definitions and creating new ones in case of semantic drift are natural bottlenecks of the project.

LLMs offer an obvious way to draft this step, and can be involved across the workflow from HTR (handwritten text recognition) correction through morphosyntactic parsing to sense disambiguation. But this comes at a cost – financial, since frontier models are priced steeply for chronically under-resourced academic projects, and non-financial, since heavy reliance on them is environmentally costly and adds organisational pressure on already strained lexicographic teams. We therefore ask how far sense assignment can be drafted reproducibly by reusing an existing dictionary as a sense repository and employing mid-tier, self-hostable

open-weight models rather than flagships – models a budget historical-lexicography project can actually run, maintain, and reproduce.

We test this in two experiments, which also inform the open question of whether the DMLP model can be kept. The first is dictionary-internal: we ask whether deployable models can reproduce DMLP's own sense assignments on its quotations. The second experiment moves to a more realistic scenario, assigning concordance lines from a 17th-century Neo-Latin corpus to the DMLP inventory against independent human gold-standard annotation. Within our single-domain corpus the target lemmas prove to be attested largely in senses continuous with that inventory; therefore, detecting genuinely diverged senses needs a broader, multi-domain corpus, which we take up in a companion study. To support reproducibility, we release the prompts together with both labelled datasets – the Experiment 1 dictionary quotations and the Experiment 2 concordance lines – in a public repository[1].

2 Background

Lexicography has long adopted advances in NLP to cope with corpus scale and to dispense with its most routine tasks, so it is no surprise that lexicographers were among the early adopters of LLMs. One approach that has emerged is not to replace expert work end to end but to use language models for drafting while leaving post-editing to humans (Lew, 2024). Jakubíček & Rundell (2023) add a further point: even though a general-purpose model can rival the dedicated post-editing tools currently in use, lexicography should still pursue task-specific solutions – models tuned for particular subtasks – rather than rely on a general system. Many studies applying LLMs in lexicography followed (Tiberius et al., 2024; Widmann, 2025), with Lugli (2025) recently demonstrating a sample workflow in practice: using generative LLMs to automate semantic annotation for a corpus-driven dictionary of a low-resource historical language. Many such applications rely on commercial frontier models, with the cost, reproducibility, and long-term availability this entails left largely out of view. Our paper calls this dependence on flagship models into question.

We test LLMs on word sense disambiguation, a core task of dictionary-making, and one open to the lexicographer's subjective judgement. We follow the framing of Meconi et al. (2025), which asks whether LLMs understand word senses and separates sense discrimination from definition generation. Our task is primarily the former — assigning a use to an existing inventory — with the 0-assignment, where the model signals that no listed sense fits; proposing a definition for such cases is left to a later study.

Navigli (2026) argues that WSD is not obsolete in the LLM era but has taken on a new role: less a stand-alone module and more a diagnostic for how well models handle word meaning. In our setting the model must map a use onto the sense structure of a specific medieval dictionary, rather than produce the general gloss, so we can test how close it comes to human lexical competence, and the error patterns we describe (§5.2, §6.3) support Navigli's point. Paradoxically, the same setup also tests the dictionary: because our gold is an existing sense inventory, the errors and their consistency across models point to weak or doubtful entries in the dictionary's own data. We do not improve the inventory here, but this suggests a further use for LLMs, in reviewing and updating lexicographic resources.

For historical-language WSD, and for Latin in particular, Bamman & Burns's Latin BERT (2020) remains the reference model – we specify a Latin BERT as a baseline but leave it to

[1] Code and prompts: https://doi.org/10.5281/zenodo.21626003. Datasets: https://doi.org/10.5281/zenodo.21625867. Development repository: https://github.com/novacellus/latinllms26dialexpol

future work (§8). Lendvai & Wick (2022) fine-tune it for WSD against the *Thesaurus Linguae Latinae*, using the dictionary's own nested sense structure as the label set – the design we adopt in Experiment 1. For Latin specifically, Ghinassi et al. (2024) pivot sense information from parallel corpora. Closer to ours, Ghizzota et al. (2025) evaluate zero-shot and fine-tuned LLMs on a SemEval-derived Latin sense dataset and find the gains come from fine-tuning rather than zero-shot prompting.

Our approach differs on three counts:

(i) we anchor on a published dictionary's sense inventory rather than a lexical-semantic-change benchmark;

(ii) we stay with deployable, self-hostable models in a zero/few-shot scenario rather than fine-tuning; and

(iii) we test against human gold on Neo-Latin corpus lines.

No prior study to our knowledge evaluates deployable LLMs for sense assignment on a medieval-Latin dictionary inventory applied to Neo-Latin material against human gold.

3 Data

For this study we selected 53 Latin nouns from the sociopolitical domain, spanning power (*potestas* 'power', *maiestas* 'majesty'), rulers and officials (*princeps* 'prince', *episcopus* 'bishop'), honours and estates (*nobilitas* 'nobility'), law (*lex* 'law', *libertas* 'liberty, freedom'), assemblies and institutions (*conventus* 'assembly', *comitia* 'electoral assembly'), peoples (*populus* 'people', *natio* 'nation'), and war and peace (*pax* 'peace'). Many of the terms are abstract and polysemous in ways that challenge lexicographic practice: *civitas* denotes both 'city' and the 'right of citizenship', while *officium* is both a specific 'office' and the general 'duty' attaching to it. The lemmas are all attested in the DMLP which we use as a sense inventory in Experiment 1; only some of them occur frequently enough in the small Neo-Latin corpus to be used in Experiment 2. Importantly, they are also attested in Classical and early medieval Latin, supporting a long-term diachronic study (McGillivray & Nowak, 2025).

3.1 Dictionary definitions and quotations (Experiment 1)

We drew 53 entries from the TEI XML edition of the DMLP (seven volumes, letters A–Q). Curating them proved non-trivial given the dictionary's nested structure: an initial parse yielded 912 sense nodes, of which 416 (46%) were non-definitional structural nodes (usage and domain labels, grammatical or collocational subdivisions, sense numbering) and were excluded. The remaining definitional senses were checked for parent–child relations and, where a definition was collocation-bound, reconstructed to stand alone (e.g. *COMES rusticalis* → *scultetus/sołtys* 'village headman').

The definition levels are a product of evolving editorial practice, and they are as much convenience-based grouping as claims about sense relations. Flattening sense hierarchy produces two patterns that may impact model performance and annotator confidence alike: some fine-grained senses sit at the same level as coarse ones (e.g. *comes*, ranging from 'magnate' to 'village headman'), and frequent fine referents are now buried in a general sense (*princeps*, where 'emperor', 'king', and 'duke' collapse into a single 'ruler, prince, or king'). The curated inventory that accompanies this study records, per sense, the lemma, a sense ID, Polish and

Latin definitions, and the sense's hierarchical level (the experiments, however, use the first-level glosses only).

Quotations in the DMLP reflect its differential character: it quotes chiefly what medieval usage adds to or diverges from Classical Latin. Therefore, quotation counts are not usage counts because a sense that is common in real texts, including the inherited classical one, may carry few quotations in the DMLP. The most-frequent-sense (MFS) baseline (§4.2) must be read with this in mind. The 53 extracted entries held 1,304 quotations. Curation removed 170 of them – 39 not tied by the dictionary to a single sense, 108 collocational or fragmentary, and 23 on annotator review – leaving 1,134 usable quotations across 52 lemmas (the two homonymous entries for *comitia* were merged into a single lemma with three senses, namely 'county', 'assembly', and 'elections'). The quotations are generally short (median 9 words; 8.7% under five), because the lexicographers abbreviate heavily. Three further interventions were needed before Experiment 1 could run: (i) elided headwords were expanded (*c-em* → *c[ivitat]em*); (ii) ellipsis marks (. . .) were removed; and (iii) the *sc.* comprehension notes (from Lat. *scilicet* 'namely'), which the lexicographers added to supply context, were stripped — a removal that can reduce a quotation's intelligibility, a difficulty that resurfaces in the error analysis (§5.2). We restrict the set to the 38 lemmas with ≥10 quotations, below which per-lemma accuracy and the MFS baseline are too unstable to interpret, totalling 1,041 usable quotations against 303 L1 senses (mean 8 senses per lemma; range 3–23).

3.2 Sense annotated concordance lines from a Neo-Latin corpus (Experiment 2)

Experiment 2 draws on four 17th-century Neo-Latin prose works on the state, law, and warfare, totalling roughly 416,000 words. The texts were edited by Iwona Krawczyk and Jagoda Marszałek, who also provided high-quality manual transcriptions (evaluation on noisier automatic transcription is left to future work). They are only lightly processed, preserving original orthography, and used raw – with no prior lemmatisation or POS tagging – to gauge whether the lexicographic pipeline would need such pre-annotation. Concordance lines for each lemma are extracted by regular expression, manually cleared of false positives (chiefly proper names and derived forms), and sampled under a fixed seed; annotators and models alike see the KWIC line – the full sentence containing the target word.

Of the 53 terms (§3.1), we select 14 for Experiment 2. The number was set by the annotation budget: it keeps each annotator near their ceiling of roughly 1,000 records while leaving enough records per lemma for a reliable per-lemma agreement estimate. The selected lemmas fall into two groups. The first (Tier A) contains eight lemmas — *civitas, populus, dignitas, imperator, natio, episcopus, comitia, praefectus* — that had too few dictionary quotations to be used in Experiment 1, but which in Experiment 2 let us ask whether a medieval inventory still frames early-modern usage. The second group of six (Tier B) — *officium, lex, princeps, libertas, bellum, dux* — are drawn from the quotation-rich entries used in Experiment 1. Four of them (*bellum, dux, lex, princeps*) were among Experiment 1's held-out (Phase B) lemmas and so can be compared across the two experiments on independent data; the other two (*officium, libertas*) served as Phase A development lemmas.

Concordance lines are allocated by group: 80 per Tier A lemma and 58 per Tier B lemma, 988 in total. The draw enforces a five-word minimum and takes one record per sentence (a single occurrence even where a sentence contains several) so that agreement is computed on independent items. A fixed seed makes the sample reproducible. Source dominance is unbalanced (*civitas* draws ≈50% from one author), and because the corpus is single-domain each lemma

likely appears in only a narrow subset of its senses — which raises the most-frequent-sense baseline the accuracies must clear and motivates a broader corpus in future work.

TABLE 1 The four 17th-century Neo-Latin works in the Experiment 2 corpus

Author	Work	Year	Genre (subject)	Words
A. M. Fredro	*Militaria*	1668	treatise (military, economy)	≈58,000
A. M. Fredro	*Vir Consilii*	1673	treatise (rhetoric, political system)	≈158,000
K. Hartknoch	*De Rep. Pol.*	1678	treatise (history, law)	≈152,000
J. Sachs	*De Scopo Reip. Pol.*	1665	polemic (political system, law)	≈48,000

Two annotators independently labelled all 988 records in Argilla[2], marking for each:

- match_status (is the word genuinely an instance of the target lemma);
- quote_quality (is there enough context to judge the sense);
- a sense_id (a listed sense, or 0 with a required note where none fits) and
- a confidence rating (1–5).

A record is analysable only where both annotators judge it a valid match with an assessable quote; 837 of the 988 records met this bar. Conventions were fixed not by a written codebook but in a joint working meeting. A single gold was resolved by consensus where annotators agreed; disagreements were adjudicated by a curator (Krzysztof Nowak), who also reassigned mislabelled 0s to their covering senses. Remaining 0s, now pointing to the real definitional gaps, were categorised as semantic novelty, insufficient context, or model error. Realised counts and inter-annotator agreement are reported in §6.1 and §6.2.

4 Experimental Design

4.1 Shared task, prompt structure, and variables

Both experiments share the assignment task and four prompting templates – zero-shot, few-shot, structured-justification (the model explains its choice), and decomposed (the choice split into steps). The senses are given as a flat numbered list (1. `<def> ... N. <def>`), the model returns an integer (0 = no sense fits), the inventory order is shuffled per item under a fixed seed to control positional bias, and the target word is marked `<WSD> ... </WSD>`.

Experiment 1 varies two axes – prompting template and definition language (Latin, Polish) – using plain text output and only the first-level senses, fixed by prior screening (§5.1). The effect of sense granularity level is left to future work. Experiment 2 keeps the four templates and the definition-language axis, fixes output to plain and granularity to first-level senses, and adds two more axes: context length (the KWIC concordance line only vs the line plus the previous sentence) and source metadata (author, title, genre, period; on vs off). Combining the four templates, two definition languages, two context lengths, and two metadata settings gives 32 configurations, run for all five models on the 837 records (§6.1). The main comparison uses the eight template × definition-language configurations, with context length and source

[2] https://argilla.io/

metadata held at their default (KWIC-only, metadata off); the effect of those two axes is examined separately in §6.3.

4.2 Baselines and metrics

Following Meconi et al. (2025), we take the per-lemma most-frequent sense (MFS) as the floor: a configuration that does not clear it convincingly is not a finding. As shown below, where one sense dominates a lemma, MFS is hard to beat and effectively becomes a ceiling (§6.3). For every configuration we report accuracy, with 95% bootstrap confidence intervals to assess how uncertain the estimate is. We also report Cohen's κ against the gold (agreement corrected for chance, where 1.0 is perfect and 0 is chance level); macro-F1 (the precision–recall balance, averaged equally across senses so rare senses are not drowned out); and the gain over MFS in percentage points.

Throughout, macro figures average across lemmas with equal weight each (so smaller lemmas are not outweighed) and micro figures average across all occurrences (so lemmas with more records count for more). The floor differs sharply between the two experiments: macro 48.7% in Experiment 1, against a much higher macro 82.2%/micro 81.6% on the Experiment 2 gold annotations. A no-context LLM and Latin BERT (Bamman & Burns, 2020) are left as further baselines for future work (§8).

For each answer we record whether it is correct, wrong, or a format-failure (a reply from which no valid sense number can be parsed). We compare configurations using paired significance tests corrected for multiple comparisons (McNemar with Holm–Bonferroni); per-sense precision, recall, F1, and confusion matrices are reported for descriptive purposes only. Each experiment adds metrics of its own:

- Experiment 1: Spearman's ρ for rank stability between Phase A and Phase B (§4.3);
- Experiment 2: the inter-annotator κ that sets the human ceiling, a breakdown of the 0-assignment rate (novelty/insufficient context/model error), and an accuracy-against-confidence check as a validity test.

4.3 Model lineup and decoding

Experiment 1 runs in two phases. **Phase A** screens the combinations of output format (plain text vs. JSON), prompt template, and definition language, using only first-level senses, on 11 development lemmas. Every configuration whose confidence interval (CI) overlaps the leader's is promoted to the next stage. **Phase B** then reports the headline results on 27 disjoint held-out lemmas (with ≥10 quotations each). A prior pilot fixed first-level sense granularity and Phase A fixed plain output; both screening results are reported with Experiment 1 (§5.1). Experiment 2 runs single-stage on all 14 lemmas, reusing the configurations promoted in Experiment 1.

Both Experiment 1 phases run the same lineup: one commercial mid-tier anchor – gpt-4o-mini – plus nine open-weight models (Llama-3.1-8B, Llama-3.2-3B, Llama-3.3-70B, Llama-4-Maverick; Gemma-3-4B, Gemma-4-31B; Qwen3-235B-A22B; Phi-4; DeepSeek-V3). Budget constraints left only one commercial model; others are deferred to future work (§8). Llama-3.2-3B is retained only for reference as a model that essentially fails the task (≈0% accuracy, high format-failure). Experiment 2 narrows this to the five models (Gemma-4-31B, Qwen3-235B-A22B, Llama-4-Maverick, Llama-3.3-70B, DeepSeek-V3) whose Phase B accuracy

cleared Experiment 1's MFS floor (macro 48.7%) by ≥5 pp on at least one configuration: when these models answer 0, it can be read as a possible genuine gap in the inventory rather than as noise from a weak model.

Decoding is identical throughout (temperature 0, max_tokens 1024, model endpoints pinned), and prompts are assembled from a shared specification with the full templates in the repository. In total this comes to 133,920 calls for Experiment 2 (32 configurations × 5 models × 837 analysable records, §6.1), all on open-weight models. Experiment 1's compute is modest – the full Phase A grid of runs ≈ $5.2 and Phase B (35,600 calls) ≈ $1.08, with each open-weight model under $1.

5 Experiment 1: Dictionary-Internal Sense Assignment

The task is as described in §4.1: the model receives a DMLP quotation and the candidate DMLP senses for its lemma; the correct answer is the sense the dictionary itself assigned to that quotation.

5.1 Phase A screening and Phase B results

Phase A fixed the configuration. A prior pilot had already compared two levels of sense detail: the top-level senses alone (L1) versus the full nested set including sub-senses (L1+L2). The coarser L1 level worked better (using sub-senses cost 4.8 percentage points), so we adopted L1. Phase A itself then fixed the remaining choices. The output format mattered most: plain text scored 34.5% against 17.9% for JSON — a 16.6-point gap (confirmed by a paired significance test) — so we used plain text throughout. Among the four prompt templates, few-shot scored highest (34.6%), ahead of structured-justification (33.6%), zero-shot (30.8%), and decomposed (30.1%); Polish definitions slightly outscored Latin. These screening results fixed the setup carried into the Phase B test: plain-text output, L1 senses, few-shot prompting, and Polish definitions.

On the 27 held-out Phase B lemmas the best configuration – **Qwen3-235B-A22B with few-shot prompting and Polish definitions** – reaches 68.3% [63.8, 72.8] ($\kappa = 0.63$, macro-F1 = 0.61), +19.6 pp over the per-lemma MFS floor (48.7%, §4.2). A small group of mid-to-large open-weight models follows – Gemma-4-31B (67.6%) and Llama-4-Maverick (67.0%), all on Polish definitions – and 34 of the 80 model–configuration runs clear MFS (Figure 1). The single commercial model, gpt-4o-mini, comes in the middle (best ≈ 48%, below MFS in several model-configuration runs), and Llama-3.2-3B marks the floor near 0%.

Pooled across all 80 model–configuration runs, accuracy is only 46.0% – hence below MFS – but that average is uninformative, mixing the broken small models in with the leading cluster. The held-out result of Phase B is not just an artifact of how Phase A selected configurations: model rankings are near-identical across the two phases ($\rho = 0.976$) and template rankings reproduce exactly ($\rho = 1.000$). The definition-language effect carries over too – Polish beats Latin by +3.0 pp in Phase A and +2.9 pp in Phase B – and its cross-experiment behaviour is taken up in §6.4.

Models largely agree on which lemmas are hard (mean pairwise correlation $r = 0.66$), but this agreement mostly reflects the baseline: a lemma is hard precisely when no single sense dominates it. Per-lemma MFS runs from near-unbeatable lemmas, where one very frequent sense covers most uses (*hereditas* 84%, *privilegium* 83%, *bellum* 81%), down to genuinely polysemous ones where no sense dominates (*militia* 19%, *iudicium* 20%, *pax* 25%).

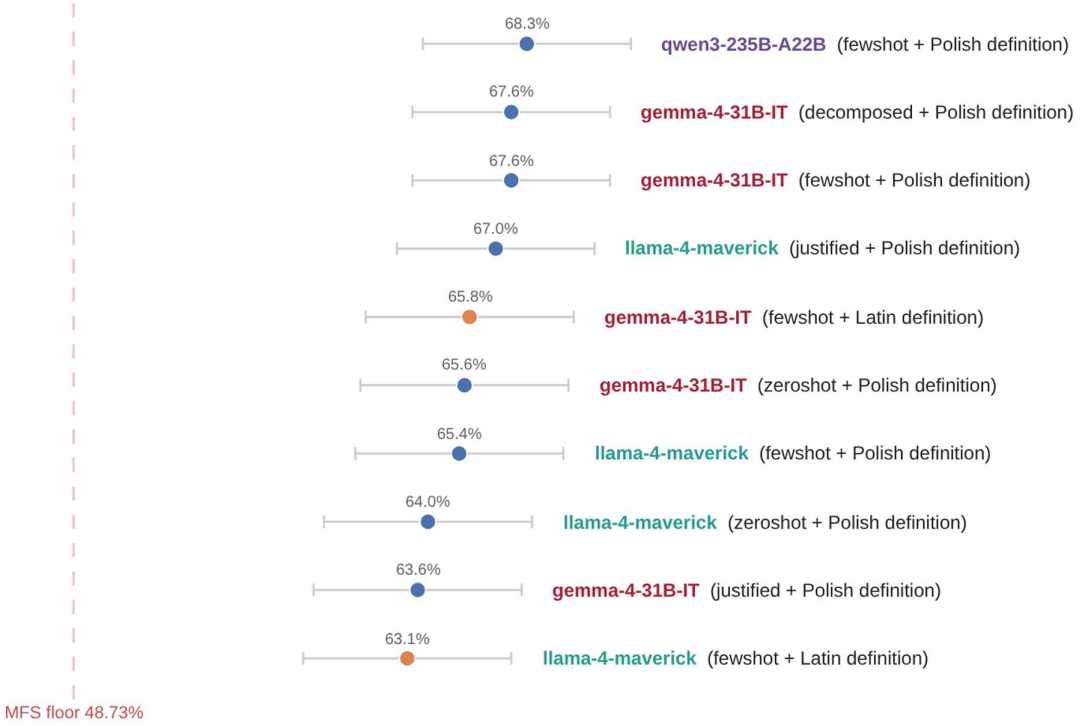


FIGURE 1 Top 10 configurations by accuracy (%) on Experiment 1 Phase B, ranked against the macro averaged MFS baseline

5.2 Error analysis

Of the 445 Phase B quotes, 93 are hard (with ≥80% of model–configuration runs wrong) — 15 of them outright consensus failures (all 80 model–configuration runs wrong) — and no quote is unanimously correct; a further 196 are a mixed bag (30–70% of model–configuration runs correct). The 15 consensus failures may be explained by:

- insufficient context, where the fragment is too short to fix the sense even for a human — and where the DMLP's own *scilicet*-glosses, which we strip, would have helped (*scilicet*-flagged quotes score −12.3 pp);
- missing external knowledge, where the disambiguator lies outside the Latin (an Old-Polish gloss, or Polish institutional realia);
- missing collocational cues, where the construction we removed from the gloss carries specific meaning that may be difficult to recover from the major-sense definition. An example is *pacem habere cum* 'cease relations with someone'', which the flattening subsumed under the major sense 'cessation, end'.
- contestable gold, where the model's reading is defensible and the gold debatable;
- and near-synonymous senses, barely separable within an entry.

Although the above typology is illustrative and not validated, it nevertheless aims to separate artifact types (insufficient context, contestable gold), which bound performance, from signal types (missing knowledge, sense granularity, collocational cues), which inform lexicographic practice (§7.1).

What drives the differences in performance between lemmas is mostly polysemy: lemmas with ≥ 10 candidate senses score far below those with ≤5 (33.7% vs 53.0%). Accuracy is also lowest (30.2%) where there are fewer than 1.5 quotations per candidate sense. We did not examine which senses these are, but given the DMLP's differential character (§3.1) a low count need not mean a genuinely rare sense — it may equally mark an inherited classical sense the dictionary does not quote — so this figure should be read with that ambiguity in mind. Quote length shows no systematic effect (long 50.1%, medium 45.2%, short 46.3%), so shortness of context is not the main driver. Genre and century were investigated too, but their effect seems small and largely a proxy for which lemma dominates each source. The most frequent confusions (*nobilitas* sense 4 → sense 3, *privilegium* 4 → 2, *bellum* 1 → 2, *decretum* 4 → 1) are adjacent fine senses within an entry – but, as below, the error mass overall is not.

Two further measurements – exploratory, and computed only on the four strongest models (Qwen3-235B-A22B, Gemma-4-31B, Llama-4-Maverick, Llama-3.3-70B) – refine the picture. First, **failures are shared across models**. On easy quotes the models agree closely both with each other and with the dictionary ($\kappa \approx 0.98$–0.99). On hard quotes they still agree with each other ($\kappa = 0.61$) but not with the dictionary ($\kappa = 0.05$): they depart from it together. To rule out that the order affected their choices, we presented the candidate senses shuffled; the mean normalised rank was 0.42, close to the 0.5 expected under no position preference.

Second, **the errors are not merely over-splitting** (choosing a finer sense over a more general one). We take two senses to be "close" when the dictionary files them under the same major sense: by that measure only 51% of wrong picks stay within the gold's major sense (against 46.5% by chance), while the rest cross into a different one; higher-accuracy models, when wrong, stay nearer within the dictionary's own hierarchy ($r = -0.53$).

Format failures concentrate in the bottom-of-range Llama-3.2-3B (25.0%, 93% of all failures); Llama-3.3-70B is the only other model with a non-negligible rate (1.9%), and remaining models are ≈ 0%. These rates are low because Phase B uses plain output only; the heavy format failures were a Phase A phenomenon (§5.1).

6 Experiment 2: Concordance Lines with Human Gold

6.1 Annotation outcomes

The annotation protocol (§3.2) was applied to the 988 sampled concordance lines for 14 lemmas. They resolved to 837 analysable records: 151 were removed (149 by annotators — 128 as occurrences of another lemma, 22 as unassessable, with 1 record flagged on both), and 2 dropped for pipeline errors. The occurrences of the lemma *natio* 'nation' were overwhelmingly false positives, leaving only 9 clean records – retained as a Tier A lemma but too few to interpret. Gold resolution gave 583 consensus and 254 adjudicated records (30.3% adjudication), with 69 high-confidence disagreements flagged as genuine ambiguity.

6.2 Inter-annotator agreement

Inter-annotator agreement sets the human ceiling. Read per lemma, raw agreement is high for most lemmas (up to 1.00), with *libertas* the low outlier at 0.52 (Table 2); the pooled Cohen's κ is 0.838 (observed agreement 0.847, n=738), the ceiling against which §6.3 measures the LLMs. The macro-averaged κ is far lower, 0.378, but this is not the consequence of annotator unreliability: where a single sense covers nearly every record, two annotators agree by chance almost as often as they actually do, and κ credits only agreement beyond chance — so it collapses

TABLE 2 Experiment 2 annotation summary: sample size, number of source texts the lemma is attested in, first-level sense count, MFS, raw and Cohen's κ agreement, and adjudication rate

Lemma	n	Sources	L1 senses	MFS	K	Agree	% adjud
Tier A							
civitas	79	4	4	0.82	0.51	0.87	29.1
dignitas	79	4	7	0.73	0.56	0.81	25.3
imperator	79	4	3	0.95	0.20	0.73	96.2
populus	78	4	3	0.92	0.28	0.89	16.7
praefectus	78	4	2	0.59	0.76	0.88	21.8
comitia	73	3	2	1.00	0.00	0.98	15.1
episcopus	51	2	4	1.00	–	1.00	3.9
natio	9	3	4	0.89	0.00	0.86	33.3
Tier B							
princeps	58	4	9	0.97	0.17	0.86	24.1
libertas	56	4	9	0.59	0.23	0.52	53.6
officium	56	4	23	0.46	0.62	0.71	39.3
dux	51	4	3	0.63	0.91	0.95	19.6
bellum	50	4	3	0.96	0.41	0.90	12.0
lex	40	3	12	1.00	0.26	0.87	17.5

toward zero even under near-unanimous labelling. *Comitia* 'electoral assembly' is the extreme case: 98% agreement but MFS 1.00, so one sense covers everything and $\kappa \approx 0$ – the lemma is trivially easy, with nothing to disagree about. Its mirror image is *libertas*, the one genuinely contested lemma (agreement 0.52, κ 0.23, MFS 0.59), where the closely related political senses of *libertas* ('freedom', 'right to do something') resist separation even for expert annotators. We therefore report three metrics per-lemma – agreement, κ, and MFS share together – rather than κ in isolation, and read the human ceiling at the pooled level. Restricting to high-confidence records (both annotators ≥4) lifts the pooled κ only slightly, to 0.885, so the agreement is not merely an effect of annotators selecting the obvious sense when unsure; it holds on the records they were most confident about. Annotators also agree on what to discard: the cleanup carries pooled $\kappa = 0.987$ on match_status (agreement 0.997) and, for quote_quality, the same prevalence pattern as the senses ($\kappa = 0.045$ but raw agreement 0.886, so we report agreement).

6.3 LLM vs. Human comparison

Against the resolved human gold, the best configuration – **gemma-4-31B with structured-justification on Polish definition** – reaches 85.9% accuracy [83.5, 88.2] ($\kappa = 0.854$). All five models' best configs are structured-justification on Polish definition (Table 3). However, since the most-frequent-sense floor is now higher (micro 0.816), the best configuration's gain over MFS is only +4.3 pp, and only 3 of the 40 model–configuration runs beat the baseline even at the bottom of their confidence interval. Raw accuracy is therefore uninformative on its own; we report the margin over MFS and recall on the minority senses which guessing the dominant sense gets wrong. On pooled κ the best setup (0.854) sits at the human ceiling (0.838), but only when lemmas are pooled: per-lemma (macro) κ is dragged down by the same effect described in §6.2 – κ collapsing where one sense dominates – not by genuine model–human disagreement.

TABLE 3 Experiment 2 per-model best-config accuracy (all models in structured-justification on Polish definition)

Model	Acc %	95% CI	κ	Δ-MFS (pp)
gemma-4-31b-it	85.9	[83.5, 88.2]	0.854	+4.3
qwen3-235b-a22b-2507	82.0	[79.3, 84.6]	0.810	+0.4
llama-4-maverick	81.5	[78.9, 84.0]	0.805	−0.1
deepseek-chat-v3-0324	81.5	[78.9, 84.0]	0.804	−0.1
llama-3.3-70b-instruct	78.7	[76.0, 81.5]	0.776	−2.9

The gain over MFS is concentrated, not uniform (Figure 2): the best model beats the per-lemma MFS on only 6 of 14 lemmas, and those are the low-MFS, polysemous ones – *dux* (+33.3 pp), *praefectus* (+29.5), *officium* (+17.9, 23 senses), *libertas* (+14.3). Where one sense already dominates, MFS is effectively a ceiling and the model tends to over-split and lose ground: *princeps* (MFS 0.97 → 0.79, −18 pp), *civitas* (−12.7), *lex* (−12.5). This split – value added exactly where the inventory is genuinely polysemous, but forfeited where one sense is near-universal – is the central finding and feeds our discussion in §7.1. The definition-language effect is large on corpus text: pooled accuracy is 79.2% for Polish definitions against 71.3% for Latin, a 7.9 pp gap (paired McNemar, $p \approx 0$).

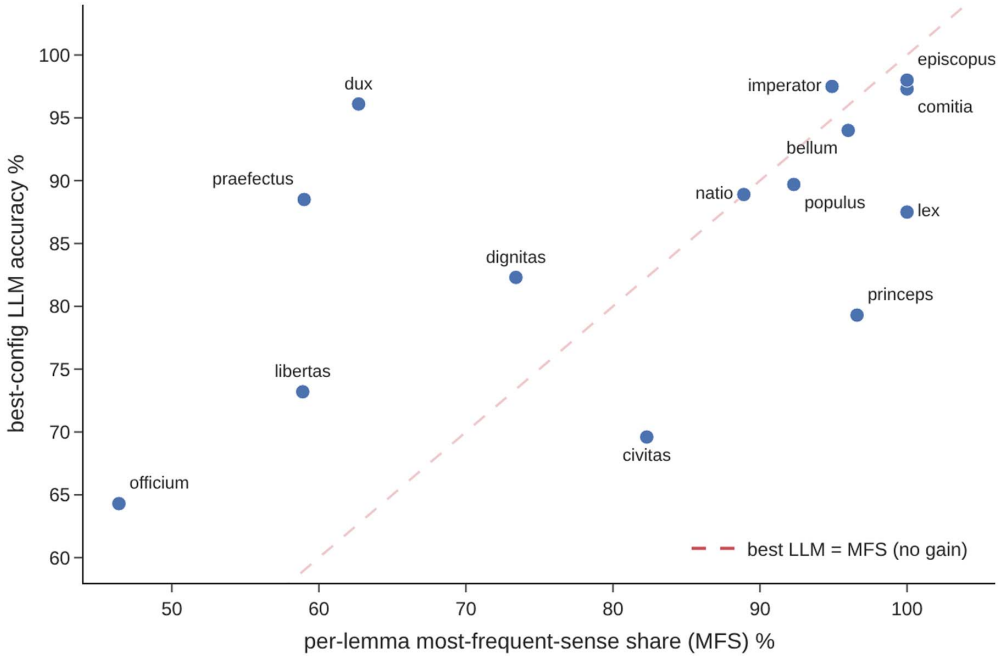


FIGURE 2 Per-lemma best-configuration accuracy against the per-lemma most-frequent-sense baseline (points above the MFS baseline are lemmas where the LLM beats MFS)

We also probed two further variables – providing a preceding sentence of context and source metadata (author, title, genre, period): both moved accuracy only modestly here (<1 pp),[3] but whether period and genre cues aid disambiguation is the diachronic question we leave to a future study.

[3] Context (KWIC+prev) added +0.79 pp on records with a preceding sentence (n = 732; $p \approx 5\times10^{-10}$); source metadata added +0.97 pp pooled, up to +2.1 pp on gemma.

Finally, model accuracy tracks human difficulty: the best configuration scores 89.5% on consensus records, 77.6% on adjudicated ones, and 69.6% on the genuine-ambiguity group with high-confidence human disagreement. It also rises with annotator confidence, from 42% (confidence 1) to 81% (confidence 5). **The models fail where the humans themselves split.**

6.4 Comparison with Experiment 1

Experiment 2 keeps the task framing of Experiment 1 but shifts both the domain — corpus lines rather than dictionary citations – and the gold provenance – human annotation rather than the DMLP's own assignments. Read raw, corpus accuracy is higher: model-pooled accuracy rises from 64.4% to 82.3% for the leader (gemma-4-31B, and on the four lemmas evaluated in both (*bellum, dux, lex, princeps*) it is 22–33 pp higher. This does not mean the task is easier: the Neo-Latin corpus task sits on a far higher most-frequent-sense floor (0.82 vs 0.49). The comparable quantity is the margin over MFS, and it collapses from +19.6 pp for the best Experiment 1 configuration to +4.3 pp here – the medieval inventory does not so much break down on Neo-Latin lines as leave the models little room above a dominant baseline. These two margins should therefore be read merely directionally, not as a strict comparison.

Two of Experiment 1's effects carry over and one inverts. **The Polish > Latin definition-language effect carries over and amplifies**, from ≈+3 pp to ≈+8 pp. Model ordering is broadly stable at the extremes (Gemma-4-31B and Qwen3-235B-A22B lead, Llama-3.3-70B trails in both; the middle reshuffles). But the template preference inverts: few-shot, marginally strongest on dictionary quotes, is weakest on corpus lines (71.2% vs justified's 78.1%), so the winning configuration shifts from Experiment 1's Qwen3-235B-A22B few-shot to Experiment 2's Gemma-4-31B justified – plausibly because dictionary-style few-shot demonstrations transfer poorly to unredacted concordance lines.

The 0-assignments, albeit scarce, provide another valuable insight. The medieval dictionary sense inventory frames Neo-Latin usage well, at least within the narrow domain of sociopolitical terms: almost every use matches a listed sense, and genuinely new senses are rare (13 cases, all in *civitas*). Each word appears mostly in a single sense, so the models' difficulty is not a dictionary that lacks the needed senses, but a corpus where one sense dominates and there is little left to disambiguate. Whether this holds in a more diverse corpus is left to future work.

7 Discussion

7.1 Implications for lexicographic practice

For a project compiling a resource like DiaLexPoL, two results bear directly on practice. The first is that **the tool need not be expensive**: self-hostable open-weight mid–sized and large models give the strongest results and beat the single commercial model at negligible cost. So an under-funded project – as historical-lexicography projects usually are – can fix and reproduce its pipeline, and even self-host both the model weights and the corpus data rather than depend on a flagship and store data in non-EU locations. The second is even more consequential: **LLM use earns its place on the polysemous lemmas** (low most-frequent-sense baseline), whereas on dominant-sense lemmas it adds little over simply assigning the most frequent sense and can lose ground by over-splitting.

Two configuration choices follow from our results. **Senses should be presented in the richer-resource language** – Polish here – which helps robustly, and more on corpus lines than on pre-edited quotes. Few-shot prompting helped on dictionary quotations but not on corpus

lines, where a structured-justification prompt did better, as noted in §6.4. **Model accuracy tracks human difficulty** (§6.3), but because we did not record the model's confidence in its own answers, we cannot use it to flag individual judgements.

Where the model fails, the remedy is often **not a larger model but external resources**. Some errors turn on information genuinely outside the Latin text — Polish institutional realia, interlinear glosses — pointing toward "retrieval-augmented lexicography", supplying the model with external reference material at inference time. Others turn on information the dictionary already carried but that the model under-used or that we had stripped from the input; these call not for retrieval but for preserving and enriching the dictionary glosses shown to the model with collocational and other cues the entry included for a reason, rather than stripping them.

Support for this comes from the effect of input quality. Experiment 1's lower accuracy partly reflects decontextualised, abbreviated fragments, whereas on full corpus sentences added context helped only marginally, so the payoff of restoring context may be register-dependent. It should be borne in mind, however, that the Experiment 2 corpus consists of discursive treatises, where a thought is usually developed fully within a single sentence; results may be worse on poetry or administrative prose, where a wider context window likely matters more.

The error analysis (§5.2) shows two further things. First, the failure is not pure over-splitting: only about half of the wrong answers stay within the original dictionary's major sense, while the other half land in a different one. So 'models over-split' is only part of the story: they also make coherent coarse errors. Second, a subset of disagreements is contestable gold, where the model's reading is defensible against a single unchecked dictionary decision in Experiment 1; so the measured error rate overstates the true one. The models' systematic disagreement with single-source dictionary assignments (§7.2) is itself a signal: editors can use it to flag legacy entries for re-examination. It also points to future work on gloss engineering: adapting and enriching the definitions shown to the model rather than using the dictionary's glosses unchanged, and even drafting new glosses for the database from scratch.

7.2 Validity and limitations

We did not test directly whether the models had seen our texts in pretraining, but the design limits how much that could matter. Experiment 2's gold is newly produced human annotation on the co-authors' own transcriptions of previously unedited works, so the gold labels are new and cannot have been seen in pretraining; and the consistency between the two experiments – stable model ordering and the robust Polish > Latin effect — argues that leakage is not what drives the results.

A related concern is the Experiment 1 gold itself: DMLP's definitions may be viewed as single-source assignments, with no inter-annotator check. On hard quotes the strongest models agree with one another far more than with that gold (κ 0.61 vs 0.05, §5.2); when models agree with each other but not with the dictionary, we cannot tell whether the models are wrong together or the dictionary's single, unchecked assignment is – there is no second judgement to settle it.

Experiment 2's multiply-annotated, adjudicated gold removes that ambiguity. More serious is the fact that the corpus is confined to a single domain – political, legal, and military prose – so the target lemmas, though frequent, are attested in only a narrow subset of their senses, concentrating usage on a few dominant ones. This raises the most-frequent-sense floor to 0.82 and inflates raw accuracy, which is why the margin over MFS, not raw accuracy, is the more instructive metric throughout.

8 Conclusion

We tested whether deployable LLMs can draft sense assignments for a diachronic Neo-Latin database built on a medieval-Latin inventory, in two experiments – reproducing the dictionary's own assignments (Experiment 1) and matching human annotation of corpus lines (Experiment 2). The answer is qualified. Self-hostable open-weight models give the best drafts at negligible cost, outperforming the commercial mid-tier, and the strongest corpus configuration matches human inter-annotator agreement when results are pooled across lemmas – though not on every lemma, where agreement statistics are distorted by dominant senses (§6.2). Gains over a most-frequent-sense baseline are modest and concentrate on genuinely polysemous lemmas – exactly where lexicographers most need help – while on dominant-sense lemmas the models over-split, marking where the lexicographer's contribution is required. We therefore treat LLM sense-assignment as a targeted aid in a post-editing framework, most useful on polysemous lemmas, not an autonomous step.

The inherited medieval inventory largely frames early-modern usage within this domain – evidence that the DMLP baseline can be built on, at least here – though flattening its structure produces artefacts which call for reconceiving that inventory. Candidate novel senses were too sparse in this single-domain corpus to test diachronic change directly.

Future work will broaden the corpus across domains (lowering the baseline and giving novelty detection more to work with) and add per-lemma fine-tuning, a Latin BERT baseline, and gloss engineering. It should also augment the model with knowledge resources – Old Polish dictionaries, realia, cross-references – rather than relying on larger models alone, and develop a validated error taxonomy that separates measurement artefacts, such as reduced context and debatable gold, from the genuine model errors that ought to inform lexicographic practice.

References

Bamman, D. & Burns, P. J. (2020). Latin BERT: A Contextual Language Model for Classical Philology. ArXiv:2009.10053. https://arxiv.org/abs/2009.10053

Ghinassi, I., Tedeschi, S., Marongiu, P., Navigli, R. & McGillivray, B. (2024). Language Pivoting from Parallel Corpora for Word Sense Disambiguation of Historical Languages: A Case Study on Latin. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation* (*LREC-COLING 2024*) (pp. 10073–10084). ELRA and ICCL. https://aclanthology.org/2024.lrec-main.880/

Ghizzota, E., Basile, P., Siciliani, L. & Semeraro, G. (2025). The Meaning of Beatus: Disambiguating Latin with Contemporary AI Models. In C. Bosco, E. Jezek, M. Polignano & M. Sanguinetti (Eds.), *Proceedings of the Eleventh Italian Conference on Computational Linguistics (CliC-it 2025)* (CEUR Workshop Proceedings, Vol. 4112, paper 45). CEUR-WS.org. https://ceur-ws.org/Vol-4112/45_main_long.pdf

Jakubíček, M. & Rundell, M. (2023). The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography? In M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček & S. Krek (Eds.), *Electronic Lexicography in the 21st Century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 Conference* (pp. 518–533). Lexical Computing CZ.

Lendvai, P. & Wick, C. (2022). Finetuning Latin BERT for Word Sense Disambiguation on the Thesaurus Linguae Latinae. In M. Zock, E. Chersoni, Y.-Y. Hsu & E. Santus (Eds.), *Proceedings of the Workshop on Cognitive Aspects of the Lexicon (pp. 37–41)*. Association for Computational Linguistics. https://doi.org/10.18653/v1/2022.cogalex-1.5

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, 11, 426. https://doi.org/10.1057/s41599-024-02889-7

Lugli, L. (2025). The Mangalam Dictionary of Buddhist Sanskrit: Automating Lexicographic Data with Generative LLMs. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda & A. Salgado (Eds.), *Electronic Lexicography in the 21st Century (eLex 2025): Intelligent Lexicography. Proceedings of the eLex 2025 Conference* (pp. 766–782). Lexical Computing CZ.

McGillivray, B. & Nowak, K. (2025). Tracing the semantic change of socio-political terms from Classical to early Medieval Latin with computational methods. In G. Galdi, S. Aerts & A. Papini (Eds.), *Varietate delectamur: Multifarious Approaches to Synchronic and Diachronic Variation in Latin (Selected Papers from the 14th International Colloquium on Late and Vulgar Latin*) (pp. 357–370). Brepols.

Meconi, D., Stirpe, S., Martelli, F., Lavalle, L. & Navigli, R. (2025). Do Large Language Models Understand Word Senses? In C. Christodoulopoulos, T. Chakraborty, C. Rose & V. Peng (Eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing* (pp. 33897–33916). Association for Computational Linguistics. https://doi.org/10.18653/v1/2025.emnlp-main.1720

Navigli, R. (2026). Is Word Sense Disambiguation Dead in the LLM Era? In S. Koenig, C. Jenkins & M. E. Taylor (Eds.), *Proceedings of the 40th AAAI Conference on Artificial Intelligence* (Vol. 40, No. 46) (pp. 39753–39762). https://doi.org/10.1609/aaai.v40i46.41331

Tiberius, C., Heylen, K., de Does, J., Vanroy, B., Vandeghinste, V. & van Doeselaar, J. (2024). LLMs and Evidence-Based Lexicography: Pilot Studies at INT. In S. Krek (Ed.), *Book of Abstracts of the Workshop Large Language Models and Lexicography* (pp. 44–48). Centre for Language Resources and Technologies, University of Ljubljana.

Widmann, T. (2025). A Pipeline for Automated Dictionary Creation with Optional Human Intervention. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda & A. Salgado (Eds.), *Electronic Lexicography in the 21st Century (eLex 2025): Intelligent Lexicography. Proceedings of the eLex 2025 Conference* (pp. 30–40). Lexical Computing CZ.

Contact information

Krzysztof Nowak
Institute of Polish Language (Polish Academy of Sciences)
krzysztof.nowak@ijppan.pl

Iwona Krawczyk
Institute of Polish Language (Polish Academy of Sciences)
iwona.krawczyk@ijp.pan.pl

Jagoda Marszałek
Institute of Polish Language (Polish Academy of Sciences)
jagoda.marszalek@ijp.pan.pl

Acknowledgments

We gratefully acknowledge the Polish high-performance computing infrastructure PLGrid (HPC Center: ACK Cyfronet AGH) for providing computer facilities and support within computational grant no. PLG/2024/017402 for the preliminary testing phase of the research presented in this paper.

Author contributions

Krzysztof Nowak conceived and designed the study, designed and implemented the experiments, conducted all analyses, prepared the datasets for correction and annotation, adjudicated unresolved annotation disagreements, and wrote the first draft of the paper; he is primarily responsible for all sections of the manuscript. Iwona Krawczyk (IK) and

Jagoda Marszałek (JM) corrected the structure of and edited DMLP definitions; edited and quality-checked the dictionary quotations, including their sense assignment; and edited the Neo-Latin source material. Both IK and JM annotated corpus lines with quality labels and sense assignments, contributing especially to Sections 3.2 and 6.1–6.3. All authors reviewed, corrected, and approved the final manuscript.

Saghar Sharifi & Abolfazl Alamdar

TOWARDS A HANDBOOK OF PERSIAN LEXICOGRAPHY

A Lexicographical-Functions Analysis of 20th–21st Century Dictionaries

Abstract This article examines four major Persian dictionaries—*Dehkhoda* (1940), *Mo'in* (1972), *Sokhan* (2003), and *Aryanpur* (2009)—through Tarp's Function Theory of Lexicography (2008). Using a descriptive-analytical mixed-methods design, the study evaluates the dictionaries in terms of four lexicographical functions: text reception, text production, translation, and knowledge acquisition. The analysis is based on 12 lexical and grammatical categories, with three sampled lexical items from each category, yielding 144 dictionary-entry observations across the dataset. The findings reveal clear functional asymmetries. *Dehkhoda* performs most strongly in knowledge acquisition and text reception; *Mo'in* preserves much of this interpretive orientation in a more compact form; *Sokhan* demonstrates the strongest monolingual performance in text reception and contemporary communicative usability; and *Aryanpur*, as a bilingual dictionary, is highly effective for translation but very weak in knowledge acquisition. Overall, the study shows that modern Persian lexicography has moved gradually toward greater user orientation, though without achieving balanced multifunctionality. It concludes by arguing for stronger dialogue between Persian lexicography and international lexicographical scholarship and recommends the compilation of a *Handbook of Persian Lexicography*.

Keywords Persian Dictionaries; Function Theory; Handbook of Lexicography; Lexicographical Functions

1 Introduction

Lexicography, as both a practical and theoretical discipline, has increasingly moved toward user-oriented and function-sensitive approaches in recent decades. In contemporary metalexicography, dictionaries are no longer viewed merely as passive repositories of lexical items and meanings; rather, they are understood as reference tools designed to meet specific user needs in particular consultation situations. Within this broader development, Tarp's Function Theory of Lexicography (2008) has emerged as one of the most influential theoretical models for understanding how dictionaries serve communicative and cognitive purposes.

Despite these developments, Persian lexicography remains relatively underrepresented in international theoretical discussions of dictionary making. Although Persian possesses a long and intellectually rich lexicographical tradition, extending from early lexical traditions before Islam to major post-Islamic dictionary works, its achievements are only marginally reflected in dominant handbooks and reference works in lexicography. Early evidence of this tradition can be associated with pre-Islamic lexical materials such as *Oim* (pre-Islamic period), while after Islam one of the foundational monuments of Persian lexicography is Asadi Tusi's *Logāt-e Fors* (c. 1073 CE). Major publications such as *The Bloomsbury Handbook of Lexicography* (2nd ed., 2022), *The Cambridge Handbook of the Dictionary* (2024), *The Oxford Handbook of Lexicography* (2016), and *The Routledge Handbook of Lexicography* (2020) have helped shape the global landscape of lexicographical scholarship, yet Persian dictionary traditions receive little sustained analytical attention within them. This absence is noteworthy not because Persian lexicography lacks substance, but because it suggests a broader imbalance in the global representation of lexicographical traditions.

At the same time, Persian lexicography itself has often developed without sustained engagement with modern meta-lexicographical theory. Many important Persian dictionaries have been compiled in response to practical, philological, literary, pedagogical, or translational needs, but without being systematically described in terms of dictionary functions, user situations, or consultation scenarios. This creates a productive field of inquiry: if Persian dictionaries are examined through a contemporary theoretical framework, what kinds of functional strengths and limitations become visible? To what extent do major Persian dictionaries respond to different user needs? And how might such an analysis contribute both to Persian lexicography and to lexicographical theory more broadly?

The present article addresses these questions by analyzing four major Persian dictionaries—*Dehkhoda* (1940), *Mo'in* (1972), *Sokhan* (2003), and *Aryanpur* (2009)—through the lens of Tarp's Function Theory of Lexicography (2008). These dictionaries were selected because they represent distinct moments and orientations within the development of modern Persian lexicography: historical-documentary, general monolingual, contemporary user-oriented, and bilingual-translational. Their comparison makes it possible to identify broader tendencies in Persian dictionary-making over the last century.

The study focuses on four major lexicographical functions emphasized in Tarp's framework (2008): text reception, text production, translation, and knowledge acquisition. These functions provide a structured basis for assessing how dictionaries assist users in understanding texts, producing texts, transferring meaning across languages, and acquiring broader lexical or encyclopedic knowledge. By applying these categories to the selected dictionaries, the article seeks not only to compare the works themselves but also to assess the extent to which Persian lexicography has moved toward a function-based and user-oriented model.

The significance of the study is twofold. First, it contributes to the still-limited body of meta-lexicographical research on Persian dictionaries. Second, it opens a broader discussion about the need for stronger dialogue between Persian lexicography and global lexicographical scholarship. In this sense, the article is not only an evaluative study of four dictionaries, but also an argument for the fuller theoretical visibility of Persian lexicography.

2 Function Theory of Lexicography

The present study is grounded in Tarp's Function Theory of Lexicography, one of the most influential approaches in contemporary meta-lexicography. Within this framework, a dictionary is defined not merely by the quantity of lexical data it contains, nor by its formal structure alone, but primarily by its capacity to fulfill specific functions in relation to users' needs. Dictionaries are thus conceived as tools designed to assist users in concrete consultation situations. Their success depends on how appropriately their structure and contents respond to these situations.

A central principle of the theory is that lexicographical works should be evaluated in relation to their intended use rather than against an abstract or purely formal ideal. The value of a dictionary therefore lies in its ability to provide relevant information efficiently to a particular type of user who encounters a particular type of problem. Such a perspective shifts attention from the dictionary as an autonomous textual object to the dictionary as a functionally motivated reference instrument.

Tarp's theory emerged partly in response to earlier tendencies in lexicography that focused heavily on linguistic description while paying insufficient attention to actual user needs. In contrast, function theory places users at the center of dictionary design. It asks what kinds

of problems users face, what kinds of information can solve those problems, and how that information should be presented in order to maximize accessibility and usefulness. In this respect, the theory is especially relevant for evaluating dictionaries across traditions, since it provides a principled comparative framework rather than one tied to a single language or lexicographical culture.

2.1 Lexicographical Functions

Within Tarp's model (2008), lexicographical functions are generally divided into two broad types: communicative functions and cognitive functions. Communicative functions relate to situations in which users need assistance with communication, such as reading, writing, or translating a text. Cognitive functions, by contrast, relate to situations in which users seek knowledge for its own sake, including lexical, cultural, encyclopedic, or subject-related information.

The present article focuses on four major functions that are particularly relevant to the selected Persian dictionaries:

1. Text reception refers to the dictionary's ability to help users understand words, expressions, and structures encountered in spoken or written texts. This includes support through definitions, sense differentiation, contextual guidance, stylistic or register labels, and semantic clarification.

2. Text production refers to the dictionary's usefulness in helping users produce appropriate and acceptable language. This function depends on the provision of information such as grammatical behavior, collocational patterns, phraseology, register or stylistic guidance, and context of use.

3. Translation concerns the dictionary's role in enabling users to transfer lexical meaning between languages. This function is especially central in bilingual dictionaries and involves the provision of equivalents, semantic distinctions, contextual differentiation, and translational guidance.

4. Knowledge acquisition refers to the dictionary's capacity to provide broader lexical, cultural, historical, or encyclopedic knowledge. This function extends beyond immediate communicative needs and includes information such as etymology, literary citation, cultural explanation, and domain-specific or historical background.

These four functions provide an analytically useful framework for comparing different dictionary types. Not every dictionary is expected to fulfill all functions equally, and some functions may be more central than others depending on the nature of the work. Nevertheless, the distribution of functional strengths and weaknesses reveals much about the lexicographical priorities underlying a given dictionary.

2.2 Relevance to Persian Lexicography

The application of Tarp's Function Theory (2008) to Persian lexicography is especially productive for several reasons. In the first place, Persian dictionaries have historically been compiled for diverse purposes: literary interpretation, preservation of classical language, lexical explanation, education, bilingual mediation, and cultural documentation. These purposes align closely with the functional perspective proposed by Tarp, even when they were not originally

theorized in those terms. A function-based analysis, therefore, makes it possible to reinterpret Persian dictionary traditions in a more systematic and comparative way.

In addition, Persian lexicography presents a particularly interesting case because it combines historical depth with functional diversity. Monumental works such as *Dehkhoda* (1940) were not designed primarily as user-oriented modern reference tools, yet they fulfill important cognitive and interpretive functions with exceptional richness. More recent works such as *Sokhan* (2003) and *Aryanpur* (2009), by contrast, show stronger tendencies toward communicative usability and user-oriented structuring. This range allows the present study to trace not only functional variation but also any potential shifts in lexicographical priorities over time.

More importantly, the Persian case also highlights certain limitations in the universal application of existing theory. If a dictionary such as *Dehkhoda* (1940) performs exceptionally well in documentary, historical, and cultural terms, then the Persian tradition may also encourage further reflection on how fully current categories of lexicographical function capture the role of dictionaries in preserving and transmitting linguistic-cultural heritage. In this sense, Persian lexicography is not merely a passive object of theoretical application; it also has the potential to enrich and challenge lexicographical theory itself.

For these reasons, Tarp's framework (2008) offers a valuable starting point for the present investigation. It allows the selected dictionaries to be examined not only as lexical compilations, but as functionally structured tools whose strengths, limitations, and historical orientations can be described in relation to actual or implied user needs.

3 Methodology

3.1 Corpus

This study adopts a descriptive-analytical mixed-methods design to examine the functional performance of four major Persian dictionaries within the framework of Tarp's theory (2008). The sample consists of the following dictionaries:

1. *Loghatnāmeh-ye Dehkhodā* (1940), here called *Dehkhoda*;
2. *Farhang-e Fārsi* (1972), here called *Mo'in*;
3. *Farhang-e Bozorg-e Sokhan* (2003), here called *Sokhan*;
4. *Farhang-e Pishrow-e Aryanpur* (2009), here called *Aryanpur*.

These dictionaries were selected through purposive sampling based on their prominence, accessibility, continued scholarly or practical use, and representativeness of different lexicographical orientations within modern Persian. *Dehkhoda* represents the historical-documentary and encyclopedic tradition; *Mo'in* represents the general monolingual tradition in a more concise and systematized form; *Sokhan* represents a more contemporary and user-oriented monolingual model; and *Aryanpur* represents the bilingual and translation-oriented tradition.

This selection makes it possible to compare not only individual works, but also broader tendencies within Persian lexicography over the past century.

3.2 Sampling and Data Extraction

To evaluate the dictionaries in a balanced and function-sensitive way, the study selected 12 representative lexical and grammatical categories, with three lexical items sampled from each category. This yielded a total of 36 sampled lexical units for each dictionary and 144 dictionary-entry observations across the full dataset.

The sampling framework is category-based rather than information-type-based. However, the selected categories were intentionally chosen because they are likely to elicit different kinds of lexicographical information relevant to Tarp's (2008) four functions. In other words, although the classification is organized around lexical and grammatical categories, its analytical purpose is function-sensitive.

The 12 categories and their sampled items are as follows:[1]

1. **Polysemous nouns:** *ʃɪr* 'lion/milk', *sær* 'head', *rah* 'way; path'[2]
2. **Simple verbs:** xændɪdæn 'to laugh', tærsɪdæn 'to fear', mɔrdæn 'to die'
3. **Light verbs:** zædæn 'to strike; to hit', kærdæn 'to do; to make', gereftæn 'to take; to seize'
4. **Compound verbs:** hærf zædæn 'to speak', zæmɪn xɔrdæn 'to fall down', *gʊʃ dadæn* 'to listen'
5. **Adjectives:** rɔʃæn 'bright; clear', sængɪn 'heavy', tɔnd 'sharp; severe; fast'
6. **Adverbs:** nɑgehɑn 'suddenly', kɑmelæn 'completely', hærgez 'never'
7. **Prepositions:** æz 'from', bɑ 'with', bæraje 'for'
8. **Idiomatic expressions:** æz kʊre dær ræftæn 'to lose one's temper', dæst be sær kærdæn 'to get rid of someone', pɑdærmɪjɑnɪ kærdæn 'to mediate'
9. **Scientific and technical terms:** vɪrʊs 'virus', tæværrɔm 'inflation', dɪ en ej 'DNA'
10. **Synonyms and near-synonyms:** dælɪr 'brave', ʃɔd͡ʒɑ 'courageous', bɪkælleh 'rash; foolhardy'
11. **Labelled entries:** xɔvɑlɪgær 'courtier/flatterer' (archaic or stylistically marked), xæfæn 'awesome/stifling' (colloquial or slang, depending on context), ʃæræfjɑb ʃɔdæn 'to be honored by attendance' (honorific/formal)
12. **Culture-specific items:** hæftsɪn 'the traditional Nowruz table with seven symbolic items', tæ'ɑrɔf 'ritual politeness/sociolinguistic etiquette', rʊdærbɑjestɪ 'social hesitation arising from politeness or obligation'

The selection of the above categories was intended to ensure diversity across semantic, grammatical, phraseological, stylistic, and cultural dimensions of the lexicon. This diversity was necessary because a function-based evaluation requires examining how dictionaries respond to different types of lexical complexity and user need. For example, polysemous nouns and idioms test sense differentiation and interpretive richness; verbs and prepositions probe grammatical

[1] For full entry-level documentation of the 144 dictionary observations underlying the present analysis, readers may consult the online supplementary dataset available at: https://drive.google.com/file/d/1uDiPx2ePie3g0ulptYb3LwzODqJ19BzV/view?usp=drive_link

[2] The phonetic transcriptions of the Persian lexical items were provided using the IPA symbols available at: https://www.ipachart.com

and combinatory guidance; scientific terms assess domain-specific clarity; and culture-specific items test the dictionary's ability to mediate culturally embedded meaning.

3.3 Evaluation Instrument and Criteria

Based on Tarp's functional framework (2008), each entry was examined with regard to the type and quality of information it provided in relation to the four major functions of text reception, text production, translation, and knowledge acquisition, as follows:

- structural and grammatical information, including part of speech, morphological indication, syntactic behavior, and combinatory guidance;
- sense differentiation and semantic presentation, including clarity of definition, distinction between meanings, treatment of polysemy, and semantic precision;
- collocational and pragmatic information, including phraseological patterns, usage contexts, register, stylistic labeling, and appropriateness of use;
- translational information, including the provision of equivalents, differentiation between possible equivalents, and contextual adequacy in bilingual consultation;
- extralinguistic and encyclopedic information, including etymology, historical citation, cultural explanation, domain-specific knowledge, and broader conceptual elaboration.

Using these criteria, each dictionary was assigned a score for each of the four functions. The scoring was carried out on a 0–5 scale, where higher scores indicate stronger functional performance.

The analysis proceeded in two stages. In the first stage, each of the 36 selected lexical items was examined in each of the four dictionaries. For every entry, observations were recorded regarding the presence, absence, and quality of information relevant to each of Tarp's (2008) four functions. In the second stage, these entry-level observations were aggregated by dictionary and function in order to arrive at comparative scores.

The final scores, therefore, represent average functional performance across the 12 lexical categories. This procedure made it possible to identify not only isolated strengths or weaknesses in individual entries, but also broader lexicographical tendencies in each dictionary.

For interpretive clarity, the numerical scores were associated with qualitative labels ranging from very weak to very strong. The primary analytical scale of the study is the original 0–5 functional rating scale. In addition, for the sole purpose of visual comparison in Section 4, these scores were converted into percentages on a 0–100 scale. This conversion was not methodologically necessary for the analysis itself, but it improves readability and facilitates cross-dictionary comparison in the bar chart by placing all results on an immediately interpretable visual scale.

It should also be emphasized that the function of translation is not directly applicable to the three monolingual dictionaries (*Dehkhoda*, *Mo'in*, and *Sokhan*) in the strict interlingual sense. For this reason, it is marked as N/A rather than assigned a numerical score. By contrast, *Aryanpur* received a score of 0.000 in knowledge acquisition because this function was considered evaluable in principle, yet the dictionary offered virtually no substantial data in that domain.

The scoring scale should be understood as an interpretive functional rating scale rather than a strictly interval-based mathematical scale. In particular, a score of 0.0 does not simply indicate "a very low degree" of functional support; it indicates the absence or near-absence of meaningful relevant information for the function under evaluation. By contrast, a score such as 2.0 indicates that some relevant information is present, though in a limited, incomplete, or only weakly usable form. Intermediate values such as 2.5 were used when performance fell between the adjacent qualitative categories.

4 Data Analysis and Findings

4.1 Quantitative Analytical Procedure and Evaluation Scale

The present study evaluates four major Persian dictionaries through the lens of Tarp's Function Theory of Lexicography (2008). Within this framework, dictionaries are assessed in relation to four major lexicographical functions: text reception, text production, translation, and knowledge acquisition.

To operationalize this model, the analysis was conducted on the basis of the 12 representative lexical and grammatical categories introduced in the methodology section—namely polysemous nouns, simple verbs, light verbs, compound verbs, adjectives, adverbs, prepositions, idiomatic expressions, scientific and technical terms, synonyms and near-synonyms, labelled entries, and culture-specific items. For each category, the corresponding entries were examined in the four selected dictionaries: *Dehkhoda*, *Mo'in*, *Sokhan*, and *Aryanpur*. The evaluation focused on the extent to which each dictionary provides information relevant to the fulfillment of the four functions, including definitional clarity, sense differentiation, grammatical guidance, collocational and pragmatic information, translational equivalents where relevant, and encyclopedic or extralinguistic detail.

Each function was assigned a score based on the degree to which the dictionary effectively addressed the relevant user need. These scores were then averaged across the 12 categories in order to obtain a final score for each dictionary in each functional domain. The primary evaluation was carried out on a 0–5 scale, and these scores were subsequently converted into percentages for visual comparison in the bar chart. Thus, for instance, a score of 4.625 corresponds to 92.5%.

It should be noted that, for the three monolingual dictionaries (*Dehkhoda*, *Mo'in*, and *Sokhan*), the function of translation is not directly applicable (N/A) in the strict interlingual sense, since these dictionaries do not provide target-language equivalents. This non-applicability is reflected consistently in both the tables and the figure. By contrast, *Aryanpur* received a score of 0.000 in knowledge acquisition. Although this function was considered evaluable in principle, the dictionary offered virtually no meaningful data in that domain. Table 1 illustrates the quantitative scoring based on Tarp's Functional framework (2008).

TABLE 1 Quantitative scoring scale based on Tarp's Functional Framework

Score	Qualitative level	Description
0.0	Very weak	No meaningful relevant information provided, or only negligible information
2	Weak	Limited information is present, but it is incomplete or of little practical use
2.5	Weak to moderate	Performance falls between weak and moderate
3	Moderate	Basic information is provided, but it lacks depth or adequacy
3.5	Moderate to strong	Performance is acceptable and exceeds the average level
4	Strong	Information is comprehensive, systematic, and functionally useful
4.5	Strong to very strong	Performance is excellent and close to fully satisfactory
5	Very strong	Information is exceptional, complete, and exemplary

It should also be noted that the percentage conversion used in the figure serves a purely presentational purpose. The analytical judgments themselves were made on the 0–5 functional scale, while the 0–100 conversion was introduced only to enhance visual clarity for readers. Likewise, the scoring labels should be read as interpretive categories rather than as claims of exact interval equivalence between all adjacent values.

4.2 Results

4.2.1 Dehkhoda

The functional profile of *Dehkhoda* as a historical-encyclopedic dictionary can be observed in Table 2.

TABLE 2 *Dehkhoda* dictionary in terms of Tarp's Lexicographical Functions

Function	Final score	Qualitative level
Text reception	4.625	Strong to very strong
Text production	1.833	Weak
Translation	N/A	Not applicable
Knowledge acquisition	4.791	Very strong

Dehkhoda, as a historical and encyclopedic lexicographical monument, offers the strongest realization of the knowledge-acquisition function among the dictionaries examined in this study. Its extensive treatment of etymology, historical usage, literary citation, and cultural or encyclopedic information makes it an exceptionally rich resource for users seeking lexical, textual, and civilizational knowledge. In this regard, its performance approaches the upper limit of Tarp's cognitive dimension.

Its score in text reception is also very high. The dictionary's careful differentiation of senses, broad semantic treatment, and abundant citation from classical and literary sources enable users to interpret difficult, archaic, and stylistically marked texts with a high degree of confidence. For philological and interpretive reading, *Dehkhoda* remains indispensable.

By contrast, its performance in text production is distinctly weak. Although it is highly informative in historical and semantic terms, it offers relatively little practical support for modern language production. Information about syntactic behavior, collocational preference, productive phraseology, and contextual appropriateness is limited or unsystematic, and examples are not generally designed to guide contemporary writing. This imbalance reflects the dictionary's foundational orientation: it was conceived primarily as a documentary and interpretive repository rather than as a user-oriented tool for active language production. In functional terms, *Dehkhoda* excels as a dictionary for reading and knowledge acquisition but remains weak as an aid to writing.

4.2.2 Mo'in

Based on data gathered from *Mo'in*, this dictionary appears more structured, though it remains primarily interpretive. Table 3 provides evidence for this claim.

TABLE 3 *Mo'in* dictionary in terms of Tarp's Lexicographical Functions

Function	Final score	Qualitative level
Text reception	3.791	Strong
Text production	2.166	Weak
Translation	N/A	Not applicable
Knowledge acquisition	3.875	Strong

Mo'in may be regarded as a more compact and systematized continuation of the earlier Persian lexicographical tradition represented by *Dehkhoda*. Like *Dehkhoda*, it is primarily oriented toward text reception and knowledge acquisition, though in a more concise and structurally accessible format. Its clearer definitions, more regularized entry structure, and inclusion of basic grammatical and orthographic information make it more usable for the general reader than its monumental predecessor.

Its score in knowledge acquisition is likewise relatively high. Although it does not match *Dehkhoda* in historical breadth or documentary depth, it still preserves substantial lexical, etymological, and cultural information. This enables it to function not merely as a decoding tool, but also as a source of broader lexical understanding.

Nevertheless, *Mo'in* remains weak in text production. Despite its improved structural clarity, it still lacks many of the features associated with a genuinely productive dictionary. Information about syntactic patterns, collocational behavior, argument structure, and context-sensitive lexical choice is relatively limited in this dictionary. Usage examples are not sufficiently rich or systematically designed to support writing decisions, and stylistic or pragmatic guidance remains underdeveloped. As a result, *Mo'in* may be regarded as a useful interpretive and semi-cognitive dictionary, but not as a strongly user-oriented tool for productive language use.

4.2.3 Sokhan

In *Sokhan*, this functional turn toward contemporary language use is reflected in its emphasis on describing how words and expressions are actually used in modern contexts rather than prescribing how they should be used. Table 4 illustrates how this approach shapes entries in *Sokhan*.

TABLE 4 *Sokhan* dictionary in terms of Tarp's Lexicographical Functions

Function	Final score	Qualitative level
Text reception	4.750	Strong to very strong
Text production	3.750	Moderate to strong
Translation	N/A	Not applicable
Knowledge acquisition	2.083	Weak

Sokhan represents a significant shift in Persian lexicography toward more modern, user-oriented, and communicatively effective dictionary design. More clearly than the other monolingual dictionaries included in this study, it appears to have been compiled with the needs of contemporary users in mind. This orientation is reflected in its strong performance in both text reception and text production.

Its very high score in text reception results from its clear definitions, relatively precise sense differentiation, and practical arrangement of lexical information. Particularly important is its stronger inclusion of contextual guidance, collocational information, and usage-based evidence, all of which facilitate the interpretation of contemporary Persian in actual communicative settings.

The dictionary also performs comparatively well in text production, significantly outperforming both *Dehkhoda* and *Mo'in* in this respect. This suggests an important functional transition in Persian lexicography from passive lexical documentation toward active communicative assistance. By offering more relevant information on lexical choice, combinatory behavior, and context of use, *Sokhan* provides users with more effective support for producing acceptable idiomatic Persian.

This shift, however, is accompanied by a relative weakening of the knowledge-acquisition function. In comparison with *Dehkhoda* and *Mo'in*, *Sokhan* devotes considerably less attention to etymological detail, lexical history, and encyclopedic elaboration. This appears to be a strategic lexicographical choice rather than a mere deficiency: the dictionary privileges synchronic usability and communicative efficiency over documentary depth. Consequently, it emerges as the strongest monolingual dictionary in the sample for contemporary reading and writing, but not for historical or encyclopedic inquiry.

4.2.4 Aryanpur

In *Aryanpur*, the role of the dictionary as a specialized bilingual tool for translation is reflected in its emphasis on direct equivalents and translation-oriented guidance. Table 5 illustrates how this focus shapes entry information and bilingual correspondences.

TABLE 5 *Aryanpur* dictionary in terms of Tarp's Lexicographical Functions

Function	Final score	Qualitative level
Text reception	2.000	Weak
Text production	3.790	Moderate to strong
Translation	4.833	Very strong
Knowledge acquisition	0.000	Very weak

As the only bilingual dictionary included in the study, *Aryanpur* aligns most directly with the translation function in Tarp's framework. Its exceptionally high score in this domain confirms that it is primarily designed to support interlingual consultation. The provision of multiple English equivalents, together with relatively careful differentiation of meaning across contexts, makes it a highly effective tool for bilingual lexical choice and translational decision-making.

The dictionary also performs reasonably well in text production, particularly where productive needs overlap with bilingual consultation. Through its treatment of equivalents, phraseological information, and occasional contextual cues, it can support users engaged in writing or translating between Persian and English. In this respect, it offers broader practical value than a purely equivalence-driven bilingual list.

Its limitations, however, are equally clear. Its score in text reception is low, largely because Persian-side explanation is often too brief to support deeper semantic interpretation. Even more notably, its score in knowledge acquisition is 0.000, which indicates not merely weakness but a near-total absence of cognitively enriching information. The dictionary provides virtually no etymological, historical, encyclopedic, or culturally expansive treatment of entries. It is therefore best understood as a highly specialized translational instrument rather than a broader educational or interpretive lexical resource.

4.3 Cross-Dictionary Comparison

For a clearer comparative overview, in Figure 1, Tarp's lexicographical functions (2008) have been represented side by side across the four dictionaries, with scores converted from the original 0–5 scale into percentages. This conversion facilitates visual comparison by placing the results on a common scale. However, the translation function is not directly applicable to the three monolingual dictionaries (*Dehkhoda*, *Mo'in*, and *Sokhan*), since they do not provide interlingual equivalents. Accordingly, this function is marked as N/A for these dictionaries in both the tables and the figure, rather than being represented numerically as zero.

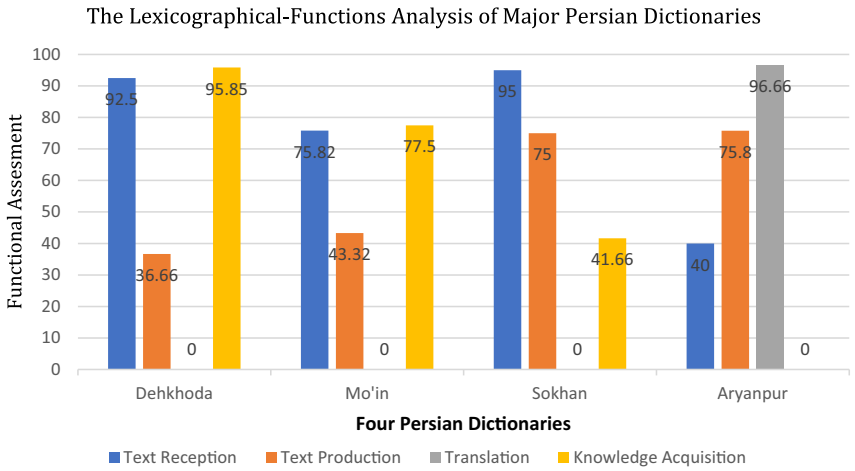


FIGURE 1 Comparative percentage scores of *Dehkhoda*, *Mo'in*, *Sokhan*, and *Aryanpur* across Tarp's four Lexicographical Functions[3]

[3] Percentages in the figure are derived from the original 0–5 evaluation scores. For the three monolingual dictionaries, the translation function is marked as not applicable (N/A) in both the tables and the figure, since these dictionaries do not directly provide interlingual equivalents. By contrast, Aryanpur's 0% in knowledge acquisition represents an evaluated function for which no meaningful data were found.

The comparative results reveal several broader patterns.

First, the findings point to a clear functional polarization within Persian lexicography. On one side stand the historically and cognitively oriented dictionaries—especially *Dehkhoda* and, to a lesser extent, *Mo'in*—which perform strongly in knowledge acquisition and text reception but weakly in text production. On the other side stand the more communicatively oriented dictionaries—*Sokhan* and *Aryanpur*—which are optimized for production and, in *Aryanpur*'s case, translation, but provide relatively limited historical or encyclopedic knowledge. None of the dictionaries examined in this study offers a fully balanced realization of all four functions.

Second, the results reveal a persistent gap between lexicographical theory and Persian lexicographical practice. In Tarp's framework, an ideal dictionary is one whose structure and data are explicitly designed in relation to user needs and consultation situations. The dictionaries examined here, however, remain largely single-function or dual-function works rather than fully integrated multifunctional tools. Their structures reflect the dominant priorities of their compilers—historical documentation, lexical explanation, or translational utility—more than a comprehensive and explicitly articulated functional design.

At the same time, the data also indicate a gradual movement toward greater functional awareness. The comparatively strong performance of *Sokhan* in text production suggests that Persian lexicography has begun to move beyond the traditional view of the dictionary as a passive lexical archive. By incorporating more collocational, contextual, and usage-based information, *Sokhan* narrows the gap between theoretical expectations and lexicographical practice and illustrates the growing relevance of user-oriented design principles in modern Persian dictionary making.

A related finding concerns the tension between specialization and comprehensiveness. *Aryanpur*'s score of 4.833 in translation and 0.000 in knowledge acquisition sharply demonstrates that a dictionary may be highly effective for one function while remaining intentionally limited in others. From this perspective, the issue is not specialization itself, but the absence of explicit functional framing and user guidance. Lexicographical effectiveness increases when a dictionary's intended function is clearly defined and systematically reflected in both its macrostructure and microstructure.

Finally, the results identify *Sokhan* as an important turning point in Persian lexicography. By partially addressing the longstanding weakness of monolingual Persian dictionaries in text production, it emerges as the most practically useful monolingual dictionary in the sample for contemporary reading and writing. Its performance lends empirical support to the increasing importance of productive and communicative functions in current lexicographical thought.

At the same time, *Aryanpur* serves as a revealing example of functional narrowing. Its excellence in translation is achieved at the expense of interpretive depth and cognitive enrichment. This makes it highly valuable as a technical bilingual tool, but less suitable as a general educational or developmental lexical resource.

A further point emerging from the comparison is that the lexicographical functions do not operate in complete isolation from one another. Certain functions appear to correlate positively in some dictionaries: for example, the strong text-reception performance of *Dehkhoda* and *Mo'in* is closely linked to their relative strength in knowledge acquisition, since fuller definitions, historical detail, and documentary citation support both interpretation and learning. In *Sokhan*, stronger text production appears to be associated with richer collocational and usage-based information, which also contributes secondarily to reception. In *Aryanpur*, translation

and text production partially overlap where bilingual equivalents, phraseological combinations, and contextual examples support not only transfer between languages but also the production of acceptable English target text. At the same time, the data also reveal functional trade-offs: dictionaries that invest heavily in documentary, historical, and encyclopedic depth tend to be less efficient in productive support, whereas dictionaries optimized for translation or communicative usability tend to reduce cognitive and heritage-oriented content. The relationship among functions, therefore, is neither one of full independence nor one of simple complementarity, but rather a dynamic pattern of overlap, reinforcement, and trade-off.

5 Conclusion

This study evaluated four major Persian dictionaries through Tarp's Function Theory of Lexicography (2008) in order to examine the relationship between modern lexicographical theory and Persian dictionary practice. The findings reveal a clear functional asymmetry in modern Persian lexicography. While Persian dictionaries have often been rich in lexical, historical, and cultural content, they have not always been designed with explicit attention to user types, consultation situations, and lexicographical functions in the modern theoretical sense.

At the same time, the results indicate a gradual movement toward greater user orientation. *Dehkhoda* is strongest in knowledge acquisition and text reception; *Mo'in* retains much of this interpretive and cognitive orientation in a more compact form;

Sakhan represents the strongest monolingual performance in text reception and the clearest shift toward communicative efficiency, while also offering stronger support for text production than *Dehkhoda* and *Mo'in*; and *Aryanpur* stands out as a highly effective bilingual tool for translation, though with very limited cognitive scope. These findings suggest that Persian lexicography is best understood not as a uniform tradition, but as a functionally differentiated field shaped by historical layering, specialization, and changing user priorities.

The study also shows that lexicographical functions may reinforce one another in some cases, while in others they involve clear trade-offs. Dictionaries rich in semantic explanation, citation, and documentary depth tend to support both text reception and knowledge acquisition, whereas dictionaries optimized for translation or communicative efficiency often provide less historical and encyclopedic information. Functional specialization, therefore, appears to be structurally embedded in the dictionaries examined here.

Overall, the study highlights the need for more explicitly function-based dictionary design in Persian lexicography and for stronger dialogue between Persian and international lexicographical scholarship. On this basis, it finally recommends the compilation of a *Handbook of Persian Lexicography* as a timely and necessary step toward documenting the field systematically and connecting its traditions with contemporary theoretical approaches.

References

Anvari, H. (Ed.) (2003). *Farhang-e bozorg-e Sokhan [The Great Sokhan Dictionary]*. Sokhan Publications.

Aryanpur Kashani, M. & Assi, M. (2009). *The Aryanpur progressive Persian-English Dictionary*. Jahan R ayaneh Computer World Co.

Asadi Tusi, A. (1965). *Loḡat-e Fors* (M. Dabirsiyaqi, Ed.). Tehran University Press.

Dehkhoda, A. (1940). *Loḡatnāme-ye Dehkhodā [Dehkhoda Dictionary]*. Tehran University Press.

Durkin, P. (Ed.) (2016). *The Oxford handbook of lexicography*. Oxford University Press.

Finegan, E. & Adams, M. (Ed.) (2024). *The Cambridge handbook of the dictionary*. Cambridge University Press.

Fuertes-Olivera, P. A. (Ed.) (2018). *The Routledge handbook of lexicography*. Routledge Taylor & Francis Group.

Jackson, H. (Ed.) (2022). *The Bloomsbury Handbook of Lexicography*. Bloomsbury Publishing Plc.

Mo'in, M. (1972). *Farhang-e fārsi-ye Mo'in [Mo'in Persian dictionary]*. Amir Kabir.

Tarp, S. (2008). *Lexicography in the borderland between knowledge and non-knowledge: General lexicographical theory with particular focus on learner's lexicography*. Max Niemeyer Verlag.

Contact information

Saghar Sharifi
Department of Teaching English and Translation, Ka.C., Islamic Azad University, Karaj, Iran
saghar.sharifi@kiau.ac.ir

Abolfazl Alamdar
Department of Linguistics, Allameh Tabataba'i University, Tehran, Iran
abalfazl_alamdar@atu.ac.ir

LLM Statement

As non-native speakers of English, the authors employed Grammarly and ChatGPT as editorial aids during the preparation of this manuscript. These tools were used to identify and rectify grammatical and orthographic errors, enhance linguistic consistency and coherence, refine the phrasing of the case studies, and standardize the formatting of the annotated references. Following the use of these resources, the authors conducted a thorough review and made any necessary revisions, assuming full responsibility for the final published version.

Jóhannes B. Sigtryggsson

THE EFFECTS OF PURISM ON A NINETEENTH-CENTURY ICELANDIC BILINGUAL DICTIONARY

Abstract The Icelandic linguist and philologist Konráð Gíslason (1808–1891) was the author of an influential Danish–Icelandic dictionary published in 1851, the first bilingual dictionary of its kind. The significance of this work lay less in its impact on lexicography than in its role in the history of linguistic purism in Iceland. One of its most distinctive features was the near absence of loanwords as Icelandic equivalents for Danish headwords, despite the widespread use of such loanwords in both spoken and written Icelandic at the time. Although the vigorous coinage of neologisms was a hallmark of nineteenth-century linguistic purism in Iceland, Gíslason's dictionary contains relatively few newly coined terms. Instead, Gíslason relied primarily on existing vocabulary and occasionally revived archaic words attested in medieval texts as equivalents for Danish lemmas. He also frequently paraphrased the meaning of the Danish words through explanatory phrases rather than resorting to loanwords. His uncompromising puristic ideals limited the dictionary's practical usefulness and made it a less accurate representation of contemporary Icelandic usage.

Keywords Purism; historical dictionaries; Icelandic; loanwords

1 Introduction

The Nordic philologist and language purist Konráð Gíslason (1808–1891) published a comprehensive Danish–Icelandic dictionary in 1851 (*Dönsk orðabók með íslenzkum þýðingum* 'Danish Dictionary with Icelandic translations', hereafter KG1851). The work appeared during a period of transition in Icelandic history, when the traditional agrarian society was undergoing modernisation. This transformation introduced a wide range of new concepts and ideas from abroad. At the same time, linguistic purism was gaining momentum in Iceland, characterised by an increasing emphasis on the creation of neologisms, particularly towards the end of the nineteenth century. These developments brought with them a substantial influx of international terms and new words, raising the question of how such elements should be treated within the conservative Icelandic linguistic tradition. The use of neologisms instead of loanwords has been a central feature of this tradition ever since.

This paper examines the extent to which Gíslason sought to avoid loanwords in his dictionary, and the strategies he employed to achieve this aim. Particular attention is paid to Danish lemmas marked with an asterisk (*) in the dictionary, which Gíslason adopted from Meyer's (1844) Danish loanword dictionary. Section 2 discusses the origins of KG1851 and the lexicographical situation in Iceland in the early nineteenth century. Section 3 provides a brief overview of the history of the purism movement in Iceland. Section 4 describes Gíslason's lexicographical methods and the various ways in which he introduced new concepts and ideas into the language without resorting to loanwords. Finally, section 5 offers concluding remarks and argues that Gíslason took his puristic ideals too far in the dictionary, resulting in an amalgam of Old and Modern Icelandic that was in many respects removed from contemporary usage.

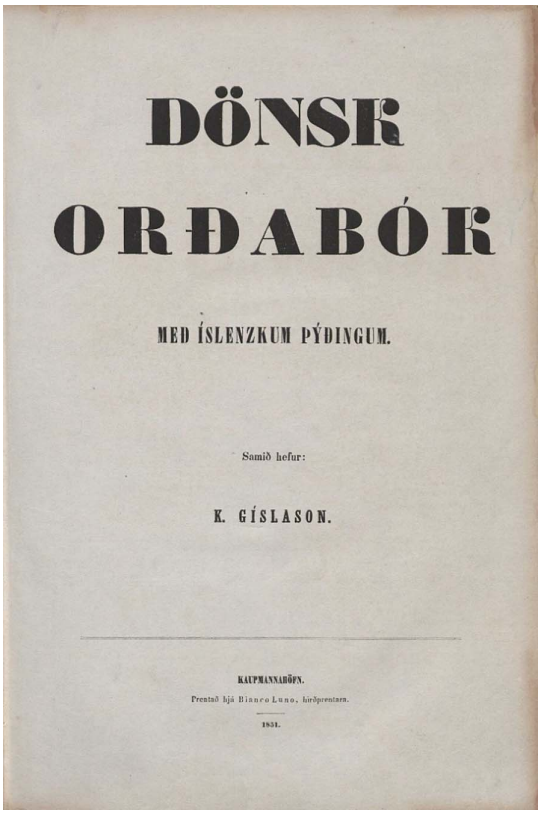

DÖNSK
ORÐABÓK
MED ÍSLENZKUM PÝDINGUM.
Samið hefur:
K. GÍSLASON.
KAUPMANNAHÖFN.
Prentað hjá Bianco Luno, hirðprentara.
1851.

FIGURE 1 The cover of KG1851

2 Dönsk orðabók með íslenzkum þýðingum (1851)

2.1 *The lexicographic situation in early 19th century Iceland*

Following the Reformation in Iceland in the mid-sixteenth century, Danish influence on both the administrative system and the Icelandic language became increasingly prominent. In nineteenth-century Iceland, Danish was widely used alongside Icelandic as an administrative language, particularly in legislation and, frequently, in commercial contexts. The highest level of governance was based in Copenhagen, as well as the High Court (Ottósson, 1990, pp. 24–26, 57–61).

Until the nineteenth century, no bilingual dictionaries between Danish and Icelandic had been published. An Icelandic–Latin–Danish dictionary was printed in 1814 (Halldórsson, 1814)[1] and the first Danish–Icelandic dictionary in 1819 (Oddsson, 1991), *Orðabók sem inniheldur flest fágæt, framandi og vandskilin orð er verða fyrir í dönskum bókum* ("Dictionary containing most rare, foreign, and difficult words occurring in Danish books"), although it was limited in scope and restricted to rare and foreign words in Danish. There was thus a pressing need for a comprehensive Danish–Icelandic dictionary.

Konráð Gíslason and Jóhann Halldórsson (1809–1844), another Icelandic academic, began work on KG1851 in 1840. They had originally intended to edit the dictionary jointly, but Halldórsson passed away in 1844. By that time, he had worked on the entries beginning with the letter *M* and parts of *N*, as Gíslason notes in the foreword to KG1851 (p. iii). Gíslason compiled the remaining sections of the dictionary and also revised and completed Halldórsson's

[1] A new edition of this dictionary was published in 1992 (Halldórsson, 1992).

contributions. The final work is therefore essentially Gíslason's and only his name appears on the title page.

Gíslason and Halldórsson applied for grants to support the project in 1841 and also circulated a letter to prospective subscribers in Iceland in 1842, in which they outlined the rationale for the work (quoted in Kristjánsson, 2003, p. 204). In this letter, they refer to the absence of such a dictionary and emphasise the importance of preserving the purity of the Icelandic language in a work of this kind.

> Icelandic laypersons have not infrequently complained of the absence of a Danish dictionary, not for the purpose of adopting words and idioms from it so as to corrupt their own language, but rather in order to fully understand Danish books, such as those dealing with agriculture and handicrafts, describing methods and the tools employed therein, or addressing matters of governance, trade, and similar subjects. Those required to learn Danish have an even greater need for such a dictionary and must rely on it much more frequently, as is the case with officials and candidates for public office. Moreover, it appears to us to be certain that a Danish dictionary with Icelandic explanations, provided that its composition does not entirely fail, may prove highly beneficial to our language. We shall therefore spare no effort to ensure that the explanations are as pure as possible (author's translation).

The completed work was published in 1851 and printed in 1,100 copies. It comprised 596 pages in large folio format, printed in a dense three-column layout, and contained a total of 52,620 headwords.

2.2 The structure and sources for the dictionary

The Danish vocabulary (lemma list) and the structure of the dictionary entries in KG1851 are primarily derived from C. Molbech's *Dansk Ordbog indeholdende det danske Sprogs Stammeord* ... (1833), as Gíslason notes in the foreword to the dictionary (KG1851, p. iii), whose structure he follows closely. Gíslason further states that he also made use, in certain instances, of the early volumes of the large Danish dictionary of the Royal Danish Society of Sciences (*Dansk Ordbog* ..., 1793–1848), which was not completed until 1905, although its use in KG1851 appears to have been very limited (cf. Kvaran, 2013, p. 99).

To this basic list of headwords and article structure derived from Molbech's dictionary, Gíslason added approximately 7600 lemmas from Meyer's (1844) loanword dictionary. The purpose of Meyer's work was to explain the meanings of various foreign words that had by that time become common in written Danish. In the preface to KG1851 (p. iv), Gíslason states that he has placed an asterisk (*) before words 'which are not Gothic or have not entered Danish from the Gothic languages. However, those words that have become so fully naturalised in Danish that they no longer appear foreign are omitted' (author's translation). He does not explicitly state that the words marked with an asterisk were taken from Meyer; however, a comparison makes this evident.

3 The purism movement in Iceland

Linguistic purism in Iceland has its origins in Humanist writings of the sixteenth century. The scholar Arngrímur Jónsson argued in Latin that Icelandic was the ancient language of Northern Europe and that Icelanders alone preserved it in an uncorrupted form (cf. Jensson, 2003; Halldórsson, 1979). In its initial phase, the purism movement aimed at replacing Danish loanwords and modifying foreign syntactic constructions, but placed less emphasis on the

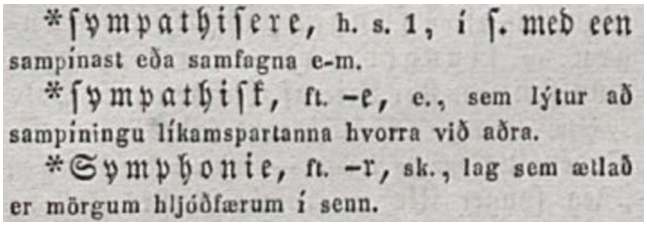

*ſympathiſere, h. s. 1, í ſ. með een sampinast eða samfagna e-m.
*ſympathiſk, ft. -e, e., sem lýtur að sampiningu líkamspartanna hvorra við aðra.
*Symphonie, ft. -r, sk., lag sem ætlað er mörgum hljóðfærum í senn.

FIGURE 2 Examples of lemmas, sympathisere 'sympathise', sympathisk 'likeable' and symphonie, borrowed from Meyer's (1844) loanword dictionary

formation of new words (Halldórsson, 1979, p. 77). In the late eighteenth century, this began to change, particularly in the publications of the Icelandic learned society, *Hið íslenska lærdómslistafélag*, where strong emphasis was placed on the creation of neologisms. The purism movement subsequently gained further momentum in the nineteenth century as part of Icelandic nationalism (Ottósson, 1990, p. 51). At that time, it was characterised by its strong focus on neologism formation, especially towards the end of the nineteenth century (Árnason, 2017, p. 363; Ottósson, 1990, pp. 100–103). The language purism movement retained its strong influence throughout the twentieth century.

There is a clear influence of linguistic purism in the aforementioned prospectus addressed to potential buyers of KG1851 (*Boðsbréf um danska orðabók*, 1842; see section 2). Gíslason and Halldórsson state in the prospectus that "we shall therefore spare no effort to ensure that the explanations are as pure as possible" (author's translation). In his biography of Gíslason, Ólsen (1891, p. 66) argues that the principal motivation for undertaking this work, i.e. KG1851, was Gíslason's desire to purge Icelandic of Danish calques and other Danish influences. Ólsen further notes that Gíslason was troubled by the fact that educated individuals often conducted much of their thinking in Danish and consequently expressed their thoughts in Danish linguistic forms, even when speaking or writing in Icelandic. Ólsen (1891, p. 66) asserts that no book published during the nineteenth century played as significant a role in the eradication of Danish influence from the language as KG1851.

The dictionary was thus compiled within a linguistic purism environment in which considerable emphasis was placed on the elimination of loanwords. As I demonstrate in section 4, however, Gíslason went even further in this respect than many of his contemporaries.

4 Konráð Gíslason's lexicographical methods

4.1 Introduction

It is interesting to look closer at the methods and means Gíslason used to express his ideas about linguistic purism in his lexicographic work. What is striking is how infrequently loanwords are used as equivalents and, instead, the frequent use of archaic lexical items as synonyms, as well as the rendering of meanings through full explanatory phrases. By contrast, neologisms are comparatively rare in Gíslason's dictionary. Icelandic equivalents appear to be employed primarily in cases where Gíslason considered them fully satisfactory.

4.2 Different font sizes

It is noteworthy to examine how Gíslason used different font sizes to distinguish between words that he approved of and others towards which he was less favourably disposed. He used

a standard-sized typeface for words attested in Old Icelandic, as well as for equivalents and neologisms that he regarded as satisfactory, whereas a smaller typeface was employed for those he considered less successful or less "pure".

Figure 3 provides examples of this practice. For the lemma *syndebyrde*, the Icelandic equivalent *syndabyrði* appears in the standard font size and is identified as archaic ("fornt"). Likewise, the equivalent *syndabætur* for the lemma *syndebod* 'redemption', which is found in the medieval manuscript AM 619 4to ("619"), is printed in the standard font size, whereas the other equivalents and paraphrases shown here appear in a smaller typeface.

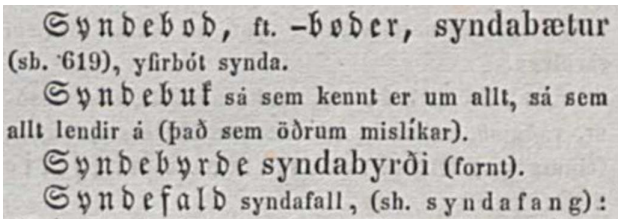

Syndebod, ft. -bøder, syndabætur (sb. 619), yfirbót synda.
Syndebuf sá sem kennt er um allt, sá sem allt lendir á (það sem öðrum mislíkar).
Syndebyrde syndabyrði (fornt).
Syndefald syndafall, (sh. syndafang):

FIGURE 3 Examples of different font sizes in four lemmas on p. 489 in KG1851

4.3 Equivalents

Icelandic equivalents, of course, frequently occur in KG1851 (see Figure 4), although less extensively than in most other bilingual dictionaries. They are presented either in a standard-sized typeface or in a smaller one (see section 4.2). Gíslason appears to have used equivalents in the standard typeface only when they were attested in Old Icelandic or when synonymous expressions were well established in the language and not of foreign origin. An example of more than one equivalent in normal font size given for a Danish headword are the adjectives *ákafur*, *fjörugur* and *öflugur* for *energisk* (KG1851, p. 102). In Figure 4 there are shown some more examples of equivalents given in a normal font size without a comment. They are all old common words in Icelandic such as *kjálki* for *kjævebeen* 'jaw'. Often there is more information or clarification added in a smaller font size. An example of that is *kjæverem*, where the specification "(á beizli)" 'of a bridle' is added to the equivalent *kverkól*.

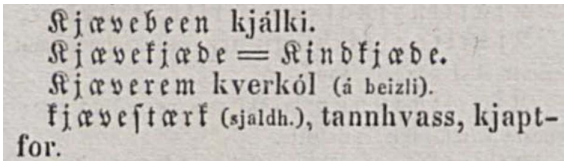

Kjævebeen kjálki.
Kjævefjæde = Kindfjæde.
Kjæverem kverkól (á beizli).
fjævestærk (sjaldh.), tannhvass, kjaptfor.

FIGURE 4 Examples of Icelandic equivalents on p. 240 in KG1851

4.4 Archaic words

Concurrent with his work on KG1851, Gíslason was also engaged in the compilation of a dictionary of Old Icelandic, a project he pursued from 1840 to 1854. This work later formed the basis of *An Icelandic-English Dictionary* (1874) (cf. Kristjánsson, 1996, pp. 19–33, 2003, pp. 190–203). In connection with this project, he collected a substantial number of examples from medieval Old Icelandic manuscripts and thus had extensive access to historical linguistic citations, which he appears to have drawn upon freely in his work on KG1851 (cf. Benediktsson, 1969, p. 105). In the preface to his Danish-Icelandic dictionary (KG1851, p. iv), Gíslason states

that he considers one of the principal merits of the work to be his consistent reference to ancient texts and the inclusion of examples drawn from them.

In general, such references derive from well-known medieval works, such as the Sagas of Icelanders. An illustrative example is the entry for the Danish word *krigsgud* ('war god'), for which three Icelandic equivalents are provided: *bardagaguð*, *orustuguð*, and *vígaguð* ("bardagaguð (Al.), orustuguð (Trójum.), vígaguð (Sn. E.)") (KG1851, p. 257)). The first word is taken from the thirteenth-century *Alexanders saga*, the second from the *Trójumanna saga*, also dated to the same century, and the third from Snorri Sturluson's *Edda* (the *Younger Edda*).

Notably, Gíslason occasionally cites medieval manuscripts or documents as sources for lexical translations, even though the texts in question had not yet been published at the time. In the preface to KG1851 (p. vi), he provides a list of the various works and manuscripts cited in the dictionary. In total, seventeen medieval manuscripts are used as direct sources. One example is the Danish word *meened* (KG1851, p. 301), for which Gíslason gives the Icelandic equivalents *meineiður*, *meinsæri*, and *ljúgeiður*. For *meineiður*, he cites the manuscript AM 619 4to, and for *ljúgeiður* the manuscript AM 656 A II 4to. The former is dated to approximately 1200–1225, and the latter to around 1250 (cf. ONPReg, pp. 457, 460). For the Danish word *fireaarig* ('four-year'), he uses the Icelandic equivalent *feræ*r and cites as its source an Icelandic medieval charter from 1395 ("ísl. skjal 1395") (KG1851, p. 115).

4.5 Meaning paraphrase

In those cases where Gíslason was unable to find suitable Old Icelandic words or established equivalents, he frequently resorted to paraphrasing the meaning rather than providing a single lexical equivalent. Benediktsson (1969, p. 105) observes that "quite often, he abandoned any attempt to translate the Danish headwords directly and instead explained their meaning, at times at considerable length" (author's translation). The following examples illustrate such paraphrastic renderings of items lacking direct Icelandic equivalents. The Icelandic explanations have been translated into English by the present author.

- *efterret* 'dessert' (KG1851, p. 97): food that is served after the main food
- *erobringsret* 'right of conquest' (KG1851, p. 105): a claim to rights or authority over something on the grounds that one has conquered and taken possession of it
- *jomfruolie* 'virgin olive oil' (KG1851, p. 228): the purest olive oil, the first portion that runs from the churn
- *negligee* 'dressing gown, negligee' (KG1851, p. 325): a garment worn indoors when no strangers are present (see Figure 5)
- *orthopædie* (KG1851, p. 349): the art of treating kinks or deformities in adolescents

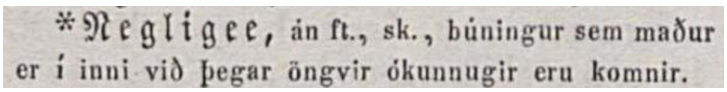

*Negligee, án ft., sk., búningur sem maður er í inni við þegar öngvir ókunnugir eru komnir.

FIGURE 5 The headword *negligee* in KG1851 and the explanation of its meaning

Introductory phrases such as *sá sem* ... ('he who ...'), *það að* ... ('that', 'the fact that ...'), *það sem* ... ('that which ...'), and *eins konar* ('a kind of') are frequently used to introduce such

semantic paraphrases. The following examples illustrate the use of these introductory formulae.

- *eclectiker* 'eclectic' (KG1851, p. 94): one who selects from the philosophies of others what he considers best
- *eddikehonning* 'oxymel' (KG1851, p. 94): a kind of beverage made from honey and vinegar
- *elastisticet* 'elasticity' (KG1851, p. 99): the property of being able to yield and return to its original form
- *enclisis* 'enclitic' (KG1851, p. 101): the fact that small words lose their stress and it is shifted onto the preceding word

In such instances, Gíslason could often have readily employed established Danish loanwords corresponding to the lemmas, yet he chose almost never to do so. For example, the loanword *edik* 'vinegar' has been attested in Icelandic at least since the fourteenth century (ONP) and was common in both spoken and written Icelandic in the nineteenth century, as it is today. The same applies to many other well-established loanwords of the period, which Gíslason nevertheless refrained from using in KG1851.

Typically, paraphrased material is presented in a smaller typeface. This typographical distinction appears to indicate that Gíslason frequently employed this strategy to provide readers with an understanding of the referent without implying that the paraphrase should function as an equivalent, usable translation of the lemma.

4.6 Neologisms

Neologisms are very rare in KG1851. Ottósson (1990, p. 101) notes that the word *vél* ('machine') is first attested in this sense in the dictionary. Gíslason occasionally also refers to more recent neologisms, as in the case of the lemma *æther*, for which he provides the equivalents *eldlopt* and *vindblá* from the dictionaries by B. Halldórsson (1814) and G. Oddsson (1819), respectively, as well as the neologism *ljósvaki* ('he who awakens the light'), coined some years earlier by Gíslason's friend, the poet Jónas Hallgrímsson.

An interesting case is the word *electricitet*, where Gíslason uses the neologism *rafmagn*. The oldest example of it in The Written Corpus of the Dictionary of the University of Iceland (Ritmálssafn) is from 1848. This seems to be an example of a very recent neologism that Gíslason uses (cf. Kvaran, 2013, pp. 109–110).

4.7 Loanwords

Although Gíslason incorporated a large number of international headwords from Meyer's dictionary (1844), he largely refrained from using established Icelandic loanwords corresponding to Danish terms as equivalents in KG1851. He went to considerable lengths in this respect, avoiding even loanwords that were widely used in nineteenth-century Icelandic. In this regard, his approach to linguistic purism may be characterised as particularly stringent.

I have examined loanwords in one section of the dictionary, namely lemmas beginning with the letter *E* (pp. 94–107 in KG1851). In total, KG1851 contains 1076 lemmas beginning with this letter, of which 291 are marked with an asterisk. In the corresponding dictionary entries, only a few clear loanwords are used as equivalents. The following is a complete list of these instances.

In the entry for *eftermale* (KG1851, p. 97) ('paint' (verb)), the Middle English loanword *penta* is used by Gíslason ("penta e-ð upp aptur – b) penta eptir e-u"), and for the corresponding noun *eftermaling*, he uses the paraphrase "það sem pentað er" ('that which is painted'). In the entry for *egn* ('land'), Gíslason includes *pláss* in parentheses as one of several equivalents for the term (KG1851, p. 99) and notes that it is commonly used but not Icelandic ("pláss algengt, en ekki íslenzka"). For the nouns *email* and *emailleur* ('enamel, enameler') and the verb *emaillere* ('to enamel/glaze') (KG1851, p. 100), he gives the equivalents *smelting*, *smelta (amaljera)*, and *smeltari*, which are considered loanwords from Middle Low German *smelt* (Íslensk orðsifjabók).

Additional examples of loanwords can be found elsewhere in the dictionary. One such case is "*bankóseðill* (*gjaldbleðill*)" (KG1851, p. 30), given as an equivalent for the Danish *bancoseddel* 'banknote'. Another example is *jarðamangari*, provided for the Danish lemma *jordpranger* 'estate agent' (KG1851, p. 229), which is considered to derive from the Old English verb *mangian* (Íslensk orðsifjabók).

5 Conclusions

As Kvaran (2013, p. 117) observes, Gíslason's tendency to provide explanatory translations rather than straightforward lexical equivalents undoubtedly disappointed many users who consulted the dictionary in search of direct Icelandic counterparts to Danish headwords. Gíslason himself regarded the dictionary's extensive references to Old Icelandic literature as one of its principal merits. However, in doing so he distanced the work from contemporary usage, with the result that the dictionary became a hybrid of archaic and contemporary language and, consequently, less user-friendly than it might otherwise have been.

The dictionary was, therefore, in some respects a successful attempt to purge Icelandic of Danish loanwords. At the same time, by using these strict purism methods, Gíslason lost sight of the primary role of dictionaries: to provide guidance on suitable equivalents in the target language and describing the contemporary vocabulary of the target language, i.e. Icelandic. Consequently, its influence was likely more limited than it might otherwise have been. The next Danish-Icelandic dictionary, published at the end of the century, was compiled by Jónasson (1896). It had a much greater impact on spreading neologisms, used equivalents wherever possible and and proved more accessible to the general public (cf. Benediktsson, 1969, p. 107).

References

*An Icelandic-English dictionary/*based on the MS. collections of Richard Cleasby, enlarged and completed by Gudbrand Vigfusson; with an introduction and life of Richard Cleasby by George Webbe Dasent (1874). Clarendon Press.

Árnason, K. (2018). Language planning and the structure of Icelandic. In A. P. Kristinsson & H. Þorgeirsson (Eds.), *Á vora tungu. Afmælisrit til heiðurs Kristjáni Árnasyni* (pp. 357–380). Háskólaútgáfan.

Benediktsson, J. (1969). Íslenzk orðabókarstörf á 19. öld. *Andvari*, 11, 96–108.

Boðsbréf um danska orðabók. (1842).

Dansk Ordbog udgiven under Videnskabernes Selskabs Bestyrelse 1793–1848. I–VI. Kiøbenhavn.

Halldórsson, B. (Ed.). (1814). *Lexicon Islandico-Latino-Danicum Biörnonis Haldorsonii.* Biørn Haldorsens islandske Lexikon. Ex Manuscriptis Legati Arna-Magnæani cura R. K. Raskii editum. I–II. Havniæ.

Halldórsson, B. (Ed.). (1992). *Orðabók íslensk – latnesk – dönsk*. Eftir handriti í Stofnun Árna Magnússonar í Kaupmannahöfn. Fyrst gefin út árið 1814 af Rasmusi Kristjáni Rask. Jón Aðalsteinn Jónsson sá um útgáfuna. Orðabók Háskólans.

Halldórsson, H. (1979). Icelandic Purism and Its History. *Word*, 30, 76–86.

Íslensk orðsifjabók: Magnússon, Á.B. *Íslensk orðsifjabók*. Retrieved May 20, 2026, from https://ordsifjabok.arnastofnun.is/

Jensson, G. Þ. (2003). Puritas nostræ lingvæ. Upphaf íslenskrar málhreinsunar í latneskum húmanisma. *Skírnir*, 177, 37–67.

Jónasson, J. (Ed.) (1896). *Ný dönsk orðabók með íslenzkum þýðingum*. Reykjavík.

Karlsson, S. (2004). *The Icelandic Language*. Translated by R. McTurk. Viking Society for Northern Research.

KG1851: Gíslason, K. (Ed.) (1851). *Dönsk orðabók með íslenzkum þýðingum*. Kaupmannahöfn.

Kristinsson, A. P. (2017). *Málheimar. Sitthvað um málstefnu og málnotkun*. Háskólaútgáfan.

Kristjánsson, A. (1996). Orðabókarstörf Konráðs Gíslasonar. *Íslenskt mál og almenn málfræði* 18, 7–36.

Kristjánsson, A. (2003). *Síðasti Fjölnismaðurinn. Ævi Konráðs Gíslasonar*. Skrudda.

Kvaran, G. (2013). Orðabókarstörf Konráðs Gíslasonar. *Orð og tunga*, 15, 95–119.

Meyer, L. (Ed.) (1844). *Fremmedord-Bog, eller kortfattet Lexikon over fremmede, i det danske Skrift - og Omgangs-sprog forekommende Ord, Konstudtryk og Talemaader; tilligemed de i danske Skrift er mest brugelige, fremmede Ordforkortelser*. Anden, forøgede og forbedrede Udgave. Forlagt af J.H. Schubothes Boghandling.

Molbech, C. (1833). *Dansk Ordbog indeholdende det danske Sprogs Stammeord, tilligemed afledede og sammensatte Ord, efter den nuværende Sprogbrug forklarede i deres forskiellige Betydninger, og ved Talemaader og Exempler oplyste*. Paa den Gyldendalske Boghandlings Forlag.

Oddsson, G. (1991). *Orðabók sem inniheldur flest fágæt, framandi og vandskilin orð er verða fyrir í dönskum bókum*. [Originally published in 1819] Ný útgáfa með íslenskri orðaskrá. Jón Hilmar Jónsson sá um útgáfuna ásamt Þórdísi Úlfarsdóttur. Orðfræðirit fyrri alda I. Orðabók Háskólans.

ONP: *ONP: Dictionary of Old Norse Prose*. Retrieved May 20, 2026, from https://onp.ku.dk/onp/onp.php?

ONPReg: *Ordbog over det norrøne prosasprog. Registre./A Dictionary of Old Norse Prose. Indices* (1989). København.

Ottósson, K. G. (1990). *Íslensk málhreinsun. Sögulegt yfirlit*. Rit Íslenskrar málnefndar 6. Íslensk málnefnd.

Ólsen, B. M. (1891). Konráð Gíslason. *Tímarit Hins íslenzka bókmenntafjelags*, 12, 1–96.

Ritmálssafn: *Ritmálssafn Orðabókar Háskólands*. Retrieved May 20, 2026, from https://ritmalssafn.arnastofnun.is/

Contact information

Jóhannes B. Sigtryggsson
The Árni Magnússon Institute for Icelandic studies
johannes.b.sigtryggsson@arnastofnun.is

Michael Spagnol, John J. Camilleri & Dwayne Ellul

FROM *ḊAMMA* TO *ĠABRA*

Four Centuries of Maltese Lexicography

Abstract This paper presents a historical and typological overview of Maltese lexicography, tracing the development of lexical resources from the earliest manuscript dictionaries to recent digital initiatives. It surveys major multilingual, bilingual, monolingual and specialised dictionaries while discussing the methodologies underlying Maltese lexicography over the years. Emphasis is given to the interaction between the hybrid lexical structure of Maltese and lexicographic practice, including issues of lemmatisation, orthography and dictionary structure. The paper also discusses recent digital initiatives, particularly *Ġabra* and *Dizzjunarju.mt,* which illustrate the growing role of computational methods in Maltese lexicography.

Keywords Maltese lexicography; digital resources; dictionary structures; lemmatisation

1 Introduction

Maltese lexicography spans almost four centuries and reflects the complex linguistic history of Maltese itself, a language shaped by extensive and prolonged language contact. Despite this tradition, there is no comprehensive overview bringing together the major lexicographic works, methodologies and recent digital developments relating to Maltese. This paper seeks to address that gap by tracing the development of Maltese lexical resources from the earliest manuscript dictionaries to contemporary online projects.

This paper is organised as follows. First, we outline the lexical stratification of Maltese. After tracing the historical development of the lexicons of Maltese, we discuss a number of methods and problems in Maltese lexicography. Following that, we describe digital resources for Maltese, focusing on *Ġabra*, an online collection of lexical resources. Finally, we mention ongoing lexicographic projects.

2 Lexical Characteristics of Maltese

Maltese is a language that typologically occupies a position between Semitic and Romance languages. While it is genealogically related to Arabic, Maltese has been in close contact with Italian and Sicilian for approximately a millennium. In addition, the past two centuries have seen significant language contact with English. This long history of contact between typologically diverse languages is mirrored in the stratification of the language, which can be broadly described in terms of three major strata.

The main stratum is Semitic, forming the basis of the phonology, morphology, and basic lexicon. Maltese shares many characteristics with Western Arabic, in particular Maghrebi dialects (among others, Aquilina, 1961, 1979). However, "[i]t is likely that successive waves of impact reached the Maltese shores from different Arab stations and at different points in the island's history" (Mifsud, 2008, p. 146). Maltese also displays several Levantine traits (Borg, 1994, 1997; Borg & Mifsud, 2002), and some researchers suggest that the main stratum of Maltese is Siculo-Arabic (Brincat, 2004; Agius, 1996).

The Romance superstratum is chiefly represented by lexical, syntactic, and, to a lesser extent, phonological and morphological influence from Sicilian and southern Italian dialects

(roughly from 1090 to 1530) and from Italian (from 1530 onwards). To some extent, the influence of Italian persists today through television, geographical and cultural proximity, as well as tourism and commerce (cf. Brincat (1998, 1999) and Caruana (2006) on the acquisition of Italian through television among Maltese speakers). For this reason, Italian can still be regarded as an adstratum in the linguistic stratigraphy of Maltese.

Alongside Italian, English constitutes a strong adstratum, contributing lexical material mainly dating from the period of British rule (1800–World War II). Its influence intensified further in the postwar period, when English emerged as a global lingua franca. Since the development of Maltese is characterised by the presence of other languages, usually of greater prestige, it can thus also be considered from a diglossic perspective, with Maltese functioning as a low variety alongside successive high varieties including Latin, Sicilian, Italian, French, and English (Brincat, 2004: pp. 25-27, 249ff.).

Over the years, other studies have attempted to quantify the extent of lexical borrowing in Maltese, though their findings vary significantly depending on the data sources and methodologies employed. Brincat (2000, 2004, 2011) provides a comprehensive etymological classification of the entries in Aquilina's (1987–1990) dictionary, thereby offering a large-scale type-based perspective. His analysis reveals that the etymological sources of the dictionary entries are distributed as follows: Semitic 32.41%, Siculo-Italian 52.46%, and English 6.12%, with the remainder comprising local formations and items of unclear origin.

However, corpus-based frequency analyses (type vs. token) of journalistic texts reveal a different picture: words of Semitic origin account for 72.92% of actual usage, primarily because prepositions, pronouns, and other high-frequency grammatical items are, in the vast majority of cases, derived from Arabic (Fenech, 1978: 140; Brincat, 2004, pp. 363-366; 2011, pp. 411-415). Comrie and Spagnol (2016), using a 1,500-word list from the Loanwords in the World's Languages project, show that lexical borrowing in Maltese varies significantly by word class: function words and verbs are predominantly Arabic, while nouns and adjectives show a much higher proportion of Romance and English origin. Their findings also reveal that Romance-origin terms outweigh English ones across all semantic domains, including those related to modern life, reflecting the deeper and longer-standing influence of Sicilian and Italian.

Despite differences in scope, all studies converge on the finding that Romance and, to a lesser degree, English have left a significant imprint on the Maltese lexicon. However, the degree of influence appears to vary depending on whether the focus is on types (dictionary entries) or tokens (actual usage). Type-based analyses highlight the sheer number of Romance-derived lexical items, while token-based data reveal the dominance of Semitic-origin words in actual use, especially among grammatical items. This divergence underscores the complexity of labelling Maltese as a "mixed language", a term often invoked but seldom precisely defined. The question of Maltese's typological classification within contact linguistics thus remains unresolved, shaped as much by methodological decisions as by linguistic facts (Stolz, 2013).

The constant interaction between Semitic and Romance structures has given rise to a dual morphological system (Spagnol, 2011). For nearly every lexical category, learners of Maltese are faced with two distinct word-formation strategies. For example, verbs, deverbal nouns and past participles, may be formed either through the combination of a consonantal root with verbal or nominal patterns (Table 1a), or through the concatenation of verbal or nominal affixes to a root or stem (Table 1b). The implications of this dichotomous morphology for lexicography are discussed below.

TABLE 1a Root-and-pattern word formation strategy

Consonantal Root	Template	Word	Meaning
√żjn	1v22v3 tv12ī3 m1v22v3	*żejjen* *tiżjin* *mżejjen*	'to adorn' 'adornment, adorning' 'adorned'

Note: Templates are represented using numbers for consonants and Vs for vowels. Numbers correspond to root consonants; doubled numbers indicate gemination. Some patterns include fixed vowels, such as long ī, marked with a macron, or fixed consonants, such as m.

TABLE 1b Concatenative word formation strategy

Syllabic Root	Suffix	Word	Meaning
√(d)dekor	-a -azzjōni -āt	*ddekora* *dekorazzjoni* *ddekorat*	'to decorate' 'decoration' 'decorated'

3 History of Maltese Lexicography

The history of Maltese lexicography dates to the first half of the seventeenth century. Drawing on Agius (1990), Aquilina (1961, 1980), Cassar Pullicino (1947), Cremona (1933), Privitera (1964), Saydon (1953), Spagnol (2006) and Camilleri (2013), we provide here a list intended to be as comprehensive as possible of the dictionaries of Maltese. The list includes not only LGP (language for general purposes) dictionaries but also a selection of children's and picture dictionaries as well as specialised dictionaries. This list is also available as a Zotero Shared Library[1].

3.1 LGP Dictionaries (in Chronological Order)

1. Thezan (a French Knight) (circa mid 17th century) *Regole per la Lingua Maltese* (Cassola, 1992, Cachia, 1994, pp. 41-55 on this earliest lexicographic work). Cachia published this dictionary as *L-Ewwel Dizzjunarju Malti: Malti-Taljan tal-Kavallier Thesan* in 1999.

2. Francesco Bardon (17th century) *Dizzjunarju Malti-Taljan-Franċiż*; Cassar Pullicino (1947, pp. 43-44) refers to this dictionary, most probably compiled in the 17th century, which however was never found.

3. (no known author) (17th century) *Il mezzo vocabolario maltese-italiano del '700*. This partial dictionary, dated between 1765 and 1775, was published by Arnold Cassola in 1996.

4. G.P.F. Agius De Soldanis (18th century) *Ḋamma tal Kliem Kartaginis mscerred fel fom tal Maltin u Ghaucin*; This is mainly in manuscript form and consists of four volumes: Maltese-Italian-Latin (vol.1) published in 2016; Italian-Latin-Maltese (vol.2-3); Latin-Italian-Maltese (vol.4). Volumes 2-4 remain in manuscript form. While technically a dictionary, it is known for its wealth of encyclopedic information (Ellul, 2024, pp. 415-416).

5. Francesc de Sentmenat-Torelles i d'Agulló († 1762), *Un Vocabulari Català-Maltès* published as *Un Vocabulari Català-Maltès Manuscrit del Segle XVIII* by Alexandre Queraltó Bartrés (2003).

[1] https://www.zotero.org/groups/6613659/maltese_lexicography/library

6. Michaelis Antonii Vassalli (1796) *Ktŷb yl-Klŷm Mâlti 'Mfysser byl-Latin u byt-Talyân*, also known as *Lexicon Melitense-Latino-Italum*. Rome; This lexicon was also published by Frans Sammut as Lexicon in 2002 and by Lawrenz Cachia as *Lexicon jew Dizzjunarju Malti-Latin-Taljan* in 2016.

7. (first half of the 19th century) *Vocabolario Francese-Italiano-Maltese*; Agius (1990, p. 173) mentions this undated dictionary found at the National Library of Malta, which contains Maltese words with a mixture of Latin and Arabic letters.

8. (second half of the 19th century) *Una raccolta di parole maltesi coi loro equivalenti in italiano ed in latino*; Agius (1990, p. 173) also mentions this undated dictionary found at the National Library of Malta, which is a short word list with eighteen folios that lists Maltese words written in Arabic letters and bears no author's name. Privitera (1964, pp. 16-20) notes four other lexicographic works dating to the same period: (a) the *Necessario Supplemento al Vocabolario del Chierico Michele Antonio Vassallo stampato in Roma* (Serracino-Inglott, 1947, p. 38) which the Count Saver Marchesi started compiling but stopped at the letter C; (b) an unpublished Maltese-Italian dictionary which is attributed to Giovann Giuseppe Bellanti; (c) a Maltese-English dictionary which George Percy Badger was meant to publish but which was never found; and (d) the *Dizionario Etimologico Maltese* by Fortunato Panzavecchia, parts of which were published shortly before he died in 1850.

9. Giorgio Grognet de Vassé († 1862) *Dizionario e Rimario Atlantico Maltese e sue ragionate spiegazioni ed etimologie.*

10. Francesco Vella (1843) *Dizionario Portatile delle lingue Maltese, Italiana, Inglese.* Livorno.

11. Giovanni Battista Falzon (1845) *Dizionario Maltese-Italiano-Inglese*. Malta.

12. Vincenzo Azzopardi (1847) *Alfabeto maltese: perfetta imitazione dell'italiano. Con un vocabulario Maltese-Italiano-Inglese.* Malta.

13. Vincenzo Azzopardi (1849) *Alfabeto maltese: perfetta imitazione dell'italiano. Con un vocabulario Maltese-Italiano-Inglese*, second and third edition. Malta.

14. Fortunato Panzavecchia (1850) *Dizionario Fraseologico Italiano e Maltese, arricchito di proverbi ed di termini tecnici, con un breve proemio sull'uso della lingua Italiana in Malta, e sulla vergenza dell'idioma Maltese all'Italianismo.* Malta.

15. Vincenzo Azzopardi (1853) *Alfabeto maltese: perfetta imitazione dell'italiano. Con un vocabulario Maltese-Italiano-Inglese,* fourth edition. Malta.

16. Vincenzo Azzopardi (1856) *Piccolo Dizionario Maltese-Italiano-Inglese*. Malta.

17. Giovanni Battista Falzon (1882) *Dizionario Maltese-Italiano-Inglese*. Malta.

18. Salvatore Mamo (1885) *English-Maltese Dictionary*. Malta.

19. Vincenzo Busuttil (1900) *Diziunariu mill Malti ghall Inglis; Diziunariu mill Inglis ghall Malti.* Malta.

20. Antonio Emanuele Caruana (1903) *Vocabolario della Lingua Maltese*. Malta.

21. Enrico Magro (1906) *English and Maltese Dictionary* (from A to L). Malta; Aquilina (1987: xxv) had in his possession a printed but unpublished second part up to the entry "penny". Magro never finished the dictionary.

22. Salvatore Mamo († 1918) *Maltese-English Dictionary*. Unpublished manuscript. Malta.

23. Salvatore Mamo *English-Maltese Dictionary*. Unpublished manuscript. Malta.
24. Vincenzo Busuttil (1932) *Diziunariu mill Malti ghall Inglis,* second edition. Malta.
25. Charles Louis Dessoulavy (1938) *Maltese-Arabic Word-List.* London.
26. Carmelo Psaila (1936-1955) *Dizzjunarju Ingliż u Malti.* Malta.
27. Giuseppe Barbera (1939-1940) *Dizionario Maltese-Arabo-Italiano*. Beirut.
28. Edward Busuttil (1941) *Kalepin (Dizzjunarju) Malti-Ingliż.* Malta.
29. Edward Busuttil (1948-1949) *Kalepin Ingliż-Malti*. Malta.
30. Edward Busuttil (1950) *Kalepin (Dizzjunarju) Malti-Ingliż,* second edition. Malta.
31. Paul Bugeja (1952) *Dizzjunarju Ingliż-Malti*, Malta.
32. Paul Bugeja (1956) *Dizzjunarju Malti-Ingliż: kif titkellem bl-Ingliż*. Malta.
33. Paul Bugeja (1957) *Ġabra ta' tagħlim: dizzjunarju Ingliż-Malti.* Malta.
34. Edward Busuttil (1976) *Kalepin: Dizzjunarju Ingliż-Malti*, third edition. Malta.
35. Edward Busuttil (1977) *Kalepin: Dizzjunarju tliet ilsna: Malti-Ingliż-Taljan*. Malta.
36. Ahmed Talaat (1978) *Dizzjunarju Malti-Għarbi.* Malta.
37. Michael Agius (1978) *Dizzjunarju tal-Lsien Malti.* This is a manuscript of a pioneering Maltese monolingual dictionary, housed in the Archives and Rare Books Collection at the University of Malta's Library (Ellul 2022).
38. Edward Busuttil (1978) *Kalepin: Dizzjunarju tliet ilsna: Malti-Ingliż-Taljan*, second edition. Malta.
39. Erin Serracino-Inglott (1975-2003) *Il-Miklem Malti*. Malta; This is the first large-scale monolingual dictionary of Maltese. Agius (1990: 177-8) notes that this dictionary includes many entries that went unrecorded in earlier dictionaries, which Serracino-Inglott collected from marginal notes written on the manuscripts of other dictionaries, and "other data compiled by the Jesuit Patri Manwel Magri, seemingly forming an addendum to Falzon's Dizionario, which were passed on to Serracino Inglott by the Gozitan Jesuit Patri Gużepp Mejlak (d. 1966), who himself was compiling a voluminous dictionary 'Maltese-Italian-English' (up to the letter /h/) which, because of his death, never came to fruition."
40. Paul Bugeja (1982) *Kelmet il-Malti: Dizzjunarju Malti-Ingliż, Ingliż-Malti.* Malta.
41. Paul Bugeja (1984) *Kelmet il-Malti: Dizzjunarju Malti-Ingliż, Ingliż-Malti,* second edition. Malta.
42. Ludovik Schembri (1986) *Damma ta' Kliem, Dizzjunarju Malti-Ingliż-Taljan.* Malta.
43. Joseph Aquilina (1987-1990) *Maltese-English Dictionary*. Malta; This is the most comprehensive dictionary of Maltese to date. It includes many lexemes previously unrecorded, especially neologisms, technical and scientific terms, as well as dialectal items in current usage in Gozo and some Maltese villages. Entries have etymological information and grammatical descriptions followed by examples illustrated with idiomatic usages and sentences.

44. Paul Bugeja (1988) *Kelmet il-Malti: Dizzjunarju Malti-Ingliż, Ingliż-Malti*, third edition. Malta.

45. Charles Attard (1989) *Damma ta' kliem: dizzjunarju Ingliż-Malti-Taljan*. Malta.

46. Paul Bugeja (1990) *Kelmet il-Malti: Dizzjunarju Malti-Ingliż, Ingliż-Malti,* fourth revised edition. Malta.

47. Paul Bugeja (1990) *Pocket dictionary and phrase book: English-Maltese*. Malta.

48. Ġużè Diacono (1991) *Dizzjunarju Ingliż-Malti* (based on Carmelo Psaila's 1936-1955 dictionary), Pawlu Aquilina & David Calleja (editors). Malta.

49. Paul Bugeja (1993) *Pocket dictionary and phrase book: English-Maltese.* Malta.

50. Emanuel Cassar & David Calleja (1998) *Dictionary Dizzjunarju Maltese-English English-Maltese* (based on Ludovik Schembri's dictionary). Malta.

51. Grazio Falzon (1998) *Maltese-English, English-Maltese: dictionary and phrasebook. N*ew York.

52. Joseph Aquilina (1999-2000) *English-Maltese Dictionary (4* volumes). Malta.

53. Manfred Moser & Christopher Meilak (1999) *Dizzjunarju Malti-Ġermaniż, Ġermaniż-Malti, Wörterbuch Maltesisch-Deutsch, Deutsch-Maltesisch. M*alta.

54. Brigitte Ohk (1999) *Dizzjunarju Malti-Ġermaniż, Wörterbuch Maltesisch-Deutsch*. Malta.

55. Pawlu Bugeja (1999) *Kelmet il-Malti: dizzjunarju Malti-Ingliż Ingliż-Malti,* Owen Bonnici & Norbert Bugeja (editors) Malta.

56. Mario Serracino-Inglott (2000) *Id-Dizzjunarju Malti u teżawru ta' Malti mħaddem.* Malta.

57. Mario Serracino-Inglott (2003) *Id-Dizzjunarju Malti u teżawru ta' Malti mħaddem,* second edition. Malta.

58. Manfred Moser, Sharon Meyer, Joe Felice-Pace (2005) *Malti-Ġermaniż: dizzjunarju kbir, Deutsch-Maltesisch: Grosses Wörterbuch.* Germany.

59. Svetislav L. Stojanovi (2005) *Srpcko-Malteski recnik, Dizzjunarju Malti-Serb.* Malta.

60. Saviour Magro (2005) *Vokabularju Tematiku Malti-Franċiż.* Malta.

61. Joseph Aquilina (2006) *Concise Maltese-English English-Maltese Dictionary*. Malta.

62. Mario Serracino-Inglott (2011) *Id-Dizzjunarju Malti u teżawru ta' Malti mħaddem,* third edition. Steve Mifsud (editor). Malta.

63. Paul Bugeja (2013) *Kelmet il-Malti: Dizzjunarju Malti-Ingliż, Ingliż-Malti,* sixth revised edition. Keith Attard (editor). Malta.

64. Charles Briffa (2015) *The English-Maltese Dictionary for the 21st Century.* Malta.

65. Mario Serracino-Inglott (2016) *Id-Dizzjunarju Malti u teżawru ta' Malti mħaddem*, fourth edition. Dwayne Ellul (editor). Malta.

66. Paul Bugeja (2017) *Kelmet il-Malti: Dizzjunarju Malti-Ingliż, Ingliż-Malti,* seventh revised edition. Keith Attard (editor). Malta.

3.2 Children and Picture Dictionaries

67. Ġorġ Mifsud-Chircop (1982) *Dizzjunarju bl-istampi għaż-żgħar*. Malta.

68. Victor Fenech (1982) *Dizzjunarju taż-żgħar.* Malta.

69. Victor Fenech (1982) *Dizzjunarju taż-żgħar: l-ewwel ktieb ta' kliem bl-istampi.* Malta.

70. Victor Fenech (1983) *Dizzjunarju taż-żgħar: it-tieni ktieb ta' kliem bl-istampi.* Malta.

71. Victor Fenech (1985) *Dizzjunarju taż-żgħar: l-ewwel ktieb ta' kliem bl-istampi,* second edition. Malta.

72. Victor Fenech (1985) *Dizzjunarju taż-żgħar: it-tielet ktieb: il-kliem f'sentenzi.* Malta.

73. Ġorġ Mifsud-Chircop (1987) *Dizzjunarju bl-istampi għaż-żgħar: l-ewwel dizzjunarju għat-tfal li għadhom kemm bdew il-Malti.* Malta.

74. Victor Fenech (1987) *Dizzjunarju taż-żgħar: it-tieni ktieb ta' kliem bl-istampi,* second edition. Malta.

75. Victor Fenech (1990) *Dizzjunarju taż-żgħar: l-ewwel ktieb ta' kliem bl-istampi*, third edition. Malta.

76. Ġorġ Mifsud-Chircop (1994) *Dizzjunarju bl-istampi għaż-żgħar.* Malta.

77. Victor Fenech (1997) *Dizzjunarju taż-żgħar.* Malta.

78. Pawlu Mizzi (1997) *L-Ewwel Vokabularju bl-Istampi: Ingliż, Taljan, Malti.* Malta.

79. (no author) (2005) *L-Ewwel Dizzjunarju Tiegħi: Malti-Ingliż, Ingliż-Malti.* Malta.

80. Albert Gatt (2011) *Dizzjunarju għall-Primarja.* Malta.

81. Yana Psaila (2013) *Frażarju Russu-Malti / Русско-мальтийский словарь-разговорник. Malta.*

82. (no author) (2014) *Dizzjunarju bil-Malti għat-tfal komplut bl-istampi u bil-kulur.* Malta.

83. Carmel Tabone (2014) *Dizzjunarju tal-Malti għall-iskejjel.* Malta.

84. Yana Psaila (2014) *Frażarju Malti-Russu (bil-Malti u bir-Russu)*. Malta.

85. Trevor Żahra (2015) *Stampakelma*. Malta.

3.3 Specialised Dictionaries

86. Fortunato Panzavecchia (1850) *Dizionario Etimologico Maltese ed Italiano, Arabo, Ebraico, Latino, Inglese, Francese, Greco, con una breve prefazione storica sulle varie vicende, e sul necessario incoraggimenti degli studi della lingua Maltese.* Malta.

87. Guido Lanfranco (1987-1995) *Dizzjunarju ta' annimali u ħxejjex madwarna: manuskritt preparatorju*. Published in periodical *Sagħtar*. Malta.

88. Robert Mifsud Bonnici (1960) *Dizzjunarju Bijo-Bibljografiku Nazzjonali.* Malta.

89. Dumink Degiorgio (1964) *Dizzjunarju Tekniku Malti-Ingliż.* Malta.

90. Joseph Aquilina (1972) *A Comparative Dictionary of Maltese Proverbs*. Malta.

91. Michael Agius (1978) *Ftit Sinonomi w Antonomi Maltin*. This is a manuscript of a pioneering Maltese monolingual thesaurus, housed in the Archives and Rare Books Collection at the University of Malta's Library (Ellul 2022).

92. Donat Spiteri (1996) *Dizzjunarju Bibliku*. Malta.

93. Oliver Friggieri (1986) *Dizzjunarju ta' Termini Letterarji*. Malta.

94. Marco De Benetti (1994) *Il-Ħolm Imfisser: l-ewwel dizzjunarju tal-ħolm bil-Malti.* Malta.

95. Michael Galea (1995) *Dizzjunarju ta' Kittieba Maltin u Għawdxin, il-bieraħ u llum.* Malta.

96. Oliver Friggieri (1996) *Dizzjunarju ta' Termini Letterarji,* second edition. Malta.

97. Albert W. Agius (2000) *It-Teżawru: Ġabra ta' Sinonimi u Tifsiriet li jixtiebhu Maltin,* Joe Felice Pace (editor). Malta.

98. Oliver Friggieri (2000) *Dizzjunarju ta' Termini Letterarji*, third edition. Malta.

99. Graziella Buhagiar & Mary Anne Portelli (2000) *Dizzjunarju ta' termini folkloristi·i għal studenti fil-livell sekondarju*, B.Ed. (Hons.) University of Malta.

100. Oliver Friggieri (2000) *Dizzjunarju ta' Termini letterarji: teorija, metrika, stilistika b'eżempji mil-letteratura Maltija, Ġorġ Mifsud-Chircop* (editor). Malta.

101. Mark Montebello (2001) *Il-ktieb tal-filosofija f'Malta: jew dizzjunarju en·iklopediku tal-ħsieb Malti.* Malta.

102. George Cassola (2002) *Dizzjunarju tal-kontabilità, Ingliż-Malti*. Malta.

103. Mario Cassar (2003) *The Surnames of the Maltese Islands: an etymological dictionary.* Malta.

104. Paul Calleja (2003) *Dizzjunarju tal-Letteratura Maltija.* Malta.

105. Charles Briffa (2007) *Dictionary for Specific Purposes, English-Maltese, Maltese-English*. Malta.

106. Michael Ellul (2009) *Maltese English Dictionary of Architecture and Building in Malta.* Malta.

107. Michael J. Schiavone (2009) *Dictionary of Maltese Biographies (2 volumes)*. Malta.

108. Albert W. Agius (2010) *It-Teżawru: Ġabra ta' Sinonimi u Antonimi*, second edition, Joe Felice Pace & Reno Fenech (editors). Malta.

109. Charles Briffa (2010) *Dizzjunarju tal-Kullana Kulturali.* Malta.

110. Keith Attard (2012) *Ballun Pinġut: Ġabra ta' espressjonijiet mil-logħba tal-futbol.* Malta.

111. Laurence Mizzi (2013) *Dizzjunarju ta' Malta fil-Gwerra.* Malta.

112. Keith Attard (2014) *Ballun Pinġut 2.* Malta.

113. Charles Briffa (2017) *A bilingual dictionary of translation*. Malta.

114. Carlston Grima (2017) *Kaxxa Infernali: Dizzjunarju tal-Logħob tan-Nar.* Malta.

115. Joseph Caruana (2019) *Dizzjunarju ta' termini nawti·i u tagħrif marittimu lokali.* Malta.

116. Dwayne Ellul (2020) *Kelma għal Kelma: Teżawru Tematiku*. Malta.

117. Lino Bruguglio (2023) *Dizzjunarju Spe·jalizzat: Ekonomija*, Michael Spagnol, Thomas Pace, Dwayne Ellul (editors). Malta.

118. Segretarjat Parlamentari għar-Riformi u l-Ugwaljanza, Human Rights Directorate (2023) *Dizzjunarju Queer: Terminoloġija u Definizzjonijiet*. Malta.

119. Edward Gilson (2026) *Dizzjunarju Speċjalizzat: Ġeografija*, Michael Spagnol, Thomas Pace, Dwayne Ellul (editors). Malta.

The above list shows that most major dictionaries are either bilingual or multilingual. There is only one fully fledged monolingual dictionary: Serracino-Inglott's *Il-Miklem Malti* [81], in ten volumes. There remains a notable lack of pronunciation dictionaries, thematic lexica, and fully developed etymological or electronic resources. This point is taken up again towards the end of this paper.

4 Methods and Traditions

In this section, we discuss methods, traditions and problems in Maltese lexicography, focusing particularly on lemmatisation, the structure of the dictionaries, and the orthography of loan words.

4.1 Lemmatisation

In Maltese, the citation form for nouns is the singular, e.g. *mejda* 'table.F', *siġġu* 'chair.M'. When the noun inflects for gender, the masculine form is used, e.g. *attur* 'actor'. The citation form for adjectives is masculine singular, e.g. *sabiħ* 'beautiful.M'. As regards verbs, Maltese has no infinitive. The citation form traditionally used in dictionaries and grammars is the third person singular masculine perfective, e.g. *kiteb* 'he wrote/to write', since it is morphologically unmarked. However, the default citation form commonly used to refer to verbal lexemes is the second person singular imperfective, e.g. *tikteb* 'you write/to write' (Mifsud, 1995: 116, Borg & Azzopardi-Alexander 1997: 155). When a verb lacks perfective forms, dictionaries tend to use hypothetical forms as **faħ* 'to smell sweet' and **xebah* 'to resemble'.

Note that in his *Dˇamma*, De Soldanis [4] lemmatises many verbs in the first-person singular imperfect, following a practice commonly found in Latin lexicography. Consequently, verbal entries are cited with the imperfect prefix *n-* or *en-*, but are alphabetised independently of this prefix, typically under the initial segment of the verbal base. For instance, *ħadem* 'to work' is entered as *n-achdem* 'I work' under A and *ħall* 'to untie' as *en-Hholl* 'I untie' under H.

4.2 Dictionary Macrostructure

Maltese dictionaries do not follow the common practice in Arabic dictionaries of entering all lemmas under a basic consonantal root, generally of the verbal base, while listing all derivatives alphabetically under the main entry. Rather, they sort lemmas alphabetically. This method disrupts the morphological and semantic relationships among words that share a common root. For instance, *ħadem* 'to work', *maħdum* 'worked', *nħadem* 'to be worked, passive', all derivatives of $\sqrt{ħdm}$, are listed under the letter ħ, m, and n respectively.

Most probably, Maltese lexicographers avoid organising their dictionaries by root because of the dual lexical and morphological structure discussed above. Words of Romance or English

TABLE 2 Morphological derivations as organised in Aquilina's dictionary

Origin	Root	Form	Gloss
Arabic	√qlb	*qalb* *qalba* *qalbien/i* *qlubi* *qlubija*	'heart' 'core' 'brave' 'valiant' 'bravery'
Romance	√stud	*studju* *studjuż* *student* *studentat* *studja*	'study' 'studious' 'student' 'studentship' 'to study'
English	√sprej	*sprej* *sprejer* *sprejjatura* *sprejja*	'spray' 'sprayer' 'a spraying' 'to spray'

origin cannot be entered under a main consonantal root unless they have been fully integrated into Maltese root-and-pattern morphology, and can thus be decomposed into a consonantal root and a verbal pattern, such as √ptr and the verbal pattern 1v22v3, yielding *pitter* 'to paint', from Sicilian *pitturi* 'painter' (Mifsud, 1995).

However, two dictionaries follow the Arabic system of entry by root. The first is Antonio Emanuele Caruana's (1903) [20], whose dictionary represents an attempt to "purify" the language of non-Semitic elements. Except for a few Romance entries, he includes only words of Semitic origin. It is therefore unsurprising that he deviates from the alphabetical arrangement adopted by earlier lexicographers. The second is Joseph Aquilina's (1987–1990) bilingual dictionary [43], which presents an interesting compromise between concatenative word formation and root-and-pattern morphology. All items are grouped under a basic entry, as in Arabic lexicography. However, for ease of reference, each lemma also appears as a subentry in alphabetical order with a basic gloss. The user is then addressed to the main entry for full semantic and etymological treatment (Aquilina, 1980; 1985a; 1985b). Table 2 shows an example of three entries from Aquilina's dictionary (with basic glosses) illustrating different etymologies and morphological structures.

4.3 Orthography of Loanwords

A longstanding issue in Maltese lexicography concerns the orthography of English loanwords, which constitute the majority of recent borrowings. Although Aquilina's dictionaries remain the most authoritative lexicographic works for Maltese, their treatment of loanword spelling no longer fully reflects current orthographic policy. Aquilina favoured a largely phonetic approach, resulting in forms such as *ajsberg* 'iceberg', *ħajdrolik* 'hydraulic' and *rawndabawt* 'roundabout', while acknowledging resistance from speakers accustomed to English spelling (Aquilina, 1980).

More recently, the Kunsill Nazzjonali tal-Ilsien Malti adopted a context-sensitive policy in *Deċiżjonijiet 2* (2018), distinguishing between fully integrated loanwords, which take Maltese spelling and morphology (e.g. *xokk* 'shock', *garaxx* 'garage'), non-integrated or compound forms, which retain English spelling (e.g. *pancake*, *soft toy*), and partially integrated forms,

where both spellings may coexist (e.g. *ħelikopter* or *helicopter*). The policy also states that English forms should not be italicised, except in beginner-level educational contexts. From a lexicographic perspective, this policy suggests that dictionaries may need to include both the Maltese-adapted and the original English form for certain entries, particularly those in the third category, thereby supporting users regardless of which orthographic form they encounter in use.

4.4 Trends in Maltese Lexicography

Across four centuries, Maltese lexicography displays a shifting typological tendency in terms of multilingual scope, coverage, audience and scale. The earliest works, from the 17th and 18th centuries, are predominantly manuscript-based and typically involve bilingual or multilingual configurations reflecting the historical context in which Maltese was mediated through Italian and Latin. These works often display uneven coverage and, in some cases, include extensive encyclopaedic information. In the 19th century, with the introduction of English following British rule, Maltese lexicography sees the emergence of printed bi- and trilingual dictionaries (Maltese-Italian-English or English-Maltese), many of which are influenced by earlier models, particularly Vassalli's work. These dictionaries tend to reflect practical needs in clerical, educational and colonial contexts, although they vary considerably in scope, structure, and degree of systematisation.

From the mid-20th century onwards, there is a marked expansion in both scale and functional range, with large bilingual works (notably Maltese-English) coexisting alongside the first monolingual dictionaries, culminating in comprehensive projects such as *Il-Miklem Malti* [39] and *Maltese-English Dictionary* [43], which offer extensive lexical, grammatical and etymological coverage. This period also witnesses increasing diversification in lexicographic output including works addressing specialised domains (technical, literary, terminological dictionaries) and different target audiences such as learners and children.

In the 21st century, the typological range further expands to include concise, domain-specific, pedagogical and digital-oriented resources, with a general shift towards monolingual and user-centred design, larger and more systematically structured macro- and microstructures, and clearer differentiation of intended users, from specialists to schoolchildren and second-language learners.

5 Digital Resources

A META-NET report on Maltese (Rosner, 2012) notes that, despite the existence of several linguistic resources, Maltese remains digitally under-represented when compared to other European languages. One of the earliest large-scale initiatives addressing this gap was the *Maltilex* project, first proposed by Rosner, Caruana and Fabri (1998) as a computational lexicon intended to provide the infrastructural basis for Maltese NLP applications, combining dictionary-based and corpus-based approaches to lexicon building (Dalli, 2002). Although projects such as *Korpus Malti* have since expanded dramatically – reaching around 470 million tokens in version 4.2 (2024) – the rich hybrid morphology of Maltese continues to make automatic processes such as POS tagging, lemmatisation and lexicon extraction particularly challenging for reliable computational lexicon-building.

5.1 *Ġabra: An Online Collection of Lexical Resources*

The online resource *Ġabra* (Camilleri, 2013) is the most significant digital lexicon project to date, maintained by the University of Malta. It was originally built as a platform for collecting different lexical resources compiled by various authors into a single searchable database with a common structure. The initial resources used for bootstrapping the database included:

1. An exhaustive enumeration of root-and-pattern verbs (Spagnol, 2011)
2. A corpus of broken plurals (Mayer et al., 2013; Schembri, 2012)
3. A list of verbal nouns from the Aquilina dictionary and others (Ellul, 2016)
4. A basic online English-Maltese dictionary (Falzon, 1997; META-SHARE, 2012)
5. A small lexicon built for the Apertium machine translation project (Forcada et al., 2011).

5.1.1 Structure

Because of the heterogeneous nature of these initial resources, *Ġabra* was designed to be structurally flexible. The primary entity in the database is the **lexeme**, corresponding to an entry in a dictionary, consisting minimally of a lemma and part of speech. Most also include an English gloss and root information, while many others also include alternative spellings, derivational form, transitivity etc. The other important entity in the database is the **word form** (linked to a parent lexeme). A word form minimally consists of a surface form, but often includes morphological features for number, gender, and tense, polarity for verbs. The lack of fixed schema means that new fields can be added without having to make any structural changes to the database.

As mentioned earlier, Aquilina includes derived forms as sub-entries of other entries, such as *libbes* 'to dress someone' being found as a sub-entry of *libes* 'to wear'. He then uses stubs to help the reader, such that anyone looking up *libbes* will be directed to the corresponding main entry. In a digital setting, the use of links for building relations between entries gives more flexibility. Thus, in *Ġabra* there is no "main" entry with a rigorous semantic treatment and stub entries which simply refer to it. Rather, all derived forms are stored as separate entries but are linked by root information. So, *libes* and *libbes* appear as separate lexemes, but they both have the same root $\sqrt{\text{lbs}}$ and thus are implicitly linked together. The user interface then takes care of making the entries reachable from one another.

5.1.2 Access and Interoperability

The *Ġabra* project was started with openness in mind. The resource is freely available online and can be used without any registration. Its lexical contents are licensed under a Creative Commons Attribution 4.0 International License, making it free to use for any purpose (with attribution). Source code is open and shared on GitHub[2] under a permissive MIT license.

The lexicon is accessible programmatically via an API which allows other services and apps to access the data directly. The database schema is fully documented and its contents can be downloaded for offline use. While its structure loosely conforms with LMF, strict adherence

[2] https://github.com/mlrs

to the standard was not prioritised by the authors. Export to the TEI Lex-0 format is similarly not possible, and the development of *Ġabra* in fact predates this specification.

5.1.3 Coverage and Accuracy

The most distinguishing feature of *Ġabra* is that it is a full-form lexicon. Using a combination of morphological generation (Camilleri, 2013; Forcada et al., 2011) and manual annotation, the database has been populated with millions of inflectional word forms, each linked to their corresponding lexeme. *Ġabra* currently contains over 20,000 lexeme entries and over 4.5 million word forms.

By comparison, Serracino-Inglott's Maltese dictionary [39] contains over 34,000 entries (Ellul, 2024), while the Maltese-English volumes of Aquilina's dictionary [43] contain some 41,016 (Brincat, 2011, p. 407). Dalli (2002) estimates that at least 30,000 lemmas must be identified in order to have significant coverage of Maltese. By considering the worst-case inflectional forms of each of the major parts of speech, together with the distribution of the different parts of speech in the Aquilina dictionary, it is estimated that a full-form Maltese lexicon could contain up to roughly 86 million forms (Camilleri, 2013). Despite this rather large number, storing this number of forms is entirely possible with today's technologies. Storing full word forms as opposed to using automatic morphological generation is simpler to manage, allows full control of the data, makes adding exceptions or irregular cases easy, and directly supports pattern-based searching and tasks such as automatic lemmatisation.

Automatically generating word forms can of course introduce errors, and it is openly stated that entries in *Ġabra* may contain mistakes. However, unlike print dictionaries which are expensive to publish and impossible to update, publishing and updating digital resources is quick and can be done continuously. Since its release in 2013, the *Ġabra* content has received various updates in the form of systematic correctness checks by researchers at the University of Malta, as well as many user-submitted suggestions for new entries. This work has however not been continuous, and to date *Ġabra's* goal of becoming a fully-fledged highly accurate dictionary for Maltese has not been realised.

5.2 Other Digital Resources

Over the past decade, several digital lexicographic and terminological initiatives have contributed to the online presence of Maltese. In 2016, *MalteseDictionary.org.mt* was launched using *Ġabra* as its backend, although the project remained limited in scope and has since been discontinued. Around the same period, work also began on *Minsel*, an online etymological dictionary for Maltese, which reached approximately 160 entries before activity ceased (Gatt, 2020). Other online linguistic resources include *Ġabra tal-Malti Qadim*, focusing on older Maltese texts, and the *Dizzjunarju Online tal-Lingwa tas-Sinjali* [3], dedicated to Maltese Sign Language. In addition, a number of terminology and language resources are available online, including the Department of Maltese's *Termini* [4] (Ellul, 2024), *ilMiklem.com* [5], and the Maltese section of IATE (EU's termbase).

[3] Both resources are available at https://mlrs.research.um.edu.mt/resources
[4] https://malti.mt/termini/glossarji
[5] https://ilmiklem.com

5.3 Future Projects

The Department of Maltese at the University of Malta is currently developing *Dizzjunarju.mt,* a function-based, polyfunctional Maltese-English e-dictionary intended as a national, authoritative online reference resource. The dictionary integrates Maltese definitions with English equivalents, pronunciation guidance, illustrative examples and etymological information. Methodologically, it adopts a two-tiered approach in which Lexical Computing's Dictionary Express platform employs artificial intelligence, natural language processing and corpus linguistics to generate draft entries based on authentic language use, which are then systematically reviewed and refined by researchers to ensure accuracy, clarity and cultural appropriateness.

6 Conclusion

In compiling this overview, we identified a number of important gaps in Maltese lexicography, including the lack of updated comprehensive monolingual and bilingual dictionaries, detailed etymological dictionaries, collocational and frequency dictionaries, thesauri, thematic LGP dictionaries, and broader LSP resources covering specialised domains. We view future work in Maltese lexicography as twofold: both the development and active maintenance of modern digital lexical resources as well as the digitisation of older manuscripts and print dictionaries, especially those that remain unpublished or difficult to access.

References

Agius De Soldanis, G. P. F. (2016). *Ġamma tal-Kliem Kartaginis Mscerred fel Fom tal Maltin u Ghaucin*. Il-Kunsill Nazzjonali tal-Ilsien Malti, il-Librerija Nazzjonali, id-Dipartiment tal-Malti, l-Akkademja tal-Malti, Heritage Malta, it-trust tad-Dipartiment tal-Innovazzjoni fir-Ri·erka u l-Iżvilupp tal-Università ta' Malta, BPC International u l-Ministeru ta' Għawdex.

Agius, D. (1996). *Siculo-Arabic*. Kegan Paul International.

Agius, D. (1990). Al-Miklem Malti: A Contribution to Arabic Lexical Dialectology. *British Society for Middle Eastern Studies Bulletin,* 17(2), 171-180.

Aquilina, J. (1958). Maltese as a mixed language. *Journal of Semitic Studies* 3(1), 58-79.

Aquilina, J. (1961). Maltese as a mixed language. *Papers in Maltese Linguistics*, 42-62. Royal University of Malta.

Aquilina, J. (1979). *Maltese Arabic Comparative Grammar*. University of Malta.

Aquilina, J. (1980). The problems of Maltese lexicography. In R.Y. Ebied & M.J.L Young (Eds.), *Oriental Studies* (pp. 159-171). Brill.

Aquilina, J. (1985a). Maltese Lexicography. *Orbis - Bulletin International de Documentation Linguistique,* II(2), 474-483.

Aquilina, J. (1985b). The problems of a bilingual dictionary. *Melanges Skok, 19-26*. Jugoslavenska Akademija Znanosti i Umjetnosti.

Borg, A. & Mifsud, M. (2002). Maltese Object Marking in a Mediterranean Context. In Paolo Ramat & Thomas Stolz (Eds.), *Mediterranean Languages* (pp. 32-46). Universitätsverlag Dr. N. Brockmeyer.

Borg, A. (1994). Language. In H. Frendo & O. Friggieri (Eds.), *Malta: Culture and Identity* (pp. 27-50). Ministry of Youth and the Arts.

Borg, A. (1996). On some Levantine linguistic traits in Maltese. *Journal of Oriental Studies. Studies in Modern Semitic Languages,* 16, 133-152.

Brincat, J. (1998). A Malta l'italiano lo insegna la televisione. *Italiano e Oltre,* XIII, 52-58.

Brincat, J. (1999). Languages Across Frontiers. The Acquisition of Italian in Malta by Viewers of Italian TV programmes. In B. Caron (Ed.), *Proceedings of the XVIth International Congress of Linguists*. Elsevier Sciences.

Brincat, J. (2000). *Il-Malti: Elf Sena ta' Storja*. PIN.

Brincat, J. (2004). *Malta. Una Storia Linguistica*. Le Mani.

Brincat, J. (2008). Malta. In K. Versteegh (Ed.), *Encyclopedia of Arabic Language and Linguistics*, III, 141-145.

Brincat, J. (2011). *Maltese and other languages. A linguistic history of Malta*. MidSea Books Ltd.

Brincat, J. (2005). *Maltese — An Unusual Formula*. MED Magazine, 27. http://macmillandictionaries.com/MED-Magazine/February2005/27-LI-Maltese.htm

Brincat, J. (2015). Aquilina's abbandinxipp. *Times of Malta,* December 8, 2015. http://www.timesofmalta.com/articles/view/20151208/opinion/Aquilina-s-abbandinxipp.594986

Cachia, L. (1994). *L-Ilsien Malti, ilbierah u llum*. Sensiela Kotba So·jalisti.

Camilleri, J. J. (2013). *A Computational Grammar and Lexicon for Maltese* [Master's Thesis, Chalmers University of Technology].

Camilleri, J. J. & Spagnol, M. (2013). An online database of root-and-pattern verbs. In *4th International Conference on Maltese Linguistics* (Lingwistika 2013).

Caruana, S. (2006). Trilingualism in Malta. Maltese, English and italiano televisivo. *International Journal of Multilingualism,* 3(3), 159-172.

Cassar Pullicino, J. (1947). Unpublished Maltese Dictionaries. *Scientia*, XIII, 3.

Cassola, A. (1992). *The Biblioteca Vallicelliana: Regole Per La Lingua Maltese*. Said International.

Comrie, B. & Spagnol, M. (2016). Maltese loanword typology. In G. Puech, & B. Saade (Eds.), *Shifts and patterns in Maltese* (pp. 315-330). De Gruyter Mouton.

Cremona, A. (1933). Lexicons and Vocabularies. *Journal of Malta University Literary Society,* 6, 257-260.

Dalli, A. (2002). *Computational Lexicon for Maltese* [Master's Thesis, University of Malta].

Ellul, D. (2022). Is-sisien ta' *Dizzjunarju tal-lsien Malti* ta' Michael Agius u aspetti mill-makrostruttura u l-mikrostruttura tiegħu. *Leħen il-Malti*, 41(91), 166-185.

Ellul, D. (2024). Maltese lexicography: a historical context and the current state. In K. S. Despot, A. Ostroški Anić & I. Brač (Eds.), *Lexicography and semantics: proceedings of the XXI EURALEX International Congress* (pp. 389-404).

Ellul, L. (2016). Verbal nouns in Maltese [Master's Thesis, University of Malta].

Farrugia, G. (1998). Il-Malti lsien imħallat. In K. Borg (Ed.), *Lingwa u Lingwistika* (pp. 227-291). Klabb Kotba Maltin.

Fenech, E. (1978). *Contemporary Journalistic Maltese: an Analytical and Comparative Study*. Brill.

Forcada, M. L., Ginestí-Rosell, M., Nordfalk, J., O'Regan, J., Ortiz-Rojas, S., Antonio Pérez-Ortiz, J., Sánchez-Martínez, F., Ramírez-Sánchez, G. & M. Tyers, F. (2011). Apertium: A free/open-source platform for rule-based machine translation. *Machine Translation* 25(2), 127-144.

Gatt, A. & Borg, C. (2011). *Using the MLRS Corpus. Technical report*. Institute of Linguistics, University of Malta.

Gatt, K. (2020). Minsel. An etymological database for lexemes of Arabic origin in Maltese. In S. Čéplö & J. Drobný (Eds.), *Maltese linguistics on the Danube* (pp. 111-117). De Gruyter Mouton.

Kontzi, R. (1981). L'elemento maltese nel maltese. *Journal of Maltese Studies,* 14, 37-42.

Kunsill Nazzjonali tal-Ilsien Malti. (2008). *Innaqqsu l-In·ertezzi 2: Il-Kliem ta' Nisel Ingliż fil-Malti*. Kunsill tal-Malti.

Mayer, T., Spagnol, M. & Schönhuber, F. (2013). Fixing the broken plural in Maltese. In A. Borg, S. Caruana & A. Vella (Eds.), *Perspectives on Maltese Linguistics* (pp. 129–158). Akademie Verlag.

META-SHARE (2012). Language resources for Maltese. Retrieved from http://metashare.metanet4u.eu/repository/search/?selected_facets=languageNameFilter_exact%3AMaltese

Mifsud, M. (2008). Maltese. In K. Versteegh (Ed.), *Encyclopedia of Arabic Language and Linguistics*, III (pp. 146-159). Brill.

Pace, T. (2012). Minding our Language. *Times of Malta*, July 29, 2012. http://www.timesofmalta.com/articles/view/20120729/technology/minding-our-language.430960

Privitera, V. (1964). *Id-Dizzjunarji Maltin* [Bachelor's thesis, University of Malta].

Rosner, M. (2009). Electronic language resources for Maltese. In B. Comrie, R. Fabri, E. Hume, M. Mifsud, T. Stolz & M. Vanhove (Eds.), *Introducing Maltese linguistics: selected papers from the 1st International Conference on Maltese Linguistics* (pp. 251–278).

Rosner, M., Caruana, J. & Fabri, R. (1998). Maltilex: A computational lexicon for Maltese. In M. Rosner (Ed.), *Proceedings of Workshop on Computational Approaches to Semitic Languages*. COLING-ACL98. Association for Computational Linguistics.

Rosner, M., Fabri, R., Attard, D.P. & Gatt, A. (2006). MLRS, a Resource Server for the Maltese Language. In J. Abela, A. Dalli & K. Guillaumier (Eds.), *Proceedings of the 4th Computer Science Annual Workshop (CSAW-2006)* (pp. 90–98). University of Malta.

Rosner, M., Fabri, R., Caruana, J., Loughraïeb, M., Montebello, M., Galea, D. & Mangion, G. (1999). Linguistic and Computational Aspects of Maltilex. In *Proceedings of ATLAS the Arabic Translation and Localization Symposium* (pp. 2-10).

Rosner, M. & Joachimsen, J. (2012). *The Maltese Language in the Digital Age*. META-NET White Paper series. Springer.

Saydon, P.P. (1953). Bibliographical Aids to the Study of Maltese. *Journal of Near Eastern Studies*, April, 1-30.

Serracino-Inglott, E. (1947). Kommenti fuq Dizzjunarji Malti. *Il-Malti,* 23, 38-45.

Schembri, T. (2012). The Broken Plural in Maltese: A Description. *Il-Lingwa Tagħna,* 3.

Spagnol, M. (2006). *Muftieħ għall-Vokabolarju Malti* [Bachelor's thesis, University of Malta].

Spagnol, M. (2011). *A Tale of Two Morphologies: Verb structure and argument alternations in Maltese* [Doctoral dissertation, University of Konstanz].

Stolz, T. (2003). Not quite the right mixture: Chamorro and Malti as candidates for the status of mixed language. In Y. Matras & P. Bakker (Eds.), *The Mixed Languages Debate. Theoretical and Empirical Advances* (pp. 71-315). Mouton de Gruyter.

Contact information

Michael Spagnol
L-Università ta' Malta
michael.spagnol@um.edu.mt

John J. Camilleri
Chalmers University of Technology and University of Gothenburg, Sweden
john.j.camilleri@chalmers.se

Dwayne Ellul
L-Università ta' Malta
dwayne.ellul@um.edu.mt

Tarrin Wills & Ellert Thor Johannsson

CAN AI-ASSISTED NORMALISATION IMPROVE ACCESS TO HISTORICAL LEXICOGRAPHIC DATA?

Abstract The topic of this article is AI-assisted normalisation and how it can be used in the context of a historical dictionary. We look at data from *A Dictionary of Old Norse Prose*, which is a lexicographic resource on the language of medieval Iceland and Norway. This dictionary has a long tradition of using non-normalised, diplomatic text editions as source for its lexicographic descriptions. As a result, the bulk of the citation material remains difficult to access for beginners and non-experts in the language. We tested several LLMs in order to assess whether it would be feasible to provide users with automatic normalisations. The article describes several phases in testing and comparing different LLMs and discussing the types of errors they made. The result show that from the initial experimental phase in 2025 the models have improved significantly and currently are able to provide fairly error-free normalisations. We also mention the option of using modern Icelandic as an intermediate stage in the normalisation process which would facilitate the use of NLP-tools. The conclusion is that LLM-assisted normalisation can support broader access to historical lexicographic data. Such tools combine automated processing with human oversight, allowing historical dictionaries to expand their usability while maintaining scholarly rigour.

Keywords Historical lexicography; language normalisation; large language models

1 Introduction

Historical dictionaries are typically based on citations from older texts that reflect non-standardised and often highly variable orthographic practices. For many lexicographic projects, this fidelity to historical spelling is essential, as it preserves important evidence about language use and textual transmission. At the same time, unnormalised citation material presents a significant barrier to dictionary use for many end-users, including students, scholars from adjacent disciplines, and non-specialists. It also limits the re-use of historical dictionary data in corpus-linguistic and computational workflows.

This paper explores whether recent developments in large language models (LLMs) can offer viable automated or semi-automated solutions for normalising historical dictionary citations while preserving their scholarly value. The study is framed as a pilot investigation into both the feasibility and the methodological implications of such an approach.

The case study is *A Dictionary of Old Norse Prose* (ONP), a comprehensive historical dictionary documenting the medieval prose language of Iceland and Norway. ONP is characterised by a high level of detail and extensive citation material drawn from diplomatic editions and, in many cases, directly from medieval manuscripts. While this approach provides valuable information about scribal practice and textual variation, it presupposes considerable expertise in Old Norse orthography and editorial conventions. As a result, many users are unable to fully exploit the dictionary's citation material.

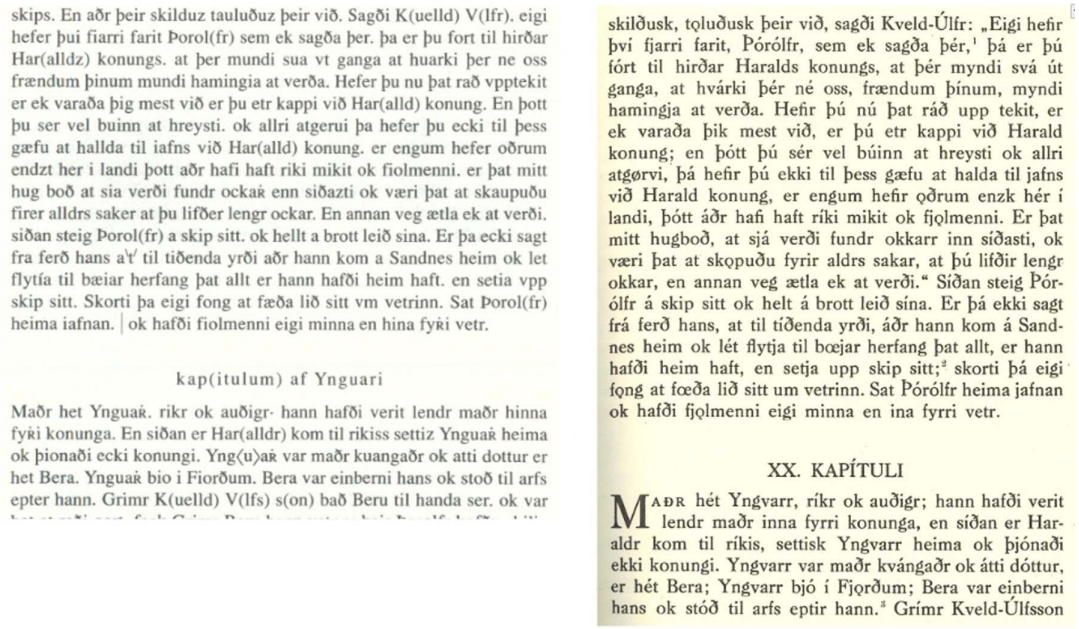
This article is structured in the following way. After this introduction we discuss normalisation in Old Norse dictionaries, how it has developed over time and what sets ONP apart from its predecessors. We then take a closer look at the use of LLMs and how they have been applied in relation to ancient languages and normalisation. We discuss the first phase of the current study which included an initial assessment of feasibility and took place in 2025. We then move on to

the second phase, which revolved around comparing different models and outputs. Finally, we account for the third and final phase, where we assess the performance of the latest models and compare it with previous efforts. The article concludes with a brief discussion and concluding remarks.

1.1 Normalisation in Old Norse Lexicography

Old Norse lexicography has a rich and long tradition going back to the works of Magnús Ólafsson (1650) and Verelius (1691). The 19th century saw a significant progress in scholarly work on the vocabulary of Old Norse as well as its lexicographic description with the publication of comprehensive dictionaries, such as the work of Cleasby-Vigfusson in 1874 and Fritzner in 1896. In these works a large part of the medieval lexis was presented in structured dictionary entries with numerous citations from various types of Old Norse texts, both original compositions as well as translations. At the same time the academic discipline of editing and publishing texts based on philological principles progressed rapidly with many new and improved text editions seeing the light of day.

In Old Norse studies, scholarly editions of texts are often produced using diplomatic form, which means that the orthography of the editions reflects many details of the scribal practices present in the original medieval manuscript or oldest text witness. This includes using several different symbols to signify the "same" letter if such was the practice of the original scribe. Another editorial approach to medieval texts is to publish them in a reconstructed "normalised" orthography, typically corresponding to early thirteenth-century Icelandic. This method is less loyal to the scribal eccentricities and results in the texts being rendered in a standardised spelling, making them more approachable and easier to read for non-experts. This latter method is widely used in editions such as those published by the Old Icelandic Text Society which has published most of the saga-literature in normalised orthography while still following other scholarly practices, such as textual commentaries and showing textual variation.



skips. En aðr þeir skilduz tauluðuz þeir við. Sagði K(uelld) V(lfr). eigi hefer þui fiarri farit Þorol(fr) sem ek sagða þer. þa er þu fort til hirðar Har(alldz) konungs. at þer mundi sua vt ganga at huarki þer ne oss frændum þinum mundi hamingia at verða. Hefer þu nu þat rað vpptekit er ek varaða þig mest við er þu etr kappi við Har(alld) konung. En þott þu ser vel buinn at hreysti. ok allri atgerui þa hefer þu ecki til þess gæfu at hallda til iafns við Har(alld) konung. er engum hefer oðrum endzt her i landi þott aðr hafi haft riki mikit ok fiolmenni. er þat mitt hug boð at sia verði fundr ockaꝛ enn siðazti ok væri þat at skaupuðu firer alldrs saker at þu lifðer lengr ockar. En annan veg ætla ek at verði. siðan steig Þorol(fr) a skip sitt. ok hellt a brott leið sina. Er þa ecki sagt fra ferð hans aꝛ̓ til tiðenda yrði aðr hann kom a Sandnes heim ok let flytía til bæiar herfang þat allt er hann hafði heim haft. en setia vpp skip sitt. Skorti þa eigi fong at fæða lið sitt vm vetrinn. Sat Þorol(fr) heima iafnan. | ok hafði fiolmenni eigi minna en hina fyꝛ̓i vetr.

kap(itulum) af Ynguari

Maðr het Ynguaꝛ̓. rikr ok auðigr· hann hafði verit lendr maðr hinna fyꝛ̓i konunga. En siðan er Har(alldr) kom til rikiss settiz Ynguaꝛ̓ heima ok þionaði ecki konungi. Yng⟨u⟩aꝛ̓ var maðr kuangaðr ok atti dottur er het Bera. Ynguaꝛ̓ bio i Fiorðum. Bera var einberni hans ok stoð til arfs epter hann. Grimr K(uelld) V(lfs) s(on) bað Beru til handa ser. ok var

skilðusk, tǫluðusk þeir við, sagði Kveld-Úlfr: „Eigi hefir því fjarri farit, Þórólfr, sem ek sagða þér,[1] þá er þú fórt til hirðar Haralds konungs, at þér myndi svá út ganga, at hvárki þér né oss, frændum þínum, myndi hamingja at verða. Hefir þú nú þat ráð upp tekit, er ek varaða þik mest við, er þú etr kappi við Harald konung; en þótt þú sér vel búinn at hreysti ok allri atgørvi, þá hefir þú ekki til þess gæfu at halda til jafns við Harald konung, er engum hefir øðrum enzk hér í landi, þótt áðr hafi haft ríki mikit ok fjǫlmenni. Er þat mitt hugboð, at sjá verði fundr okkarr inn síðasti, ok væri þat at skǫpuðu fyrir aldrs sakar, at þú lifðir lengr okkar, en annan veg ætla ek at verði.“ Síðan steig Þórólfr á skip sitt ok helt á brott leið sína. Er þá ekki sagt frá ferð hans, at til tíðenda yrði, áðr hann kom á Sandnes heim ok lét flytja til bœjar herfang þat allt, er hann hafði heim haft, en setja upp skip sitt;[2] skorti þá eigi fǫng at fœða lið sitt um vetrinn. Sat Þórólfr heima jafnan ok hafði fjǫlmenni eigi minna en ina fyrri vetr.

XX. KAPÍTULI

Maðr hét Yngvarr, ríkr ok auðigr; hann hafði verit lendr maðr inna fyrri konunga, en síðan er Haraldr kom til ríkis, settisk Yngvarr heima ok þjónaði ekki konungi. Yngvarr var maðr kvángaðr ok átti dóttur, er hét Bera; Yngvarr bjó í Fjǫrðum; Bera var einberni hans ok stóð til arfs eptir hann.[a] Grímr Kveld-Úlfsson

FIGURE 1 Examples of unnormalised (left, Einarsson, 2001, p. 27) and normalised (right, Nordal, 1933) editions of the same passage of *Egils saga*

The classic lexicographic works of the 19th century, most importantly the dictionaries of Cleasby-Vigfusson and Fritzner, prefer employing a normalised orthography in their lexical descriptions. This is true both when compiling the headword list used in the dictionary work as well as when citing medieval texts. As a result, the lexicographic data appears more accessible and streamlined and users do not have to be aware of spelling peculiarities found in some of the original sources or even know much about Old Norse scribal practices.

In 1939 *A Dictionary of Old Norse Prose* was established as a new lexicographic project based on different editorial practices from its predecessors. In this dictionary the lemma list was to be normalised in a similar manner as had been the practice of earlier lexicographic works, but the text material was to be based on philological principles and cited from the text edition most loyal to the original text. Since then, the implementation of these principles has meant that ONP consistently has prioritised diplomatic editions even sometimes preferring to cite original manuscripts rather than existing editions if those editions are not compatible with ONP's principles of philological integrity.

As ONP has progressed over the years, switching from producing printed volumes into publishing the dictionary material exclusively online, it remains committed to its philological foundation and editorial principles. This means that it prefers unnormalised versions of citation texts taken directly from diplomatic editions of the works it covers. There is a small amount of normalisation of individual characters (long s 'ſ' for s, for example) and silent expansion of abbreviations based on the editions. Still ONP's practice is a huge departure from its predecessor large-scale dictionaries (Cleasby and Vigfusson, 1874; Fritzner, 1886–1896). Those earlier dictionaries include around 100,000 citations, which is considerably less than ONP's 800,000. Therefore, the ONP dictionary contains a significant amount of unnormalised material that the users of the dictionary need to be able to read and process if they want to take full advantage of the information found in this resource.

Even though ONP lexicographic descriptions are largely based on citations from unnormalised scholarly editions the dictionary developed its own normalised orthography and standards for the list of headwords. This standard is slightly different from the main orthographic standards used in Old Norse text publications and dictionaries as is evident in Table 1 (based on Bernharðsson et al., 2019).

TABLE 1 Normalisation in dictionaries and text editions of Old Norse. The first column gives examples of original spellings from ONP citations and diplomatic editions; ÍF = Íslenzk fornrit (Old Icelandic Text Society); ONP = A Dictionary of Old Norse Prose; Baetke = Baetke (1965-1968); NO = Heggstad et al. (1975); GNO = Gammelnorsk Ordboksverk; Fritzner = Fritzner (1886-1896); Mod Ice = Modern Icelandic. This table was originally set up by Robert K. Paulsen and is taken here from Bernharðsson et al. (2019) with supplementary examples of original orthography and English translations

Orig. orthography	ÍF	ONP	Baetke	NO	GNO	Fritzner	Mod Ice	Transl.
ręð-az, røðaz, ręðaz	hræðask	hræðask	hræðask	hræðast	ræðast	hræðast	hræðast	'be afraid'
søkia, sækia, sœkia	sœkja	søkja	sœkja	søkja	sœkia	sœkja	sækja	'seek'
iorð, jaurd, iorþ	jǫrð	jǫrð	jǫrð	jǫrð	iorð	jörð	jörð	'earth'
eydzla, eyzla	eyzla	eyzla	eyðsla	eyðsla	øyðsla	eyzla	eyðsla	'excess'

TABLE 1 (*CONTINUED*)

Orig. orthography	ÍF	ONP	Baetke	NO	GNO	Fritzner	Mod Ice	Transl.
giæzla, gæzla, gætzla	gæzla	gǽzla	gæzla	gæzla	gæzla	gæzla	gæsla	'care'
snilld, snild, stnilld	snilld	snilld	snild	snild	snild	snild	snilld	'genius'
talddu, talldu, tolþo	tǫlðu	tǫlðu	tǫldu	tǫldu	toldu	töldu	töldu	'told' pl.
sæmð, sømd, sęmþ	sœmð	sømð	sœmd	sømd	sœmd	sœmd	sæmd	'honour'
olog, vlǫg, o log	ólǫg	ólǫg	ólǫg	úlǫg	úlog	úlög	ólög	'injustice'
hiolp, hialp	hjálp	hjǫlp	hjálp	hjǫlp	hiolp	hjölp	hjálp	'help'

ONP is still a work in progress. Over the years this dictionary has gone through significant changes from being a straight-forward collection of citation slips to its current form as an all-sided online lexicographic resource. Until now though normalised orthography has been limited to the list of lemmas. As technology progresses and the lexical information is no longer constrained by printed pages, limited computer capabilities and fragile digital storage, it would make sense to consider expanding the use of normalisation beyond the lemma list and on to the citation material itself. From a lexicographic perspective, normalisation of citation text is best understood not as a replacement for diplomatic citation, but as an additional access layer serving different user needs. There are many scenarios where a standardised version of citation material is desirable, particularly for beginning students, interested laypeople and for computational use cases.

Manual normalisation requires expert knowledge and is time-consuming, making it impractical for large-scale lexicographic resources with limited staffing. This motivates the exploration of automated and semi-automated approaches for normalisation. As a result it seems worthwhile for a lexicographic resource such as ONP to consider expanding the use of normalised Old Norse beyond headwords and providing its users with alternative means to interact with the dictionary material and thereby expanding its user base and practical usage. With the rapid development of LLMs such expansion has become more attractive and seems like a logical next step in the evolution of ONP.

1.2 Use of LLMs for processing ancient languages

Pre-LLM methods for normalisation involve a number of different machine learning techniques and a number of tools are surveyed and tested by Bollmann (2019). The vast majority of these approaches yield less than 90% accuracy on unseen texts, which is considerably below the standard required by a project like ONP. Even results around 95% accuracy, which certain methods can sometimes produce, are insufficient to produce texts that can be published without close to the same amount of intervention as manually normalising from scratch.

The use of machine learning in early languages is surveyed by Thea Sommerschield and colleagues (2023), who catalogue more than 240 studies applying machine learning to ancient texts across a wide range of languages and scripts. The authors propose a taxonomy of AI applications including digitisation, restoration, linguistic analysis, textual criticism, translation, and

decipherment. They argue that AI methods are transforming the study of ancient texts in ways comparable to the impact of scientific instruments in the natural sciences, and these show considerable promise with respect to Latin and Greek (e.g. Pavlopoulos, 2025).

Language normalisation as a specific goal using machine learning, however, has only been applied in limited settings. There are various reasons for this: some languages do not display significant orthographic variation (e.g. Latin) or this variation is normally removed in digital source materials; for other languages where there is significant orthographic variation, such as Middle English, there is no tradition for normalisation and NLP tools therefore are trained on unnormalised materials. An exception is Early Modern English, where texts from around the time of Shakespeare are routinely normalised for both digital and print editions.

The present authors have not found a published application of generalised LLMs to the specific normalisation of premodern languages according to established conventions. It is presumed that the use-cases are generally limited: normalisation requires the LLM to have an understanding of the syntax, lexicon and semantics of the language of the text to be normalised; if this is the case, normalisation functions as an internal intermediate stage of language processing that is normally superseded by the resulting output which translates or otherwise analyses the input text.

ONP's citations are taken from editions that were produced over a period of almost two hundred years and with highly variable practices of reproduction. Any automatic process needs to take into consideration this variability of underlying practices. The citation texts for ONP should be suitable for processing: although they are generally short, they represent a minimal unit of contextual meaning that gives an understanding of the given word in context, as selected by editors. The following processes therefore use the citation texts as the focus of testing normalisation of Old Norse.

2 Phase 1: AI Normalisation Feasibility (April 2025)

2.1. Method

The first phase of the study tested the feasibility of using an LLM to normalise ONP citations written in non-standard orthography. Citations were extracted from the ONP database, stripped of markup, and submitted to the ChatGPT API using GPT-4o.

The system prompt was:

> You are an expert in Old Norse language and grammar. You normalise Old Norse texts to their standard forms, preserving punctuation and capitalisation where appropriate.

The user prompt included five example pairs of unnormalised citations ('few-shot' prompting) and their manually normalised equivalents, followed by a new citation for processing. The resulting normalisations were recorded along with timing and model metadata. The results can be seen at https://onp.ku.dk/gefin/q.php?p=onp/info/ainorm.

2.2 Results

This process was slow, with each query taking 3-5s, and resource-intensive, using up the credits of the basic ChatGPT plan after fewer than 2000 citations. Some examples are in Table 2, where a manual normalisation made by the authors is added for reference along with a

translation. (References in square brackets use ONP's internal referencing system (siglum and page number) and can be looked up at https://onp.ku.dk).

TABLE 2 GPT-4o normalisation of Old Norse compared with manual normalisation

ONP citation	GPT-4o normalisation	Manual normalisation	Manual Translation	Notes
j hinn mikla staden Constantinopolim hverr. at hofvd stadr er ein enn meste æ ollv Gricklande [*KonReyk 292*]	í hinn mikla staðinn Konstantínópel, hverr at hǫfuðstaðr er einn enn mesti á ǫllu Grikklandi	í hinn mikla staðinn Constantinopolim, hverr at hǫfuðstaðr er einn inn mesti á ǫllu Grikklandi	In(to) the great city of Constantinople, which is the greatest capital in all of Greece	Both Lat. and ON forms are acceptable normalisation; inconsistent form of article *hinn/enn* is retained
lofanndi gud himneskan er firir sinn son gerir oll stormerki i heilogum anda a himni ok iordu [*Nik Holm perg 16 4° 38r*]	lofandi guð himneskan er fyrir sinn son gerir ǫll stórmerki í heilǫgum anda á himni ok jǫrðu.	lofandi guð himneskan, er fyrir sinn son gerir ǫll stórmerki í heilǫgum anda á himni ok jǫrðu.	. . . praising heavenly God, who for his son made all sacraments/ miracles in the Holy Spirit on heaven and earth	Correct automatic normalisation
uer skulum ok uera menn sattir uor imillum ok sam uerir ad oldri ok ad ati a þingi ok a þiodstefnu [*Heiðy 99*]	Ver skulum ok vera menn sáttir vár í millum ok samverir at aldri ok at ætti á þingi ok á þjóðstefnu.	Vér skulum ok vera menn sáttir vár í millum ok samverir at ǫldri ok at áti á þingi ok á þjóðstefnu.	We should be men reconciled between us and living in peace together in drinking and eating at the assembly and at public gatherings	Vowel length is not normalised ('vér'), mis-understanding of phrase 'at ǫldri ok at áti'
Ketill mælti. mvn eigi Barði þar vera ok ei er honum v lict ok kann ek eigi mann at kenna ef ei er hann. ok sva uar hann buinn i svmar a þingi [*Heið 85*]	Ketill mælti: Mun eigi Barði þar vera ok ei er honum v líkt ok kann ek eigi mann at kenna ef ei er hann. Ok svá var hann búinn í sumar á þingi.	Ketill mælti: Mun eigi Barði þar vera ok ei er honum ólíkt ok kann ek eigi mann at kenna ef ei er hann. Ok svá var hann búinn í sumar á þingi.	Ketill spoke: 'Might Barði not be there? It is not unlike him, and I am not able to recognise a man if it is not him; and he was dressed that way this summer at the assembly."	Model did not identify the separated negative prefix ('v lict' = 'ólíkt') but otherwise ok

The results showed that GPT-4o was generally capable of producing usable normalised Old Norse text. Some errors did occur. In rare cases these included lexical replacement, i.e. instead of correcting the unnormalised word the model would replace it with a different word. This can be observed in Table 2 in examples like unnormalised "sam uerir ad oldri ok ad ati a þingi ok a þiodstefnu" which the AI normalises as "samverir at aldri ok at ætti á þingi". The normalisation of *oldri* > *aldri* and *ati* > *ætti* indicates that the AI was not able to process the semantic content of the words in context. The word *oldri* is a spelling of standard *ǫldri* 'drinking'. As the AI does not seem to realise this it replaces the *oldri* form with the word *aldri* 'age' knowing that

sometimes *ǫ* and *a* can alternate according to phonetic rules. The logic of replacing *áti* with *ætti* is not as easy to figure out. Possibly the AI is considering the word *ætt* which is commonly also written *átt* meaning 'family' or 'lineage'. However, the form *ætti* is not an obvious inflected form of this word. The AI seems to be interpreting the meaning of the sentence as 'together in age and lineage' where the actual meaning is 'together in eating and drinking'. However, the two words it suggests would logically go together in certain contexts.

Another even more extreme example of such lexical replacement is the sentence "Þa geck Hyrockin a framstafn na⁄kqvans" which is normalised as "Þá gekk Hyrrokkin á framstafn knǫrrins". The wordform *na⁄kqvans* which should be *nǫkkvans* when normalised (from *nǫkkvi* 'a boat') is rendered as a completely different lexical item *knǫrrins* an erroneous form of *knarrarins* from *knǫrr* 'cargo ship'. This error is quite interesting as it shows replacement within the same semantic domain. It implies that the AI has some basic understanding of the context and knows that the word in question should refer to a vessel on water. As it does not seem to be able to make a connection between unnormalised *na⁄kqvans* and any known normalised form it replaces it with a different type of sea-faring craft. The fact that it uses the irregular form *knǫrrins* (probably for *knǫrrsins*) instead of the standard form *knarrarins* shows that it takes possible alternative forms into consideration if they are present in the original text.

Most other errors observed were not significant and were typically minor. These included:

- inconsistent treatment of articles (e.g. *hinn/enn*),
- occasional lexical substitutions based on misinterpretation
- failure to identify separated prefixes (e.g. *v líkt* → *ólíkt*),
- uncertainty in ambiguous forms.

In some cases, the model retained orthographic inconsistencies present in the source text. This limits its usefulness for fully standardised NLP pipelines but does not render it unsuitable for more exploratory or pedagogical uses.

The first phase also revealed significant practical constraints:

- Each query required approximately 3–5 seconds of processing time per citation.
- API usage costs were high, exhausting a basic plan after fewer than 2000 citations. This is likely due to the large number of tokens in the few-shot prompt.
- Batch processing, which may have improved processing time and reduced the number of billing tokens per text snippet, proved difficult to implement.
- Large-scale application of this approach would entail substantial financial and energy costs.

Given the scale of ONP, fully automated normalisation using a commercial LLM was not economically viable at this stage. It may in some cases be more efficient to employ trained human annotators.

3 Phase 2: Evaluation of Models and Outputs (January 2026)

Since the initial experiment, LLM performance on historical and lesser-resourced languages has improved considerably. The second phase therefore evaluates multiple models and compares two target outputs:

- Reconstructed standard Old Norse orthography (corresponding to early thirteenth century Norse-Icelandic)
- Modern Icelandic orthography

The underlying hypothesis is that models trained primarily on contemporary language data may perform better when asked to produce modern Icelandic, while models with stronger exposure to Old Norse material may perform better on reconstructed medieval orthography. Modern Icelandic output could also be suitable for subsequent processing on the wide variety of NLP tools developed for that language.

3.1 Method

Two system prompts were used:

- Old Norse: "You are an expert in Old Norse language and grammar. You normalise Old Norse texts to their standard forms, preserving punctuation and capitalisation where appropriate."
- Modern Icelandic: "You are an expert in Old Norse and modern Icelandic language and grammar. You modernise Old Norse texts to Icelandic, preserving words, punctuation and capitalisation where appropriate."

3.2 Results

A comparison of an original citation text with different models' normalisation and modernisation output is shown in Table 3.

TABLE 3 Comparison of different models on a single citation

ONP citation	þa uil ec bioða þér holmgaꝛngo. oc þat at uið berimc her a þingino. hafi sá fe þetta laꝛnd oc laꝛsa aꝛra er sigr fęr [*EgFrg*^{2}*340*]
Manual translation	Then I will challenge you to a duel, and that we fight here at the assembly. Let him have this property — lands and movable goods — who gains the victory.
Manual ON	Þá vil ek bjóða þér hólmgǫngu, ok þat at vit berimsk hér á þinginu. Hafi sá fé þetta, lǫnd ok lausa aura, er sigr fær.
GPT-4o (ON)	Þá vil ek bjóða þér hólmgǫngu, ok þat at við berimk hér á þinginu. Hafi sá fé þetta land ok lausa aura er sigr fær.
Gemini 3 (ON)	Þá vil ek bjóða þér hólmgǫngu, ok þat, at vit berimsk hér á þinginu. Hafi sá fé þetta, lǫnd ok lausa aura, er sigr fær.
DeepSeek (ON)	Þá vil ek bjóða þér holmgöngu, ok þat at við berimk hér á þinginu; hafi sá fé þetta, lönd ok lausa aura, er sigr fær.
Copilot (ON) (30.4.26)	Þá vil ek bjóða þér hólmgǫngu, ok þat at við berimsk hér á þinginu. Hafi sá fé, þetta lǫnd ok lǫysa aura, er sigr ferr.
GPT-5 (ON)	Þá vil ek bjóða þér hólmgǫngu, ok þat at vit berimsk hér á þinginu. Hafi sá fé þetta lǫnd ok lǫysa aura, er sigr fær.
Manual Icel.	Þá vil ég bjóða þér hólmgöngu, og það að við berjumst hér á þinginu. Hafi sá fé þetta lönd og lausa aura er sigur fær.

TABLE 3 (*CONTINUED*)

GPT-5 (Icel.)	Þá vil ek bjóða þér hólmgöngu, ok þat at vit berjumsk hér á þinginu. Hafi sá fé þetta lönd ok lausa aura er sigr fær.
Copilot (Icel) (30.4.26)	Þá vil ég bjóða þér hólmgöngu, og það að við berjumst hér á þinginu. Hafi sá fé þetta land og lausa aura er sigur fær.

The particular citation was chosen because of its slightly complex syntax and use of the subjunctive. The newer models resulted in accuracy (by comparison with the expert-normalised texts) close to that of GPT-4o. GPT, which was also the underlying model for Copilot, seems to have understood the demonstrative singular form *Þetta* 'that' to go with *land/lǫnd* 'land' and therefore normalised the vowel of the latter, incorrectly changing the number of the noun from plural to singular. Other errors were present in GPT/Copilot. DeepSeek produced largely correct normalisation, although using post-1250 orthography ('ö' for 'ǫ'). Gemini 3 produced correct results. On the basis of this and similar tests we deem it the best next-generation model for this particular task.

The results indicate that both normalisation strategies are viable. However, performance varies across models and target outputs:

- Some models produce strong Old Norse normalisations with good syntax.
- Some newer models produce inferior results to their predecessors (GPT 5 vs. 4)
- Others show better fluency when producing modern Icelandic.
- In some cases, "modern Icelandic" output is only partially modernised and in others Old Norse output is incorrectly modernised.
- Models occasionally adjust morphology (e.g. number agreement) based on interpretation of context.

There is some indication that LLMs may already implicitly use modern language representations as an intermediate step, suggesting that explicit use of modern Icelandic may improve accuracy. This remains a hypothesis for further testing.

4 Phase 3: Evaluation of next-generation model (April 2026)

4.1. Method

Since commencing the study LLMs have improved still further, as shown above. Comparative testing of different models revealed that Gemini 3 produced promising results for normalisation that exceeded previous tests and other models. The pricing structure and speed of this model makes it more suitable for future applications if it proves to be accurate. As the model already appears to have an understanding of standard normalisation practices, it did not seem necessary to take a few-shot approach with specific examples of normalised citations. We instead used no-shot prompts with a more detailed system prompt specifying key rules. Input was organised into batches and output requested in a specific JSON format. The system prompt was eventually refined to the following:

> You are an expert philologist specialising in Old Norse.
>
> Your task is to take unnormalised Old Norse text snippets (as found in the Dictionary of Old Norse Prose/ONP) and convert them into standard normalised Old Norse (standard academic orthography). Do not adjust or correct the grammar. Some texts may already be normalised and can be outputted as is.
>
> Key Rules:
>
> 1. Standardise vowels: Use 'ǫ' for o-caudata where appropriate, 'ø' or 'oe' correctly, and acute accents for long vowels (á, é, í, ó, ú, ý, æ, oe).
> 2. Consonants: Use 'ð' and 'Þ' correctly. Standardise 'u' to 'v' and 'i' to 'j' where they function as consonants.
> 3. Middle voice/mediopassive: Use -sk/-mk (not -z, -st, -zt)
> 4. Output JSON: { "normalisedSnippets": [{ "id": " . . . ", "original": " . . . ", "normalised": " . . . ", "notes": " . . . " }] }

The specific model used was "gemini-3-flash-preview". Other models were more problematic: "gemini-3.1-flash-lite" was very fast but output contained noticeably more errors; the 'Pro' model was slower and more expensive and did not produce better results.

4.2. Results

To compare with previous phases, 70-80 citations from three models were manually scored for errors by the authors and cross-checked. The models tested were gpt-4o (from the Phase 1 results), gemini-3-flash-preview (Phase 3) and gemini-3-flash-lite (for reference). Errors included single-letter errors, compound identification errors, grammatical/inflection errors (weighting: 1), lexical errors (weighting: 2) and modernisation errors (where the resulting normalisation belongs to an incorrect period, with weighting: 0.5). GPT-4o yielded 6.0 errors per 100 citation tokens, Gemini 3 Flash Lite 9.9 errors, and Gemini 3 Flash 3.6 errors, showing a considerable advancement over the results in phase 1.

Examples of the phase 3 results are shown in Table 4.

TABLE 4 Input and output of normalisation using gemini-3-flash-preview

Input (ONP citation text)	Output (Gemini 3 Flash)	Translation (manual)
þat muɴdi at því ósætti verþa es vísa vǫn vas at þér barsmíþer gørþisc á miþli manna es landet eyddisc af [*Íslb*^{x} 29]	þat mundi at því ósætti verða er vísa ván var at þær barsmíðar gørðisk á miðli manna er landit eyddisk af	That would lead to such discord, as it was certain that battles would take place among men such that the land would be devastated
Sigvatr hafdi avxi i hendi oc lystrr til Þorsteins, enn rossit var millvm þeira, er hann Þorsteinn helt á; oc lagðiz hann niðr hia rossino ǫðrumegin, oc com auxar skaptit a herða-toppinn oc brotnaði i svnðr, enn oxin com a herdar-blad Þorsteini, oc vard honum ecki mein at því [*Stu*^{I}*K 182*]	Sigvatr hafði øxi í hendi ok lýstr til Þorsteins, en hrossit var millum þeira, er hann Þorsteinn hélt á; ok lagðisk hann niðr hjá hrossinu ǫðrumegin, ok kom øxarskaptit á herðatoppinn ok brotnaði í sundr, en øxin kom á herðablað Þorsteini, ok varð honum ekki mein at þv.	Sigvatr had an axe in his hand and struck at Þorsteinn, but the horse that Þorsteinn was holding was between them; and he threw himself down on the other side of the horse, and the axe handle hit the top of his shoulders and broke apart, but the axe head hit Þorsteinn's shoulder blade, and he was not harmed by it.

TABLE 4 (*CONTINUED*)

Input (ONP citation text)	Output (Gemini 3 Flash)	Translation (manual)
hann for i brott ok kom til hreiðmars ok syndi oðni gvllit. en er hann sa bavginn syndiz honvm avar fagr ok tok af feno [*SnEU360*]	Hann fór í brott ok kom til Hreiðmars ok sýndi Óðni gullit. En er hann sá bauginn, sýndisk hǫnum afar fagr ok tók af fénu.	He went away and came to Hreiðmarr and showed Óðinn the gold. But when he saw the ring, it seemed to him extremely beautiful and he removed it from the treasure
sv var sett nefnd ok vandrædis fijrıng þa i stað, at griþ voro sett þann uetr með þeim ollum [*Stu BLAdd 11127166r*]	Svá var sett nefnd ok vandræðisfirring þá í stað, at grið váru sett þann vetr með þeim ǫllum.	That arbitration and avoidance of trouble was established at that time, so that a truce was formed between them all for that winter.

The errors are largely minor (underlined in the table): in the first citation the mod. Icel. *-ar* inflection (first recorded in the 16th century) is used instead of ON *-ir*; in the second *herðablað* is used for the more common and correct *herðarblað* (although both variants exist in ON). The only problematic example is in the final citation, where Gemini misinterprets the syntax of the first clause and normalises a demonstrative to an adverb ('sú'/'svá') and interprets both the accompanying noun as a verb participle and vice versa. This results in an incorrect normalisation and is the only example where Gemini appears to have misunderstood the Old Norse text itself.

The same method was applied to the examples discussed above as problematic in phase 1 ('Þa geck Hyrockin . . . ' and 'uer skulum ok . . . '). Gemini produced a normalisation identical with the expert manual normalisation.

Batches took a similar length of time (30-100 seconds) regardless of the number of citation texts in the batch (10-100). This suggests much of the latency is due to processing the system prompt, meaning that larger batches are more efficient. However, batches above around 30 texts often failed due to errors in the generated JSON output. Eventually we arrived at a batch size of 25 to ensure reliable output. The time per citation was consequently less than 2 seconds (average 1.56s/citation for batches of 25-50 citations).

To evaluate the usefulness of this approach it is important to understand both the financial cost and environmental impact compared with traditional (human) approaches. As our approach potentially provides only an additional access layer for the dictionary, the costs should ideally be minimal. To estimate the total load we use the API reporting of the Flash model, where an average batch of 25 citations uses around 10,000 total tokens. The total cost for the estimated 200 million tokens would be USD600 for the Gemini Flash model.

Calculating the energy usage and environmental impact is extremely difficult given the lack of transparency from AI service providers around these issues. Elsworth et al. (2025) is a non-peer reviewed, internal Google estimate of the energy usage and environmental impact of Google's AI based on a standard prompt, but the paper does not report on the size of this prompt. We therefore use the average reported from WildChat (Zhao et al. 2024) of around 300 tokens for each interaction. Each 25-citation batch is therefore the equivalent of 33 human interactions used for benchmarking. The total corpus would take, conservatively, the equivalent of 700,000 standard prompts. The total consumption for processing the entire corpus based on Elsworth and colleagues' data for Gemini would be 168 kWh, 21kg CO_2 and 182L water. These are not

inconsiderable financial and environmental impacts but are negligible compared with the equivalent cost of paying staff and the indirect environmental impacts of hardware, heating, building space and so on, given that such a task would take trained editors many years to perform.

There remain significant issues with the quality of the output given ONP's high scholarly standards. All output would need to be checked and corrected before it could be integrated into the official dictionary interface. However, this model works well enough to be potentially helpful for users of the dictionary to get a better grasp of the textual context.

5 Conclusion

The study demonstrates that LLM-assisted normalisation can meaningfully improve access to historically complex lexicographic resources. The results suggest several key points:

- LLMs can produce usable normalisations despite apparently limited explicit training on historical data.
- Errors are typically minor but systematic enough to require human oversight.
- Fully automated workflows are currently constrained by cost and efficiency but are rapidly improving.
- Hybrid workflows combining automated processing with expert review are promising.

The use of modern Icelandic as an intermediate layer introduces additional possibilities for linguistic enrichment and search functionality, particularly through integration with existing NLP tools.

Given that it is currently beyond the resources of the project to check all automatically normalised citations, a possible solution would be to provide a user-initiated solution, whereby users could request normalised citations via the LLM API. A similar approach is used by ONP for definitions where the entries have been only edited in Danish. In these cases, the user can request an automatic translation via an API, which is then flagged as such (Wills et al., 2021). The automatically converted text does not then appear in web searches but adds an access layer for users.

In a similar way, the evidential status of historical citation material must be preserved. Normalisation should therefore be presented as an additional access layer rather than a replacement for diplomatic citation.

Although limited in scale, this study shows that LLM-assisted normalisation can support broader access to historical lexicographic data. It enables new forms of interaction with citation material, particularly for non-specialist users and computational applications.

Rather than replacing philological expertise, such tools are best understood as components of hybrid workflows. These combine automated processing with human oversight, allowing historical dictionaries to expand their usability while maintaining scholarly rigour.

Future work will involve refining prompts, comparing workflows, and systematically evaluating the accuracy of different normalisation strategies.

References

Baetke, W. (1965–1968). *Wörterbuch zur altnordischen Prosaliteratur* (Sitzungsberichte der Sächsischen Akademie der Wissenschaften zu Leipzig: Philologisch-historische Klasse, Vol. 111). Berlin.

Bernharðsson, H., Haugen, O. E. & Berg, I. (2019). Normalisation. In O. E. Haugen (Ed.), *The Menota handbook: Guidelines for the electronic encoding of Medieval Nordic primary sources* (Version 3.0, Ch. 10). Medieval Nordic Text Archive. http://www.menota.org/handbook.xml

Bjarni Einarsson (Ed.). (2001). *Egils saga Skallagrímssonar* (Jón Helgason, preparatory work). EA A 19.

Bollmann, M. (2019). A Large-Scale Comparison of Historical Text Normalization Systems. In J. Burstein, C. Doran & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 3885-3898). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1389

Cleasby, R. & Vigfusson, G. (1874). *An Icelandic-English Dictionary*. Oxford.

Elsworth, C., Huang, K., Patterson, D., Schneider, I., Sedivy, R., Goodman, S. & Manyika, J. (2025). Measuring the environmental impact of delivering AI at Google Scale. *arXiv preprint arXiv:2508.15734*.

Fritzner, J. (1886–1896). *Ordbog over det gamle norske sprog* (Rev. ed., Vols. 1–3). Kristiania.

Heggstad, L., Hødnebø, F., & Simensen, E. (1975). *Norrøn ordbok*. Oslo.

Nordal, S. (Ed.). (1933). *Egils saga Skalla-Grímssonar* (Íslenzk fornrit 2). Hið íslenzka fornritafélag.

Pavlopoulos, J., Bakagianni, J., Pouli, K. & Gavriilidou, M. (2025). Open or closed LLM for lesser-resourced languages? Lessons from Greek. *arXiv*. https://arxiv.org/abs/2501.12826

Ross, E. A. S. (2023). A new frontier: AI and ancient language pedagogy. *Journal of Classics Teaching*, 24(48), 143–161. https://www.cambridge.org/core/journals/journal-of-classics-teaching/article/new-frontier-ai-and-ancient-language-pedagogy/A63EF69F5FE5529F0F45FB1EB655A9F7

Salazar Alva, R. G., Nuñez, M., López, C. & Martín Arista, J. (2025). AI-driven generation of Old English: A framework for low-resource languages. *arXiv*. https://doi.org/10.48550/arXiv.2507.20111

Sommerschield, T., Assael, Y., Pavlopoulos, J., Stefanak, V., Senior, A., Dyer, C., Bodel, J., Prag, J., Androutsopoulos, I. & de Freitas, N. (2023). Machine learning for ancient languages: A survey. *Computational Linguistics*, 49(3), 703–747. https://direct.mit.edu/coli/article/49/3/703/116160/Machine-Learning-for-Ancient-Languages-A-Survey

Wills, T., Battista, S. & Jóhannsson, E. Þ. (2021). Making an ongoing historical dictionary accessible to an English-speaking audience. Abstract from *Euralex XIX: Lexicography for Inclusion*. https://euralex2020.gr/wp-content/uploads/2021/10/EURALEX2020_BookOfAbstracts-Preview.pdf

Zhao, W., Ren, X., Hessel, J., Cardie, C., Choi, Y. & Deng, Y. (2024). Wildchat: 1m chatgpt interaction logs in the wild. *arXiv preprint arXiv:2405.01470*.

Contact information

Tarrin Wills
Dictionary of Old Norse Prose, University of Copenhagen
tarrin@hum.ku.dk

Ellert Thor Johannsson
The Árni Magnússon Institute for Icelandic studies
ellert.thor.johannsson@arnastofnun.is

Marija Žarković Eriksson

LEGAL LEXICON AND THE E-*DRAE* 1884

A Diachronic Study of Specialised Vocabulary in a General Dictionary Using a Digital Critical Edition

Abstract This paper presents a methodology for the diachronic study of legal lexical units related to criminal law in a nineteenth-century general-language dictionary, the twelfth edition of the *Diccionario de la lengua castellana* of the by the Real Academia Española (*DRAE* 1884), as analyzed through its critical and digital edition (e-*DRAE* 1884). The study combines a lexicographical examination of how these units are defined, structured and interconnected with their contextualization within the legal, social and ideological framework of nineteenth-century Spain, drawing on the sociocognitive theory of ideology and on the notion of the dictionary as discourse. Selected case studies from the criminal lexicon illustrate continuities and ruptures between the inherited juridical tradition and the emerging liberal order. The paper finally discusses the methodological advantages that the e-*DRAE* 1884 environment offers for research on specialized lexis in historical dictionaries.

Keywords Legal lexicon; historical lexicography; computational lexicography; *DRAE* 1884; ideology; e-*DRAE* 1884

1 Introduction

Throughout the nineteenth century, the dictionary of the Real Academia Española became a key reference work for both the history of lexicography and the history of the Spanish language as a whole (Freixas Alás, 2019). Between 1803 and 1899 the Academy published ten editions of the dictionary, in a period also marked by an unprecedented expansion of specialised vocabulary linked to the development of science and technology (Clavería Nadal, 2016). Within this sequence, the twelfth edition, published in 1884 (*DRAE* 1884), occupies a particularly prominent place: prepared between 1869 and 1884, it is the outcome of a thorough revision of the previous edition and devotes specific attention to technical vocabulary (Clavería Nadal, 2021a).

A sustained line of lexicographic research has shown that meaning is never entirely neutral, since dictionaries not only record language use but also reflect social values, cultural assumptions and particular ways of interpreting the world (Moon, 2014; Fuertes-Olivera and Tarp, 2022). The legal lexicon registered in the *DRAE* 1884 is, in this respect, a privileged observation point for diachronic research on specialised vocabulary. Issued during the final decades of a juridically and politically turbulent century, the dictionary mirrors a legal system in transition: from the Old Regime, with its plurality of fueros and jurisdictions, to a new liberal order articulated around codification, the separation of powers and individual rights (Tomás y Valiente, 2012). The way in which legal units are lemmatised, defined and labelled in a general-language dictionary therefore reflects both inherited juridical traditions and emerging conceptual models.

The aim of this paper is to illustrate, through a selection of case studies, a methodology for the diachronic semantic study of specialised legal vocabulary in the *DRAE* 1884, using the digital critical edition e-*DRAE* 1884 (Clavería Nadal et al., 2026) as the research environment. We focus on the lexicon of substantive criminal law, which constitutes one of the largest and most stable nuclei of legal terminology in the academic tradition. The article is organised as follows. Section 2 outlines the historical, juridical and lexicographical context of the *DRAE*

1884. Section 3 presents the e-*DRAE* 1884 environment and the methodology adopted for the retrieval and analysis of the criminal-law lexicon. Section 4 develops the case studies, organised around the notion and modalities of unlawful conduct and the notion and modalities of legal penalty. Section 5 discusses the results from a sociocognitive perspective, and Section 6 closes with the main conclusions and future lines of research.

2 Context: penal reform and the DRAE 1884

2.1 Historical and juridical context

The nineteenth century was, for Spain, a period of deep juridical reorganization. Reason became the critical and universal tool of a new culture and was applied to law with the same analytical method as to scientific research, breaking with the older deductive system (Rodríguez Ennes, 2012). At the legislative level, this led to a growing prevalence of statute law as the only legitimate source of law, weakening the previous legal pluralism and establishing a strict legal monism. After the War of Independence, the Spanish legal system absorbed ideas derived from the Enlightenment and from French law (cf. Cobo del Rosal, 2011) and embarked on a long process of codification that culminated in the *Decreto de Unificación de Fueros* of 1868 (a decree that abolished special jurisdictions and legal privileges), the *Código Penal* of 1870, and the *Ley de Enjuiciamiento Criminal* of 1882.

This transformation had direct consequences for the lexicon of criminal law. New concepts had to be coined, older terms re-defined and inherited categories rearranged. The *DRAE* 1884, prepared during this period of intense legal reform, is therefore a particularly rich source for observing how legal concepts circulate beyond specialized normative texts and enter the sphere of general linguistic codification.

2.2 The DRAE 1884 as a lexicographical milestone

From a metalexicographical point of view, the *DRAE* 1884 represents a major step in the systematisation of the lexicographical codification of Spanish (Clavería Nadal, 2021a). The revision of the eleventh edition was carried out between 1869 and 1884, a period of marked political and social instability. Despite this turbulent background, the Academy undertook a far-reaching reform of the dictionary on the basis of the *Reglas para la corrección y aumento del diccionario vulgar* (1870), a programmatic document that set out both the treatment of the various lexicological and grammatical groups and the general directives of the reform (cf. Rodríguez Ortiz & Garriga, 2010; Clavería Nadal, 2021b). Although the new edition was elaborated, as always, on the basis of the previous one (eleventh edition, 1869), the procedures employed were significantly more systematic, and the scope of the revision affected almost every part of the macrostructure and microstructure of the work (cf. Blanco Izquierdo & Clavería Nadal, 2021).

The reform also entailed a more consistent use of diatechnical labels, the incorporation of etymological information, and a clearer organisation of senses within the article, all of which contributed to a more structured representation of specialised vocabulary, including that of the legal field. For these reasons, the *DRAE* 1884 has been described as a lexicographical milestone within the academic series: it preserves the inherited tradition of the *Diccionario de autoridades* (1726-1739), the first dictionary published by the Academy, while introducing methodological innovations that anticipate the twentieth-century dictionary editions. This combination of continuity and innovation makes it an especially productive object for the historical study of specialised vocabulary, and in particular for the analysis of how the legal lexicon, including its criminal-law component, is selected, defined and contextualised within a general-language dictionary.

3 Methodology

The case study presented in this paper is conducted using the **e-*DRAE* 1884**, the digital critical and hypertextual edition of the twelfth *DRAE* (Clavería Nadal et al., 2026). The e-*DRAE* 1884 integrates, in a single tool, the distinctive features of four different types of edition: facsimile, interpretative, critical and annotated. The query interface allows simple and combined searches over lemma, etymology, definition text, lexicographical labels (including diatechnical labels such as *For*. for legal lexicon, *Med*. for medical or *Mil*. for military), grammatical categories, and the body of scholarly comments, with parallel display of transcribed text and facsimile image.

The methodology adopted in this study combines a descriptive and a comparative approach. The starting point is the systematic retrieval, through the e-*DRAE* 1884 search interface, of the legal lexicon of the twelfth academic dictionary. Advanced searches and filters on the platform enabled, first, the retrieval of the legal lexicon, understood as the vocabulary used to designate legal concepts and realities through specialized terms or ordinary-language words that acquire legal meanings in specific contexts (Mattila, 2012). Second, these tools enabled the identification of the subset related to criminal law, comprising vocabulary referring to criminal norms, conduct that infringes those norms, and the penalties or sanctions (*Diccionario panhispánico del español jurídico*, 2023: s. v. *derecho penal*).

Three complementary procedures were used for retrieval, since several features of this specialized lexicon had to be considered. First, all entries and senses bearing the diatechnical label were identified. Although the *DRAE* 1884 includes a broad set of lemmas and senses, 1077 in total, marked with the domain label *For*. for *forense*, and thus linked to the legal domain, many lexical units appear without any indication of specialized use. As Battaner (1996, p. 104) points out, the presence of thematic labels seems to be arbitrary in the lexicographical tradition of the RAE. This means that part of the specialized lexicon lacks this type of identification. In the legal domain, the lexical units without a label are often denominations of general-language concepts frequently used in law (cf. Alcaraz, Hughes and Gómez. 2014), such as *delito* [1][1] or *culpa*, or archaic legal concepts belonging to a remote historical reality. However, certain legal or legally relevant concepts also appear without a subject-field label, as in *concusión* [2] and *depredación* [2]. These lexical units were identified through a predefined set of keywords related to criminal law, established based on legal sources from the period, such as criminal codes, and specialized studies by legal historians. The keywords were then used to retrieve entries semantically related to criminal law but lacking an explicit subject-field label, through the in-text search function of the e-*DRAE* platform.

Each of these units is then analyzed at the macrostructural level (presence, absence, addition, removal of lemmas across editions) and at the microstructural level (definitional formula, lexicographical labels, examples, cross-references). The diachronic dimension is established by comparing the text of the *DRAE* 1884 with the previous editions of the same dictionary, particularly the *DRAE* 1869, using the ***Nuevo tesoro lexicográfico de la lengua española*** as auxiliary resource.

The interpretation of the data is informed by two theoretical frameworks: the sociocognitive theory of ideology (Van Dijk, 1998), which relates linguistic choices to socially shared systems of knowledge and belief; and the notion of the dictionary as discourse (Benson, 2002;

[1] Many legal lexical units do not appear in the dictionary as independent lemmas, but rather as senses or subentries. In this paper, when an article contains several senses, the number in square brackets indicates the position of the relevant sense in the *DRAE* 1884. The e-*DRAE* 1884 interface makes it possible to retrieve these units both through lemma searches and full-text searches.

Hornscheidt, 2008; Rodríguez Barcia, 2012; Chen, 2019), which insists on the dictionary as a regulated discursive practice rather than a neutral mirror of usage. Van Dijk's sociocognitive theory of ideology treats ideology not as a merely external political context, but as a cognitive and social structure that mediates between discourse and society. Applied to lexicography, this makes it possible to interpret definitions, labels and classifications as textual traces of broader ideological models through which a community organizes legal and social knowledge. On the other hand, the notion of the dictionary as discourse understands lexicography as a historically situated practice through which knowledge is produced and organized, so that definitions do not simply record usage but actively construct meaning through selection and recontextualization. The combination of these perspectives makes it possible to interpret changes in definitions not only as updates motivated by linguistic evolution, but also as discursive operations that reorganize and reformulate juridical knowledge.

4 Lexicon of criminal law in a nineteenth-century general dictionary

At the macrostructural level, the substantive criminal-law lexicon[2] retrieved from the *DRAE* 1884 for this study comprises 174 lexical units and senses. For the purposes of semantic analysis, the material was organized along two main categories: the lexicon of unlawful conduct (Table 1) and the lexicon of penalties (Table 2).

TABLE 1 Lexical units denoting unlawful-conduct concepts in the *DRAE* 1884, classified by protected legal interest

Concept of unlawful conduct	*atentado* [2]; *atentar* [2]; *cometer* [2]; *crimen*; *delito* [1]; *exceso*; *perpetrar* [1]; *premeditar* [2]; *conato* [3]; *reincidir* [1]; *relapso* [1]; *dolo* [2]; *traspasamiento* [1]; *traspasar* [6]; *transgredir* [1]; *infracción* [1]
Crimes against persons	*alevosía* [1]; *asesinar* [1]; *asesinato* [1]; *asesino* [1]; *fratricidio* [1]; *homicida* [1]; *homicidio* [1, 2]; *infanticida* [1]; *infanticidio* [1]; *lesión grave* (s. v. *lesión*); *matricida* [1]; *matricidio* [1]; *muerte* [3]; *parricidio* [1]; *rapto* [2]; *raptor* [2]; *violar* [2]
Crimes against honor	*calumnia; calumniar; difamar; difamación; injuria* [1, 2]
Sexual crimes	*adulterio* [2]; *estupro*; *rapto* [2]; *raptor/ra* [2]; *violar* [2]
Crimes against property	*abigeato* [1]; *baratería* [1, 2]; *colusión* [1]; *dolo* [1]; *engañar* [2]; *estafa* [1]; *estafar* [1]; *estelionato* [1]; *fraudar* [1]; *fraude* [1]; *hurtar* [1]; *hurto*; *latrocinio* [1]; *pillaje* [1]; *pillar* [1]; *robar* [1]; *robo* [1]
Crimes of falsification	*expendedor* [3]; *expender* [4]; *falsedad* [3]; *falso/sa* [1]
Crimes against State security	*alta traición* (s. v. *traición*); *pirata*; *piratería* [1]; *piratería* [2]; *traición* [1]
Crimes against the Constitution	*crimen de lesa majestad* (s. v. *crimen*); *delito de lesa majestad* (s. v. *delito*); *rebelión*; *sedición* [1]
Crimes against liberty and security	*coacción; desacato [3]; secuestro; amenazar; amenaza [1, 2]*
Customs crimes[a]	*contrabando* [1, 2, 3]; *meter* [2]
Crimes committed by public officials in office	*atentado* [1]; *baratería* [3]; *cohechar* [1]; *cohecho* [1]; *concusión* [2]; *corromper* [5]; *depredación* [2]; *malversar* [1]; *peculado* [1]; *prevaricar* [2]; *prevaricato* [1]

[a]Customs crimes were regulated mainly by special customs and tax legislation and therefore fall outside the categories established in the 1870 Spanish Penal Code.

TABLE 2 Lexical units denoting concepts related to the notion and modalities of legal penalty in the *DRAE* 1884

Concept of legal penalty	*pena* [1]; *enmienda* [3]; *suplicio* [1]; *paga* [3]; *castigo* [1]
Capital penalties	*acañaverear* [1]; *ahorcado* [1]; *ahorcar* [1]; *ajusticiado* [1]; *ajusticiar* [1]; *apedrear* [2]; *arcabucear* [2]; *aspar* [2]; *atenacear* [1]; *crucificar* [1]; *dar garrote* (*s. v. garrote*); *enrodar* [1]; *garrote* [2]; *hervencia* [1]; *lapidar* [1]; *palo* [5]; *pasar por las armas* (*s. v. arma*); *pena capital* [7]; *pena ordinaria* (*s. v. pena*); *poner á uno en un palo* (*s. v. palo*); *sentenciar á uno á garrote* (*s. v. garrote*); *último suplicio* (*s. v. suplicio*)
Corporal penalties	*aflictivo* [2]; *azotado* [2]; *decalvar* [1]
Deprivation or restriction of liberty	*arresto* [3]; *calabozo* [1, 2]; *confinación* [1]; *confinar* [2]; *deportación* [1]; *deportar* [1]; *destierro* [1]; *desterrar* [1]; *encarcelación* [1]; *encarcelar* [1]; *encerramiento; encierro* [3]; *exilio; extrañamiento* [1]; *extrañar* [1]; *galera* [2]; *penitenciaria* [3]; *prisión* [9]
Rights-depriving penalties	*degradación actual, degradación real and degradación verbal* (*s. v. degradación*); *privación* [3]
Forced labor	*cadena* [9]; *forzado* [3]; *galera* [7]; *gastador* [3]; *presidiario* [1]; *presidio* [3, 5]; *rematado* [2]; *remo* [5]
Financial penalties	*achaque* [6]; *comiso* [1]; *homicillo* [1]; *multa* [1]; *multar* [1]; *pena pecuniaria* and *penas de cámara* (*s. v. pena*)
Honor-based penalties	*afrenta* [3]; *argolla* [3]; *estigma*; *hábito de penitencia* (*s. v. hábito*); *poner en un palo* (*s. v. palo*); *sacar á la vergüenza á uno* (*s. v. vergüenza*); *vergüenza* [5]; *vindicta pública* (*s. v. vindicta*)

The lexicon of unlawful conduct, which comprises 97 lexical units and senses, is classified according to the protected legal interest affected by each offence. The categories are based primarily on the structure of Book II of the *Código penal* of 1870, especially crimes against persons, honor, property, falsification, State security, the Constitution, liberty and security, and crimes committed by public officials in office.

The lexical units related to penalties, which comprises 77 lexical units and senses, is grouped into seven categories: capital penalties, corporal penalties, deprivation or restriction of liberty, rights-depriving penalties, forced-labor penalties, financial penalties, and honor-based penalties. This classification reflects the principal punitive modalities attested in nineteenth-century Spanish criminal law, while also taking account of older punitive practices preserved in the dictionary through definitions, subentries, and phraseological units. Most of these categories were already represented in previous editions of the dictionary, although the 1884 edition reorganized and expanded them through new senses, subentries, and subject-field labels, which will be discussed in the next section.

4.1 The notion and modalities of unlawful conduct

The legal lexicon related to criminal offences in *DRAE* 1884 presents a broad and differentiated conception of unlawful conduct, structured around several semantic modalities that reflect the evolution of nineteenth-century Spanish penal thought. A first group of entries defines illegal acts in general and abstract terms through words such as *delito* [1], *crimen, exceso,*

[2] The analysis focuses exclusively on the vocabulary of substantive criminal law, that is, lexical units relating to crimes and penalties rather than to criminal procedure or judicial administration.

atentado [2], or *infracción* [1], which describe transgression as the violation of legal and social order. Alongside these nouns, verbs such as *cometer* [2], *perpetrar* [1], *transgredir* [1], and *traspasar* [6] emphasize the action of breaching a norm and reinforce the idea of unlawful conduct as a conscious act against established authority.

Spanish liberal constitutionalism sought to replace a fragmented judicial system with a more centralized and uniform one, and the reforms after 1868 aimed to bring most cases under the ordinary courts. This principle was consolidated through the *Ley Orgánica del Poder Judicial* of 1870 and the *Ley Provisional de Enjuiciamiento Criminal* of 1872, which standardized judicial organization and criminal procedure within a single legal framework (Tomás y Valiente, 2012, pp. 490-491). The 1884 edition of the *DRAE* reflects this legal context in a subtle way. In the definition of *delito* ('crime'), the replacement of "alguna ley" ('some law') by "la ley" ('the law') points to a more unified conception of legal order. The concept of crime no longer presented merely as the infringement of one rule among many, but as an offence against a unified legal order:

> **Delito**. [. . .] Culpa, crimen, quebrantamiento de la ley (*DRAE* 1884).

The definition of *culpa* ('fault') also changes in 1884. The expression "por propia voluntad" ('by one's own will'), used in the editions from 1837 to 1869, is replaced by "á sabiendas y voluntariamente" ('knowingly and voluntarily'):

> **Culpa**. [. . .] Falta más ó menos grave, cometida á sabiendas y voluntariamente (*DRAE* 1884).

This wording places greater emphasis on awareness and voluntariness, aligning the academic definition with the principle of culpability already consolidated in legal doctrine (cf. Sainz Guerra, 2004). While *delito* [1] is defined mainly through breach of the law, *culpa* is described through the conscious and voluntary conduct of the person who commits the fault. The two entries therefore complement each other: one foregrounds the legal norm that has been broken, while the other highlights the agent's responsibility.

This dictionary edition shows a similar movement towards a more precise legal conception of criminal responsibility. Entries such as *autor/ra* [7], *cómplice* [1], *reo* [2], *agresor/ra* [1], or *malhechor*, ra [1] distinguish different forms of intervention in punishable acts and reflect a more technical doctrine of authorship and participation. The entry *agresor/ra* no longer refers only to the person who physically attacks another "para matarla, herirla ó hacerle cualquier daño" ('with the intention of killing, injuring or causing any kind of harm'), but also includes two legal senses that define the aggressor as the person who violates another's right or provokes a quarrel through insult, threat, challenge, or provocation (*DRAE* 1884: s. v. *agresor/ra* [2]; s. v. *agresor/ra* [3]). This semantic expansion reflects a broader legal understanding of aggression, no longer limited to material injury but extended to the violation of rights and to legally relevant forms of provocation. The same tendency appears in the definition of *autor/ra* [7]. Whereas the previous editions equated *autor/ra* with the general term *causante* [2], *DRAE* 1884 defines it in criminal matters as the person who commits the criminal offence, forces or directly induces others to commit it, or cooperates in its execution through an act without which it would not have been carried out. The lexicographical discourse thus becomes more technical and differentiated, since it incorporates the figures of the direct perpetrator, the instigator, and the necessary cooperator. This development points to a modern doctrine of authorship and participation, in which criminal responsibility is determined not only by the material result of the offence, but also by the type and degree of intervention in the punishable act (cf. Cerezo Mir 2003).

When it comes to specific criminal modalities, the lexical treatment of offences committed by public officials in *DRAE* 1884 reflects the increasing attention paid by nineteenth-century Spanish penal law to administrative corruption and abuses of authority. After the Criminal Code of 1848, and later in the Criminal Code of 1870, this field came to include both misconduct by public officials and corrupt intervention by private individuals (cf. Oneca, 1970, p. 250; Iñesta Pastor, 2018). This development is visible in definitions of the lexical units such as *concusión* [2], *depredación* [2], and *malversar*. Starting from the 1869 edition, the dictionary of the RAE includes the second sense of *concusión* [2] defined through three legally relevant elements: arbitrary exaction, the status of the agent as a public official, and personal profit. The noun *depredación* [2] similarly links misappropriation and unlawful exaction to abuse of authority or trust, while the definition of the verb *malversar* incorporates the adverb *ilícitamente*, which explicitly marks the unlawful nature of the conduct. In this context, the case of *baratería* [3] is particularly significant. Earlier editions define the crime as accepting money or gifts to pass sentence, while *DRAE* 1884 narrows it to accepting them for passing a just sentence. The adjective *just* does not exclude illegality; rather, it shows that the punishable element lies in the corruption of the judicial function itself. The definition therefore places the offence not in the material injustice of the decision, but in the breach of impartiality caused by the judge's acceptance of a benefit. This modification illustrates how *DRAE* 1884 preserves an older juridical term while adapting it to the nineteenth-century concern with public integrity and the control of judicial corruption.

4.2 The notion and modalities of legal penalty

The legal vocabulary of punishment in the late nineteenth-century editions of the *DRAE* reflects the consolidation of a more moderate and correction-oriented penal doctrine. Influenced by liberal reforms and by the penal theories associated with the Criminal Code of 1848, punishment increasingly came to be understood not only as retribution, but also as a mechanism of moral reform and prevention (cf. López Melero, 2012; Masferrer y Alvarado, 2017, p. 198). This evolution becomes particularly visible in the dictionaries published by the Real Academia Española during the second half of the century, especially in the editions of 1869 and 1884, which introduce and redefine several entries related to corrective punishment and disciplinary regulation.

One of the most illustrative examples is the lexical unit *correccionalmente*, incorporated in *DRAE* 1869 and defined in *DRAE* 1884 in relation to punishment or procedure of a corrective nature. Its definition reflects a conception of punishment centered less on physical suffering and more on reform and social correction. A similar tendency appears in *corrección*, whose third meaning defines correction as the reprimand or censure of an offence, fault, or defect, thus linking punishment with moral regulation and social order. The same logic underlies *disciplinario/ria*, also introduced in *DRAE* 1869, where disciplinary punishment is associated with institutional order and corrective sanctions. In the twelfth edition, the legal meaning of *arresto* [3] is incorporated and defined as a brief confinement imposed as correction or punishment, which likewise points to the influence of correctionalist doctrine.

At the same time, the dictionaries preserve older conceptions of punishment inherited from earlier editions and connected with corporal penalties and exemplary public sanctions. In *DRAE* 1884, two different lexicographic strategies coexist. The first is historicization, through which a punishment is presented as belonging to a past legal order, not by an explicit label, but through grammatical and temporal markers such as verbal tense. This strategy can be observed in the treatment of the legal sense of the lexical unit *argolla* ('ring collar' or 'collar used for public punishment'). By the late nineteenth century, changes in

penal thought and political criticism of punishments considered infamous had contributed to the gradual disappearance of the *argolla* as a legal punishment. The modifications introduced in the definition of *argolla* [3] in *DRAE* 1884 respond directly to these legislative changes and reveal the close connection between academic lexicography and the evolution of criminal law:

> **Argolla**. [. . .] || Castigo público que se ejecutaba con algunos delincuentes, poniéndolos á la vergüenza metido el cuello en una argolla. En España recientemente consistía esta pena en preceder el reo á otro ú otros de pena capital, conducido en caballería, en ser colocado en un asiento sobre el cadalso y en permanecer en él mientras duraba la ejecución, asido por el cuello á un madero con una argolla. Este castigo ha desaparecido ya de nuestro código penal (*DRAE* 1884).

The use of the past tense in the opening phrase, *se ejecutaba* ('was carried out'), signals that the punishment is no longer presented as current practice, but as something that belonged to the past. The definition then describes the specific form that the *argolla* punishment had recently taken in Spanish legislation and finally refers explicitly to the *Código penal* of 1870 to state that this form of punishment had already disappeared from the legal system. In this way, the entry not only defines the lexical unit but also places it in its historical and legislative context, bringing the dictionary into line with contemporary penal reform.

Other entries, however, show the opposite tendency, namely the preservation of older punitive concepts within the dictionary. The adjective *aflictivo/va* was first incorporated in 1770 and referred to corporal punishments imposed by judicial authority. In *DRAE* 1884, the entry is divided into two distinct senses, one general and one juridical, with the latter still associated with corporal penalties and without any indication that the practice belonged to an earlier legal order. Even more revealing is the case of capital punishment. Although the nineteenth-century Spanish criminal codes recognised the garrote as the only legal method of execution, from the eleventh edition onwards the dictionary incorporated the subentries *dar garrote* and *sentenciar á uno á garrote* (s. v. *garrote*), centred on the imposition of this form of execution and thus reflecting its centrality within the codified penal system. However, the twelfth edition continues to register terms related to hanging, such as *ahorcamiento, ahorcado/da* [1], *dogal* [2] and *palo* [5], with juridical meanings, even though this practice had been abolished as a method of execution by the Royal Decree of 24 April 1832. Alongside forms of execution described in the definitions of the lexical units such as *acañaverear* ('to execute with reeds'), *apedrear* [2] ('to stone'), *atenacear* ('to torture with pincers') and *hervencia* ('boiling'), which are explicitly marked as obsolete, these entries reveal the dictionary's conservative tendency to preserve lexical material associated with older punitive practices as part of the historical and cultural heritage of the language. The coexistence of these entries with newer vocabulary related to correction and discipline reveals the overlap between older punitive traditions and modern disciplinary approaches within the legal lexicon of the late nineteenth century.

5 Discussion

The analysis of the criminal-law lexicon in the *DRAE* 1884 confirms the usefulness of approaching historical general-language dictionaries as discursive spaces in which specialised knowledge is reorganized and transmitted to a broader readership. The lexical material examined throughout this study reflects a moment of profound juridical transformation in nineteenth-century Spain, marked by codification, the consolidation of liberal legal structures, and the gradual replacement of the legal pluralism of the Old Regime. At the same time, the dictionary preserves lexical units associated with older juridical realities, revealing the coexistence of inherited legal traditions and emerging conceptual models within the same lexicographical work.

From the perspective of the sociocognitive theory of ideology, these lexicographical choices show how definitions mediate between language and social knowledge. They reflect broader ways of understanding crime, punishment, authority, and legal responsibility in nineteenth-century Spain. The more precise treatment of criminal participation in lexical units such as *autor/ra* or *cómplice*, the emphasis on intentionality in *culpa*, or the attention paid to corruption and abuses committed by public officials correspond to central concerns of liberal criminal law. In Van Dijk's terms (1998), these meanings may be understood as part of socially shared representations through which legal knowledge and institutional authority are organized and transmitted. Definitions therefore do not merely describe lexical meanings, but also contribute to the circulation of broader conceptual models of legal order and criminal responsibility.

The e-*DRAE* (1884) environment strengthens this type of analysis because it allows the researcher to recover legal meanings across lemmas, definitions, sense divisions, labels, and pluriverbal subentries. The platform thus makes it possible to reconstruct a lexical field that would remain only partially visible through a label-based search or through linear consultation of the printed dictionary or the fascimile.

6 Conclusion and future work

The combination of a descriptive lexicographical analysis with a comparative reading of the nineteenth-century criminal codes has made it possible to reconstruct the way in which the penal lexicon enters and is reorganised within a general-language dictionary. The case studies examined show that the *DRAE* 1884 is neither a passive mirror of usage nor a simple repository of legal terminology: it is a discursive space in which older categories inherited from the *Diccionario de autoridades* coexist with definitions that incorporate the conceptual framework of the criminal codes of 1848 and 1870. The twelfth edition emerges, therefore, as a particularly informative witness of the transition from an Old Regime penal culture to the liberal codified order.

Several lines of future research follow from this work. First, the analysis can be extended systematically to other branches of legal vocabulary, especially civil and procedural law, in order to obtain a more complete picture of the juridical lexicon of the *DRAE* 1884. Second, a comparison between the twelfth edition of the academic dictionary and the specialised legal dictionaries published in Spain during the second half of the nineteenth century would contribute to a better understanding of the relationship between academic and juridical lexicography. Third, the continued development of annotations and search functionalities within the e-*DRAE* 1884 platform will make it possible to refine the diachronic study of semantic change and to explore further the ideological dimensions of specialised discourse in historical dictionaries.

More broadly, the integration of philological, lexicographical and digital methodologies exemplified by the e-*DRAE* 1884 opens new perspectives for the diachronic study of specialised lexicon and for the analysis of the relations between dictionaries, social knowledge and the history of legal concepts.

References

Alcaraz Varó, E., Hughes, B. & Gómez, A. (2014). *El español jurídico*. Ariel.

Antón Oneca, J. (1970). El Código penal de 1870. *Anuario de Derecho Penal y Ciencias Penales*, 23, 229–252.

Battaner, P. (1996). La marcación técnica del léxico de especialidad. In E. Forgas (Ed.), *Léxico y diccionarios* (pp. 91–122). Universitat Rovira i Virgili.

Benson, P. (2002). *Ethnocentrism and the English Dictionary*. Routledge.

Blanco Izquierdo, M.ª Á. & Clavería Nadal, G. (Eds.). (2021). *El diccionario académico en la segunda mitad del siglo XIX: evolución y revolución. DRAE 1869, 1884 y 1899*. Peter Lang.

Cañizares Navarro, J. B. (2015). *Las penas infamantes en la codificación penal española* [Doctoral dissertation, Universidad CEU Cardenal Herrera].

Cerezo Mir, J. (2003). *Curso de Derecho Penal español*. Tecnos.

Chen, W. (2019). Towards a discourse approach to critical lexicography. *International Journal of Lexicography*, 32(3), 362–388. https://doi.org/10.1093/ijl/ecz003

Clavería Nadal, G. (2016). *De vacunar a dictaminar. La lexicografía académica decimonónica y el neologismo*. Iberoamericana/Vervuert.

Clavería Nadal, G. (2021a). La lexicografía académica en la segunda mitad del siglo XIX: tradición e innovación (*DRAE* 1869, *DRAE* 1884, *DRAE* 1899). In M.ª Á. Blanco Izquierdo & G. Clavería Nadal (Eds.), *El diccionario académico en la segunda mitad del siglo XIX: evolución y revolución. DRAE* 1869, 1884 y 1899 (pp. 15–56). Peter Lang.

Clavería Nadal, G. (2021b). América en el Diccionario de la lengua castellana de la Real Academia Española (1817–1852). *Lexis*, 45(1), 77–123. https://doi.org/10.18800/lexis.202101.002

Cobo del Rosal, G. (2011). La influencia del Code Pénal francés de 1810 en el Código Penal español de 1822. *Anuario de Derecho Penal y Ciencias Penales*, 64, 285–307.

Código Penal de España. (1848). Imprenta Nacional.

Código Penal de España. (1870). *Gaceta de Madrid*, no. 243, August 31, 1870.

Constitución de la Monarquía Española. (1876). *Gaceta de Madrid*, no. 184, July 2, 1876.

Decreto de 6 de diciembre de 1868, de unificación de fueros. (1868). *Gaceta de Madrid*, no. 342, December 7, 1868.

Diccionario panhispánico del español jurídico. (2023). Real Academia Española/Consejo General del Poder Judicial. https://dpej.rae.es

e-*DRAE 1884. Edición digital, crítica e hipertextual del DRAE 1884*. Retrieved January 2026, from https://e*DRAE*1884.uab.cat/edicion

Freixas Alás, M. (2019). La lexicografía académica de mediados del siglo XIX: el aumento de voces en la novena edición del *DRAE* (1843). In D. Azorín, G. Clavería & E. Jiménez Ríos (Eds.), *ELUA: El diccionario de la Academia y su tiempo: lexicografía, lengua y sociedad en la primera mitad del siglo XIX* (Anexo V, pp. 181–203). https://doi.org/10.14198/ELUA2019.ANEXO5.08

Fuertes-Olivera, P. A. & Tarp, S. (2022). Critical lexicography at work: Reflections and proposals for eliminating gender bias in general dictionaries of Spanish. *Lexikos*, 32(spe), 105–132. https://doi.org/10.5788/32-2-1699

Hornscheidt, A. (2008). A concrete research agenda for critical lexicographic research within critical discourse studies: An investigation into racism/colonialism in monolingual Danish, German and Swedish dictionaries. *Critical Discourse Studies*, 5(2), 107–132. https://doi.org/10.1080/17405900801990066

Iñesta Pastor, E. (2018). Delitos de funcionarios públicos en la codificación penal española del siglo XIX. *Anuario de Historia del Derecho Español*, 88, 451–498.

Ley de Enjuiciamiento Criminal. (1882). Real Decreto de 14 de septiembre de 1882. Gaceta de Madrid, no. 260, September 17, 1882.

López Melero, M. (2012). Evolución de los sistemas penitenciarios y de la ejecución penal. *Anuario de la Facultad de Derecho de la Universidad de Alcalá*, 5, 401–448.

Masferrer, A. & Alvarado Planas, J. (2017). *La codificación penal en España: tradición e influencias extranjeras*. Thomson Reuters Aranzadi.

Mattila, H. E. S. (2012). Legal vocabulary. In P. M. Tiersma & L. M. Solan (Eds.), *The Oxford Handbook of Language and Law* (pp. 27–38). Oxford University Press.

Moon, R. (2014). Meanings, ideologies and learners' dictionaries. In A. Abel, C. Vettori, & N. Ralli (Eds.), *Proceedings of the XVI EURALEX International Congress* (pp. 85–105). EURAC.

Real Academia Española. (1869). *Diccionario de la lengua castellana* (11th ed.). In Nuevo tesoro lexicográfico de la lengua española. Real Academia Española. https://ntlle.rae.es/ntlle/SrvltGUILoginNtlle

Real Academia Española. (1884). *Diccionario de la lengua castellana* (12th ed.). In Nuevo tesoro lexicográfico de la lengua española. Real Academia Española. https://ntlle.rae.es/ntlle/SrvltGUILoginNtlle

Rodríguez Barcia, S. (2012). El análisis ideológico del discurso lexicográfico: una propuesta metodológica aplicada a diccionarios monolingües del español. *Verba. Anuario Galego de Filoloxía*, 39, 135–159.

Rodríguez Ennes, L. (2012). *Historia jurídica de España*. Dykinson.

Rodríguez Ortiz, F. & Garriga Escribano, C. (2010). Las Reglas para la corrección y aumento del diccionario vulgar (1870) y la lexicografía académica del siglo XIX. *Revista de Lexicografía*, 16, 145–166.

Sainz Guerra, J. (2004). *La evolución del derecho penal en España*. Universidad de Jaén.

Tomás y Valiente, F. (2012). *Manual de historia del derecho español*. Tecnos.

Van Dijk, T. A. (1998). *Ideology: A Multidisciplinary Approach*. SAGE.

Contact information

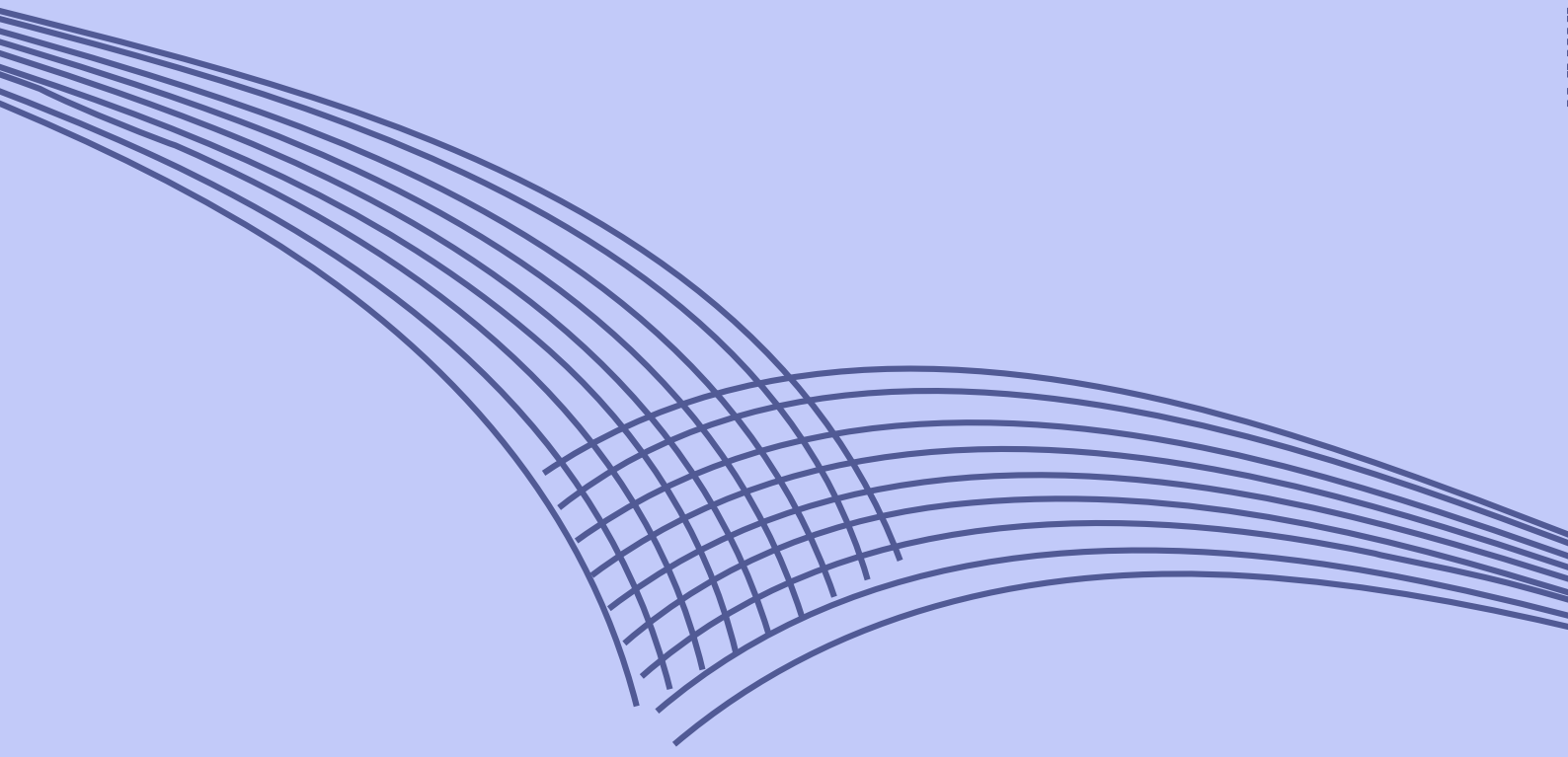
Marija Žarković Eriksson
Universitat Autònoma de Barcelona
marija.zarkovic@autonoma.cat

Acknowledgements

This work was developed within the coordinated project *Modelo de una edición digital e hipertextual del DRAE 1884* (PID2022-136666NB-C21/C22). I would like to thank Prof. Gloria Clavería for her insightful comments.

LEARNER'S LEXICOGRAPHY & DICTIONARY USE

EURALEX 2026



Ida Dringó-Horváth & Katalin P. Márkus

DICTIONARY USE IN A TECHNOLOGY-DRIVEN ERA

Longitudinal Evidence from EFL and GFL Graduates

Abstract This paper presents findings from the third phase of a longitudinal repeated cross-sectional study examining dictionary-use patterns among graduates of English as a Foreign Language (EFL) and German as a Foreign Language (GFL) programmes in Hungary. Drawing on questionnaire data collected in 2020, 2023, and 2026, the study investigates how dictionary use is reshaped in an increasingly digital and AI-mediated reference environment, with particular attention to generative AI-based tools. The results indicate both continuity and change in reference practices. Online dictionaries and search engines remain widely used, but their dominant position is increasingly complemented by machine translation systems and generative AI tools. Print dictionaries continue to be widely owned, yet their frequency of use shows a gradual decline, pointing to a growing gap between ownership and actual use. The 2026 data provide early evidence that chatbots have become part of graduates' language-support repertoires, functioning as multifunctional tools that combine dictionary-like, translation, and text-support functions. The findings suggest a redistribution of reference practices within a hybrid digital environment and highlight the need to reconsider dictionary didactics in light of learners' expanding engagement with AI-based language technologies.

Keywords Dictionary use; dictionary skills; dictionary didactics; longitudinal research; AI-based tools

1 Introduction

Over the past several decades, research on dictionary use has played a central role in shaping modern lexicography by foregrounding users' needs, behaviours, and reference skills (see e.g. Barnhart, 1962; Fajt et al., 2025; Lew, 2011; Lew & De Schryver, 2014; Nesi, 2015; Quirk, 1973). While the insights gained from user studies have contributed substantially to the development of more user-friendly dictionaries, particularly in digital environments, the pedagogical dimension of dictionary use, often referred to as dictionary didactics, has received comparatively less systematic attention (cf. Lew & Galas, 2008; Nesi, 2015; Nied Curcio, 2022). This imbalance is especially evident in formal education, where learners' increasing reliance on digital reference tools is not always matched by structured training in their effective and critical use. The rise of AI-based reference tools further underscores the need for dictionary didactics, as learners' expanding use of generative systems is rarely accompanied by systematic instruction on how to integrate these tools meaningfully with traditional dictionary resources (cf. Lew & De Schryver, 2014; P. Márkus & Dringó-Horváth, 2023; Ptasznik & Lew, 2025).

2 Background to the research

To address these issues, the present study forms part of an ongoing longitudinal research project, employing a repeated cross-sectional design, which investigates dictionary use habits, dictionary skills, and attitudes towards dictionary didactics among graduates of English as a Foreign Language (EFL) and German as a Foreign Language (GFL) programmes in Hungary. The project is designed to capture both continuity and change in dictionary use at a time of rapid technological development, marked by the widespread availability of online dictionaries, search engines, and, more recently, AI-based machine translation tools (for further details on

the project, see P. Márkus et al., 2023; Dringó-Horváth & P. Márkus, 2024). Two phases of data collection have already been completed (in 2020 and 2023), using an identical questionnaire instrument to ensure comparability across cohorts. The findings of these phases provide the empirical and conceptual foundation for the current third phase.

The first phase of the research, conducted in 2020, offered a detailed snapshot of dictionary use and dictionary didactics among university graduates. The results revealed a clear dominance of digital resources: although most participants owned print dictionaries, online dictionaries and search engines were by far the most frequently consulted tools. At the same time, respondents demonstrated a strong reluctance to pay for dictionaries and made only limited use of advanced digital features, such as specialised search techniques, personalisation options, or supplementary materials (P. Márkus et al., 2023). These patterns mirrored trends reported in international dictionary use research and suggested that convenience and accessibility outweighed considerations of lexicographic added value (cf. Kosem et al., 2018; Müller-Spitzer & Koplenig, 2014; Müller-Spitzer et al., 2011).

The second phase of the research, carried out in 2023, replicated the original study with a new cohort of graduates and introduced a comparative perspective that marks the transition towards a longitudinal design. Overall, the findings pointed to a stability in dictionary use habits over time. Online dictionaries and search engines remained the most frequently used resources, while print dictionaries continued to lose practical relevance despite high ownership rates. At the same time, a significant increase was observed in the use of machine translation tools, reflecting broader developments in AI-driven language technologies and their growing integration into everyday language-related tasks. Despite these technological shifts, the use of advanced dictionary features remained consistently low, suggesting that increased exposure to digital tools does not automatically lead to more sophisticated reference behaviour (Dringó-Horváth & P. Márkus, 2024). As in all previous phases of the project, data relating to dictionary didactics and the teaching of dictionary skills were also collected in the present phase; however, these aspects fall outside the scope of the current paper and will be addressed in separate, focused analyses.

Taken together, the results of the first two phases depict a complex picture characterised by both continuity and gradual change: stable preferences for online resources, a rapid rise in machine translation, and a persistent underuse of lexicographic features beyond basic meaning lookup or translation (such as grammatical information, usage notes, or collocations). Building on these findings, the present third phase of the research adopts a longitudinal repeated cross-sectional design to examine changes in dictionary-use patterns across successive cohorts over time, with the aim of contributing empirically grounded insights to dictionary design, dictionary didactics, and teacher education. While the core structure of the previously validated questionnaire was retained to ensure comparability, the instrument was slightly adapted and extended in the third phase to capture developments arising from rapid technological change, particularly the emergence of AI-based language tools. By collecting data from a new cohort of graduates using this updated but methodologically consistent instrument, the study seeks to determine whether the patterns identified so far represent stable characteristics or ongoing shifts. A further aim of the present phase is to document how new AI-based tools are integrated into graduates' everyday reference practices and how these tools coexist with, complement, or partially replace traditional dictionary resources within a hybrid digital reference environment.

Beyond extending the existing dataset with a third measurement point, the present study contributes to dictionary-use research in two important ways. First, it provides one of the earliest longitudinal repeated cross-sectional accounts of how generative AI tools

are becoming integrated into established dictionary and reference practices. Rather than examining dictionaries, machine translation systems, or chatbots as isolated resources, the study conceptualises them as coexisting elements within an increasingly hybrid digital reference environment. Second, by comparing three cohorts across a six-year period, the research makes it possible to distinguish between stable patterns of dictionary use and emerging changes associated with the growing adoption of AI-based language technologies. In this respect, the study moves beyond describing individual tools and contributes to a better understanding of how reference practices are being reorganised in response to technological change. A further aim of the present phase is to document how new AI-based tools are integrated into graduates' everyday reference practices and how these tools coexist with, complement, or partially replace traditional dictionary resources within this evolving hybrid digital reference environment. The findings of the present study also have practical implications, as they provide a basis for adapting the content of lexicology, lexicography, and dictionary-didactics courses offered at Károli Gáspár University of the Reformed Church in Hungary to changing user needs and to the new competencies required by digital and AI-based reference tools.

3 Research into dictionary use in hybrid digital environments

3.1 Aims and participants

The present research is motivated by the need to better understand how dictionary-use practices develop and change over time in response to technological, educational, and pedagogical factors. The study builds on earlier waves of data collection by incorporating the new 2026 survey, which allows dictionary-use patterns to be examined in the context of recent technological change. Particular attention is paid to the impact of artificial intelligence, especially the use of generative chatbots as dictionary-like and translation tools. While the broader research project addresses three interrelated research questions, the present paper focuses exclusively on one of them. Accordingly, the analysis centres on the following question:

RQ: *What patterns can be identified in the dictionary use habits of EFL and GFL university graduates, with particular regard to the use of AI-based tools?*

As indicated above, a longitudinal repeated cross-sectional design is employed, drawing on samples collected in 2020, 2023, and 2026 from graduates of EFL and GFL programmes in Hungary. The present data collection represents the third phase of the project, which is designed to continue with further measurement waves at three-year intervals in the future, allowing for the systematic examination of changes in dictionary-use patterns over time at the cohort level. This design allows us to examine changes in dictionary-use patterns over time at the group level, while it does not support claims about individual developmental trajectories. As this longitudinal research is continued and further data collection phases are completed, variables that show a consistent shift in the same direction across successive measurement waves may be subjected to trend analysis, following the approach described by Müller-Spitzer & Wolfer (2024, p. 266). The table below (Table 1) displays the participants' personal characteristics.

TABLE 1 Personal characteristics of participants (2020, 2023, and 2026)

Characteristic		2020 (n=197)	2023 (n=110)	2026 (n=103)
Age	**Range**	21 – 63	20 – 58	20 – 59
	M (SD)	34.00 (11.00)	31.00 (11.00)	33.00 (13.00)
Gender, %	**Female**	85%	84%	79%
	Male	15%	16%	21%
Foreign language teaching experience	**Participants with at least one year of experience, n (%)**	149 (75.6%)	64 (58.2%)	58 (56.3%)
	Years of experience, M (SD)	12.00 (9.00)	9.00 (9.00)	12.00 (10.00)

Note. M = mean; SD = standard deviation. Percentages for teaching experience were calculated using the total number of respondents in each survey year as the denominator

Based on the data provided, the samples are homogeneous in terms of gender distribution ($\chi^2 = 2.153$, df = 2, p = .341) but show heterogeneity in terms of age (Kruskal-Wallis $p < .001$) and teaching experience (Kruskal-Wallis $p < .001$). The average age exceeded 30 years in all samples (2020: M = 34.00, SD = 11.00; 2023: M = 31.00, SD = 11.00; 2026: M = 33.00, SD = 13.00), largely due to the inclusion of postgraduate, practising (in-service) teachers who were enrolled in part-time postgraduate programmes to obtain a higher-level qualification. Teaching experience was also widespread, with the 2020 sample being older and more experienced on average, compared to the 2023 and 2026 samples. These differences may influence findings related to dictionary use and the teaching of dictionary skills and therefore represent a potential limitation. Future research should consider controlling for age and teaching experience to support more robust comparisons across sampling periods. All participants had a target language proficiency of C1 according to the CEFR (the Common European Framework of Reference for Languages), in line with national Training and Output Requirements. Foreign language teacher trainees receive uniform training in language pedagogy and lexicography across specialisations, ensuring a comparable educational background in dictionary use, in accordance with Hungary's Training and Output Requirements.

3.2 The research instrument

The questionnaire draws on both Hungarian and international research traditions in studies of dictionary use. It is grounded in established Hungarian empirical work (e. g. Dringó-Horváth, 2017; Gaál, 2016; Márkus & Szöllősy, 2006) and informed by widely used international methodological frameworks, particularly those developed at the Institute for the German Language in Mannheim (Müller-Spitzer et al., 2011, 2012; Müller-Spitzer & Koplenig, 2014). The questionnaire also relates to research efforts aimed at developing and validating standardised instruments for measuring dictionary-use behaviour and dictionary-use skills. In particular, Gavriilidou (2013) developed and validated the Strategy Inventory for Dictionary Use (S.I.D.U.), while Mavrommatidou et al. (2019) developed and validated the Strategy Inventory for Electronic Dictionary Use (S.I.E.D.U.), two standardised instruments designed to assess learners' dictionary-use strategies and skills. While the present study does not adopt these instruments directly, its focus on dictionary-use behaviour, reference strategies, and user awareness is conceptually aligned with this line of research.

At the same time, the scope of the present questionnaire is broader, as it examines not only dictionary-use skills and strategies but also attitudes towards dictionary didactics, teaching practices, and the integration of emerging digital and AI-based reference tools. The inclusion of these dimensions makes it possible to investigate dictionary use within a wider educational and technological context. To ensure direct comparability, the same questionnaire was employed in all three studies, with only minor modifications in the third phase to reflect recent technological developments, which are described in more detail below. The questionnaire instrument is described in detail in P. Márkus et al. (2023) and is publicly available for download. It comprises 21 background questions and a total of 69 attitudinal statements. Of these, 16 items address the types of dictionaries and digital resources used by participants, while 53 items are organised into 10 scales that capture different aspects of dictionary use, habits, and attitudes. The first four scales (USAGE1–USAGE4) focus on the frequency and functional dimensions of dictionary use, and scales 5–6 (ATTITUDE1–ATTITUDE2) examine general attitudes towards dictionary use. Scales 7–10 (TEACHING1–TEACHING4) address attitudes towards the teaching of effective dictionary use. While the full instrument allows for a comprehensive analysis of dictionary use and dictionary didactics, the present conference paper draws selectively on those parts of the questionnaire that are relevant to the current research focus, namely patterns of dictionary use among EFL and GFL graduates, with particular attention to the use of AI-based tools. All responses were measured on a five-point Likert scale (Agreement: 1=Strongly disagree, 5=Strongly agree; Frequency: 1=Never, 5=Regularly), enabling quantitative analysis of usage patterns across cohorts.

3.3 Data collection and analysis

Data collection was conducted in 2020, 2023, and 2026, each time during the final weeks of the academic year, using an online questionnaire (Microsoft Forms). As for the inclusion criteria, all final-year students enrolled in the selected programmes in the year of the study were invited to participate. Access details were distributed centrally via the university's academic database on three occasions. In the 2026 wave, a total of 185 students from the English and German programmes were invited to participate, resulting in a response rate of 55.7% (N = 103), which is considered satisfactory for an online questionnaire survey in a higher-education context. To complement the questionnaire data, 18 semi-structured interviews were also conducted with participants. These interviews provide deeper insight into users' reference practices, decision-making processes, and perceptions of changes in dictionary and AI-supported tool use, thereby helping to contextualise and interpret the quantitative findings. The analysis of the interview data falls outside the scope of the present paper and will be reported in a separate publication.

The data were coded following collection, and all reversed items were recoded to ensure consistent scale directionality within each construct. Reversed items were included at the questionnaire design stage to enhance response validity and attentiveness and to allow post hoc consistency checks after recoding. For nominal-level variables, proportions were calculated. Differences across the three data collection waves were tested using Pearson's chi-square test. Effect sizes for chi-square analyses were estimated using Cramér's V. When global comparisons indicated significant differences, Bonferroni-adjusted pairwise post hoc comparisons were conducted between individual years. For quasi-interval variables, means (M) and standard deviations (SD) were computed. As assumptions of normality were violated, group comparisons were performed using the Kruskal–Wallis test, followed by Bonferroni-corrected Mann–Whitney post hoc tests where appropriate. Effect sizes for Kruskal–Wallis analyses were estimated using eta-squared ($\eta^2 = [H - k + 1]/[N - k]$). To examine the potential influence of

demographic characteristics on selected significant results, supplementary binary logistic regression and ANCOVA analyses were conducted with age and teaching experience included as control variables. Spearman's rank correlation coefficient was applied for correlation analyses. All statistical analyses were carried out using SPSS version 27.0. Data handling complied fully with GDPR regulations, and access to the dataset was restricted to members of the research project. 3.4 Reliability of the scales

As the analyses presented in this paper focus on the USAGE scales, the internal consistency of these multi-item scales was assessed using Cronbach's alpha coefficients (Dörnyei & Taguchi, 2010). Most scales showed acceptable reliability, with alpha values exceeding the recommended threshold of .60 (see Table 2). The only exception was the USAGE2 scale, which displayed lower reliability in earlier phases of the project. In the 2026 survey, however, this scale was no longer included because the search strategies it measured have become increasingly integrated into digital dictionary interfaces and are often performed automatically without users' conscious awareness. In addition, two items were removed from the USAGE1 scale as they were no longer considered relevant in contemporary digital dictionary use. The reliability coefficients presented in Table 2 indicate that the scales analysed in the present study meet accepted standards of internal consistency (Dörnyei & Taguchi, 2010).

TABLE 2 Internal consistency of the USAGE multi-item scales

Scale		Nr. of items	α (2020)	α (2023)	α (2026)
USAGE1	Use of unique features relating to digital dictionaries	6	.647	.692	-
USAGE1 (reduced)*	Use of unique features relating to digital dictionaries	4	-	-	.715
USAGE2**	Use of search methods relating to digital dictionaries	5	.590	.607	-
USAGE3	Willingness to pay	4	.643	.628	.675
USAGE4	Conscious use of the prefatory material in dictionaries	4	.862	.800	.788

* In 2026, two of the six items were removed.
** In 2026, the scale was no longer used.

4 Results and discussion

4.1 Ownership and use of dictionaries and digital reference tools

First, respondents were asked about the types of dictionaries they owned. The results of these items are presented below (see Table 3). All conclusions and trends identified across the three measurements should be interpreted strictly within the scope of the respective samples and do not imply generalizable patterns beyond the studied population.

TABLE 3 Types of dictionaries owned by respondents (in order of frequency based on the 2026 survey)

Questionnaire item	% 2020	% 2023	% 2026	χ^2	df	*p*	*V*
print dictionary	95%[a]	88%[a]	89%[a]	6.299	2	.043*	.124
offline dictionary application on some smart device	53%	47%	51%	.873	2	.646	.270
machine translation software subscription	9%[a]	31%[b]	31%[b]	29.992	2	< .001*	.270
offline digital dictionary on their computer	27%	21%	22%	1.633	2	.442	.063
online dictionary subscription	9%	9%	6%	.939	2	.625	.048

*significant change at α = .05
a/b There is no significant difference between values sharing the same superscript, based on Bonferroni-adjusted pairwise Mann–Whitney comparisons (p < .017)
V = Cramér's V effect size.

Most participants reported owning a print dictionary across all three measurement points. A decline is observable from 95% in 2020 to 88% in 2023, followed by a slight increase to 89% in 2026, suggesting a period of stabilisation rather than a continued decrease. Overall, the pattern points to a gradual long-term shift away from traditional print resources. Although the global chi-square test indicates a significant overall difference ($\chi^2 = 6.299$, df = 2, p = .043), Bonferroni-adjusted pairwise comparisons did not reveal significant differences between individual years. This trend, therefore, warrants further monitoring. Across all three measurements, offline dictionary applications on smart devices were consistently more widely used (roughly twice as high) than offline digital dictionaries installed on computers, while both formats showed stable patterns over time without significant change. This indicates a clear preference for mobile-based offline solutions, with computer-based offline dictionaries remaining a more limited but steady option. Reported ownership of online dictionary subscriptions remains low and largely unchanged (9% in 2020 and 2023, 6% in 2026), with no significant difference across years. This indicates that paid online dictionary services have not become more widespread among respondents despite broader digitalisation trends. The consistent but low willingness to pay for such services corroborates findings from other studies suggesting that generally, there is a reluctance to invest money into dictionaries, including online resources (cf. Gaál, 2016; Lew, 2016; Nied Curcio, 2022).

In contrast, ownership of machine translation software shows a pronounced and statistically significant increase, rising from 9% in 2020 to 31% in both 2023 and 2026 ($\chi^2 = 29.992$, p < .001). Bonferroni-adjusted post hoc comparisons confirm that the 2020 value differs significantly from those of the later samples, while no difference is observed between 2023 and 2026. This pattern suggests a rapid shift that has stabilised at a substantially higher level of adoption, indicating that respondents have developed a growing preference for automated translation options. Additional binary logistic regression analyses controlling for age and teaching experience confirmed that survey wave remained a significant predictor of machine translation software ownership, whereas neither age nor teaching experience showed a significant effect. However, it is important to note that ownership of a given resource (whether a print dictionary, a digital application, or a subscription-based reference tool) does not necessarily entail regular or effective use (see Table 4). More broadly, access to or possession of reference resources does not always correspond to actual consultation practices, nor does it reliably reflect users' stated preferences or habitual behaviour, which is a discrepancy that has been widely documented in dictionary-use research (cf. Kosem et al., 2018). This pattern

is clearly reflected in the 2026 data: although print dictionaries are still owned by a large majority of respondents (89%; see Table 3), their mean frequency of use has further declined (M = 2.08; see Table 4). At the same time, digital resources dominate everyday practice, with online dictionaries (M = 3.99), search engines (M = 3.92), machine translation software (M = 3.19), and chatbot-based tools (M = 2.80) used more frequently, reflecting an ongoing change of dictionary use in favour of readily accessible digital and AI–supported tools. Together, these findings point to a growing gap between ownership and use and underscore an ongoing transition of reference practices toward flexible, multifunctional digital tools, in line with broader shifts in digital information consumption (cf. Baron, 2021; Liu et al., 2025; Ptasznik & Lew, 2025). Data presented in Table 4 reveal clear shifts in the frequency of use of different dictionary types across the three data collection points.

TABLE 4 Use of different applications as a dictionary or translator (in order of frequency based on the 2026 survey)

Questionnaire item	Mean 2020	SD 2020	Mean 2023	SD 2023	Mean 2026	SD 2026	Kruskal-Wallis p	η^2
online dictionaries	4.63[a]	0.78	4.57[a]	0.93	3.99[b]	1.29	<.001*	.065
search engines (e.g. Google)	3.82	1.20	3.93	1.15	3.92	1.20	.690	.000
machine translation software (e.g. DeepL Translate)	2.53[a]	1.53	3.42[b]	1.41	3.19[b]	1.45	<.001*	.064
chatbot (e.g. ChatGPT)	-	-	-	-	2.80	1.39	-	-
offline dictionary application on some smart device	2.44	1.50	2.29	1.47	2.28	1.52	.557	.000
print dictionaries	2.68[a]	1.35	2.31[ab]	1.19	2.08[b]	1.19	<.001*	.036
offline digital dictionary on their computer	1.76	1.25	1.85	1.34	1.71	1.19	.639	.000

*significant change at alpha = 0.05
a/b There is no significant difference between values sharing the same superscript, based on Bonferroni-adjusted pairwise Mann–Whitney comparisons ($p < .017$)
η^2 = effect size estimate based on the Kruskal–Wallis statistic.

Online dictionaries and search engines remain the most frequently used tools across all three measurement points. For online dictionaries, no significant difference is observed between 2020 and 2023, but a significantly lower mean is recorded in 2026 (M = 3.99). This decrease is noteworthy, as it may indicate a redistribution of digital consultation practices, possibly due to the use of new possibilities, such as chatbots or AI-driven translation software, as it is described below. This interpretation is also supported by the interview data, which suggest that participants increasingly combine traditional dictionary resources with chatbots and AI-based translation tools for everyday language-related tasks. Search engines, on the other hand, show consistently high and stable mean values across all years, with no significant differences ($p = .690$), underscoring their continued importance as general-purpose lexical resources. These tools continue to dominate users' reference practices, which is in line with earlier findings and previous research (cf. Gaál, 2016; Lew, 2015; Müller-Spitzer et al., 2012; Nied Curcio, 2015; P. Márkus et al., 2023; Reder, 2016; Töpel, 2015). The use of machine translation software shows the most pronounced change. Mean values increase sharply from 2020 to 2023 (from M = 2.53 to M = 3.42), a difference confirmed as statistically significant, while no

further significant change is found between 2023 and 2026 (M = 3.19). This pattern suggests a rapid adoption phase followed by stabilization, indicating that machine translation tools have become an established component of respondents' language support repertoire. By contrast, the use of offline dictionary apps and offline computer-based dictionaries remains largely unchanged across the three time points, indicating that these tools are used occasionally for specific purposes rather than as primary reference resources.

Although print dictionaries continue to be widely known and used, their mean frequency of use shows a gradual decline, with only the difference between 2020 and 2026 reaching statistical significance, suggesting a slow but cumulative shift away from print dictionaries rather than an abrupt change (cf. Ptasznik & Lew, 2025; Sycz-Opoń, 2024). However, supplementary ANCOVA analyses controlling for age and teaching experience indicate that this effect is sensitive to differences in sample composition, as the survey-wave effect for print dictionary use was no longer statistically significant after adjustment. The observed decline should therefore be interpreted with caution and may partly reflect cohort differences rather than a purely time-related trend.

4.2 Chatbots and translation applications

Following the emergence of large language model-based chatbots in late 2022 and their rapidly increasing use in foreign-language contexts, including translation, dictionary consultation, and text styling, the questionnaire was extended to include items targeting these tools as dictionary and/or translation resources. The 2026 data thus introduce chatbot use as a distinct category, revealing a relatively high mean frequency of use (M = 2.80). Although no longitudinal comparison is possible at this stage, the inclusion of chatbots points to an emerging trend that calls for systematic investigation across future cohorts as part of the repeated cross-sectional research design.

Accordingly, the use of individual chatbots and translation applications is examined in more detail below (Table 5). ChatGPT emerges as the most frequently used generative AI tool (M = 2.93, SD = 1.37), while Gemini, Microsoft Copilot, and Grok follow far behind and remain marginal in comparison. As for the translator tools, Google Translate stands out (M = 2.83, SD = 1.48), while DeepL and QuillBot show a more moderate level of use. This distribution suggests that chatbot use in 2026 is not evenly spread across platforms, but rather concentrated around one dominant tool, and that generative AI tools are already competing directly with established machine translation solutions (e. g. DeepL and Google Translate).

TABLE 5 Use of specific software and chatbots for dictionary- and translation-related purposes (ordered by frequency, N=103)

Questionnaire item	Mean 2026	SD 2026
ChatGPT	2.93	1.37
Google Translate	2.83	1.48
DeepL	2.11	1.52
Gemini	1.80	1.25
QuillBot	1.27	0.76
Microsoft Copilot	1.21	0.65
Grok	1.04	0.14

AI-driven translation systems now support a wide range of tasks, including text and speech translation, contextual interpretation, terminology support, and adaptation to cultural nuance. Integrated into web platforms, mobile applications, and professional translation tools, these systems have made high-quality translation more accessible, efficient, and cost-effective. Consequently, the user base of AI translation software has expanded from professional translators to include students, researchers, businesses, and casual users (cf. Huang et al., 2025; Liu et al., 2025; Ruoqi et al., 2023). Recent research suggests that they may produce more fluent, context-sensitive, and culturally adaptive translations than conventional machine translation tools, particularly for rhetorically or culturally complex texts and when appropriate prompts are used (e. g. Chen & Lin, 2025). However, these systems may also introduce semantic shifts or over-adapt the source text, which makes human review necessary in academic and professional contexts. Findings support a hybrid approach, combining machine translation tools with systematic human oversight to monitor and correct output, ensuring optimal translation quality, especially in languages with complex morphology and culturally embedded expressions (cf. Mohammed, 2025).

Taken together, these findings indicate that digital language-support practices are becoming more diverse, with translation tools and AI-based applications increasingly used alongside traditional dictionaries. By combining dictionary, translation, and text-support functions within a single interface, chatbots may help account for both the slight decline in online dictionary use observed in 2026 and the stabilisation of machine-translation use. Rather than signalling the abandonment of earlier resources, the results point to a gradual reorganisation of reference practices as users adapt to a changing digital environment.

5 Conclusions

This study examined patterns of dictionary use among EFL and GFL graduates in Hungary within a longitudinal repeated cross-sectional framework, with particular attention to the growing role of AI-based tools. The findings reveal both continuity and change in reference practices. While online dictionaries and search engines remain central to everyday language support, their dominant position is increasingly complemented by machine translation systems and generative AI-based chatbots. The 2026 data provide early evidence that chatbots have become part of respondents' language-support repertoire, functioning as multifunctional tools that integrate dictionary-like, translation, and text-production functions within a single interface. Rather than replacing traditional resources outright, these tools appear to contribute to a redistribution of reference practices within a hybrid digital environment. In an international perspective, these findings align closely with results reported in dictionary-use research from other educational contexts, which likewise document the dominance of digital resources, the growing reliance on machine translation, and the increasing use of AI-based language tools alongside dictionaries (cf. Lew & De Schryver, 2014; Liu et al., 2025; Müller-Spitzer & Koplenig, 2014).

The Hungarian data thus confirm broader international trends while also illustrating how rapidly emerging AI technologies may accelerate ongoing shifts in reference behaviour. More importantly, the findings suggest that contemporary dictionary use is increasingly embedded in a hybrid digital reference environment, in which traditional dictionaries coexist with machine translation systems and generative AI applications. From a theoretical perspective, this indicates that dictionary use can no longer be examined exclusively within the boundaries of lexicographic resources but must be understood as part of a broader network of interconnected digital reference tools. At the same time, these developments underscore the need to reconsider dictionary didactics in response to changing reference practices and to address learners' growing reliance on AI-mediated tools through targeted, evidence-based pedagogical

guidance (cf. Chi, 2020; Dringó-Horváth & P. Márkus, 2024; Law, 2024; P. Márkus, 2023; Rundell, 2016).

Limitations

Several limitations should be taken into account when interpreting the findings. First, the study follows a longitudinal repeated cross-sectional design rather than a true panel design; as a result, changes can be identified at the cohort level but not at the level of individual developmental trajectories. Second, the data are based on self-reported questionnaire responses, which may not fully reflect actual dictionary-use behaviour in practice. Third, the sample is limited to graduates of EFL and GFL programmes at a single Hungarian university. Although the participants share educational characteristics typical of Hungarian foreign-language programmes, the sample cannot be considered representative of all Hungarian EFL and GFL graduates. Consequently, the findings should be interpreted primarily within the context of the participating institution and regarded as indicative rather than fully generalisable to the national level. Finally, chatbot use was examined only at one measurement point (in 2026), making it impossible to assess longitudinal change in this area at present. Future rounds of data collection will be necessary to determine whether the observed patterns represent stable developments or transitional phenomena.

References

Barnhart, C. (1962). Problems in editing commercial monolingual dictionaries. In F. W. Householder & S. Saporta (Eds.), *Problems in Lexicography*. Indiana Research Centre for Language and Semiotic Studies. Reprinted in R.R.K. Hartmann (Ed.) (2003). *Lexicography: Critical Concepts*. Volume 1: Dictionaries, compilers, critics, and users (pp. 285–301). Routledge.

Baron, N. S. (2021). *How we read now: strategic choices for print, screen, and audio*. Oxford University Press.

Chen, S. & Lin, Y. (2025). A multidimensional comparison of ChatGPT, Google Translate, and DeepL in Chinese tourism texts translation: fidelity, fluency, cultural sensitivity, and persuasiveness. *Frontiers in Artificial Intelligence*, 8, 1619489. https://doi.org/10.3389/frai.2025.1619489

Dringó-Horváth, I. (2017). Digitális szótárak — szótárdidaktika és szótárhasználati szokások. [Digital Dictionaries – Dictionary Didactics and Dictionary Use Habits.] *Alkalmazott Nyelvtudomány*, 17(5), 1–27.

Dringó-Horváth, I. & P. Márkus, K. (2024). Using dictionaries in teaching (and learning) English as a Foreign Language – the beginning of a longitudinal research project. In A. Klosa-Kückelhaus & M. Nied Curcio (Eds.), *Dictionary Use and Dictionary Teaching: New Challenges in a Multilingual, Digital and Global World* (pp. 9–28). De Gruyter. https://doi.org/10.1515/9783111373294-002

Fajt, B., Bánhegyi, M. & P. Márkus, K. (2025). A Systematic Literature Review of Over Three Decades of Dictionary Use Research in Lexikos. *Lexikos*, 35(2), 423–442. https://doi.org/10.5788/35-2-2105

Gaál, P. (2016). *Online szótárak és használóik*. [*Online dictionaries and their users*.] University of Pécs.

Gavriilidou, Z. (2013). Development and Validation of the Strategy Inventory for Dictionary Use (S.I.D.U.). *International Journal of Lexicography*, 26(2), 135–153. https://doi.org/10.1093/ijl/ect007

Huang, Y., Li, D., & Cheung, A. K. F. (2025). Evaluating the linguistic complexity of machine translation and LLMs for EFL/ESL applications: An entropy weight method. *Research Methods in Applied Linguistics*, 4(3), 100229. https://doi.org/10.1016/j.rmal.2025.100229

Kosem, I., Lew, R., Müller-Spitzer, C., Ribeiro Silveira, M. & Wolfer, S. (2018). The image of the monolingual dictionary across Europe. Results of the European survey of dictionary use and culture. *International Journal of Lexicography*, 32(1), 92–114.

Law, Wai-on (2024). Dictionary Use Training in Secondary School EFL Textbooks in Taiwan. *Lexikos*, 34(1), 77–98. https://doi.org/10.5788/34-1-1880

Lew, R. (2011). Studies in dictionary use: recent developments. *International Journal of Lexicography*, 24(1), 1–4.

Lew, R. & De Schryver, G-M. (2014). Dictionary users in the digital revolution. *International Journal of Lexicography*, 27(4), 341–359.

Lew, R. & Galas, K. (2008). Can Dictionary Use be Taught? The Effectiveness of Lexicographic Training for Primary School Level Polish Learners of English. In E. Bernal & J. DeCesaris (Eds.), *Proceedings of the XIII EURALEX International Congress* (pp. 1273–1285). Universitat Pompeu Fabra.

Lew, R. (2015). Research into the Use of Online Dictionaries. *International Journal of Lexicography*, 28(2), 232–253.

Lew, R. (2016). Dictionaries for Learners of English. *Language Teaching*, 49(2), 291–294. https://doi.org/10.1017/S026144481500049X

Liu, R., Chen, X. & Xu, Y. (2025). Beyond Replacement: How Large Language Models Influence Dictionary Usage Patterns Among Chinese English Learners. *International Journal of Lexicography*, 38(4), 342–364. https://doi.org/10.1093/ijl/ecaf017

Mavrommatidou, S., Gavriilidou, Z. & Markos, A. (2019). Development and Validation of the Strategy Inventory for Electronic Dictionary Use (S.I.E.D.U.). *International Journal of Lexicography*, 32(4), 393–410. https://doi.org/10.1093/ijl/ecz015

Márkus, K. & Szöllősy, É. (2006). Angolul tanuló középiskolásaink szótárhasználati szokásairól. [The Dictionary Use Habits of Hungarian Secondary School Students Learning English as a Foreign Language.] In T. Magay (Ed.), *Szótárak és használóik*. [*Dictionaries and their Users*] (pp. 95–116). Akadémiai Kiadó.

Mohammed, T. A. S. (2025). Evaluating Translation Quality: A Qualitative and Quantitative Assessment of Machine and LLM-Driven Arabic–English Translations. *Information*, 16(6), 440. https://doi.org/10.3390/info16060440

Müller-Spitzer, C. & Koplenig, A. (2014). Online Dictionaries: Expectations and Demands. In C. Müller-Spitzer (Ed.), *Using Online Dictionaries* (pp. 143–188). Lexicographica Series Maior 145. Walter de Gruyter.

Müller-Spitzer, C., Koplenig, A. & Töpel, A. (2011). What makes a good online dictionary? Empirical insights from an interdisciplinary research project. In I. Kosem & K. Kosem (Eds.), *Electronic lexicography in the 21st century: New applications for new users. Proceedings of eLEX2011* (pp. 203–208). Institute for Applied Slovene Studies.

Müller-Spitzer, C. & Wolfer, S. (2024). 9 Research into Dictionary Use. In A. Klosa-Kückelhaus (Ed.), *Internet Lexicography: An Introduction* (pp. 261–304). De Gruyter. https://doi.org/10.1515/9783111233758-010

Nesi, H. (2015). Thirty Years of User Studies– And what we still need to find out. In L. Li, J. McKeown & L. Liu (Eds.), *Words, Dictionaries and Corpora: Innovations in reference science: Proceedings of ASIALEX 2015* (pp. 1–8). Hong Kong Polytechnic University.

Nied Curcio, M. (2022). Dictionaries, foreign language learners and teachers. New challenges in the digital era. In A. Klosa-Kückelhaus, S. Engelberg, Ch. Möhrs & P. Storjohann (Eds.), *Dictionaries and society: Proceedings of the XX EURALEX International Congress* (pp. 71–84). IDS-Verlag.

P. Márkus, K. & Dringó-Horváth, I. (2023). Developing Dictionary Skills through Monolingual and Bilingual English Dictionaries at Tertiary-level Education in Hungary. *Lexikos*, 33(1), 350–381.

P. Márkus, K. (2023). *Teaching Dictionary Skills*. L'Harmattan.

P. Márkus, K., Fajt, B. & Dringó-Horváth, I. (2023). Dictionary Skills in Teaching English and German as a Foreign Language in Hungary: A Questionnaire Study. *International Journal of Lexicography*, 36(2), 173–194. https://doi.org/10.1093/ijl/ecad004

Ptasznik, B. & Lew, R. (2025). Dictionaries versus AI Tools through the Eyes of English Majors. *International Journal of Lexicography*, 38(2), 140–158. https://doi.org/10.1093/ijl/ecaf005

Quirk, R. (1973). The social impact of dictionaries in the UK. In R.I. McDavid & A.R. Duckert (Eds.), *Lexicography in English* (pp. 76–88). Academy of Sciences. Reprinted in R.R.K. Hartmann (2003) (Ed.), *Lexicography: Critical Concepts. Volume 1: Dictionaries, compilers, critics, and users* (pp. 312–326). Routledge.

Reder, A. (2016). Wörterbücher und Digitale Tools als Ratgeber bei der Kollokatorsuche. *Beiträge zur Fremdsprachenvermittlung* 58, 18–30. http://www.vep-landau.de/fileadmin/user_upload/bzf/Hefte/bzf_2016_58.pdf

Rundell, M. (2016). Dictionaries and crowdsourcing, wikis and user-generated content. In P. Hanks & G-M. De Schryver (Eds.), *International Handbook of Modern Lexis and Lexicography*. Springer-Verlag.

Ruoqi, S., Yuan, X. & Gochuico, M. T. (2023). Translators and Their Use of ChatGPT. *Academia Lasalliana Journal of Education and Humanities*, 5(1), 49–59. https://doi.org/10.55902/plru9197

Sycz-Opoń, J. (2024). Print Dictionaries Are Still in Use: A Survey of Source Preferences by Polish Translators. *International Journal of Lexicography*, 37(3), 301–321. https://doi.org/10.1093/ijl/ecae004

Töpel, A. (2015). Das Wörterbuch Ist Tot – Es Lebe das Wörterbuch? *Info DaF* 5, 515–534. https://www.doi.org/10.1515/infodaf-2015-0508

Contact information

Ida Dringó-Horváth
Károli Gáspár University of the Reformed Church in Hungary
dringo.horvath.ida@kre.hu

Katalin P. Márkus
Károli Gáspár University of the Reformed Church in Hungary
p.markus.kata@kre.hu

Acknowledgements

This research was supported by the project *"Oktatásinformatika a felsőoktatásban III."* (Project ID: 20794B800) – Károli Gáspár University of the Reformed Church in Hungary.

Christine Ganslmayer & Martina Nied Curcio

HOW GERMAN LANGUAGE TEXTBOOKS ADDRESS THE DIGITAL TRANSFORMATION OF LEXICOGRAPHY

Abstract The digital transformation of lexicography has fundamentally changed lexicographic resources available for language learning and teaching. However, their integration into German language teaching – both as a first language (GL1) and as a foreign language (GFL) – remains limited. This paper investigates how the digital transformation of lexicography is reflected in contemporary German language textbooks through a qualitative and quantitative analysis (GFL: 2011–2024; GL1: 2003–2021). Drawing on a comparative textbook corpus, the study combines quantitative frequency analysis with qualitative content analysis of textbook tasks and their didactic functions. The findings show that lexicographic resources are underrepresented, predominantly conceptualised in traditional (print-based) ways, and rarely embedded in meaningful communicative or strategy-oriented tasks. Online lexicographic information systems, hybrid resources, and AI-supported language tools are almost entirely absent from the analysed textbooks. Furthermore, the development of lexicographic competence and media literacy is largely neglected. The paper argues that textbooks do not adequately reflect the digital transformation of lexicography and calls for a reconceptualisation of dictionary didactics that systematically integrates contemporary lexicographic information systems while also preparing learners to use AI-supported language tools critically and appropriately. In doing so, it contributes to current discussions on the digital transformation of lexicography and its implications for language education.

Keywords Digital transformation of lexicography; lexicographic competence; digital lexicography; dictionary use; language didactics; German as a Foreign Language (GFL); German as a First Language (GL1); media literacy

1 Introduction and research questions

Dictionaries have always been an aid in (foreign) Language Teaching with the aim of acquiring, facilitating and improving language skills (cf. Wiegand et al., 2010). Today, however, the lexicographic landscape has changed fundamentally. Traditional printed dictionaries have increasingly evolved into digital lexicographic information systems that integrate lexical data with corpora, grammar resources, translation functions and, more recently, AI-supported language tools. This transformation has fundamentally changed not only the resources available to users but also the ways in which lexical information is accessed, processed and evaluated. In Foreign Language Teaching (FLT) with the communicative method in the 1970s/1980s, dictionaries (especially bilingual ones) were often banned from FLT classrooms. Research into dictionary use shows that learners used dictionaries at all times, even during the communicative method, regardless of whether or not working with dictionaries was actually encouraged in class. The consequence, however, is that learners are left to their own devices to decide which dictionary to use in which learning phase, for which purpose and in which way it can be used successfully – which does not necessarily lead to satisfactory results.

As a result, dictionary use in secondary education remains largely restricted to orthographic purposes, while the broader potential of dictionaries for language reflection, vocabulary development, text production, and the promotion of textual literacy remains underused. This limited and largely instrumental approach also makes it difficult to establish a sustainable, lifelong relationship between learners and lexicographic resources. However, the current competence-oriented approach to teaching is bringing dictionaries back into focus

(cf. Haß, 2019; Merten, 2022b; Ganslmayer & Nied, 2026a, pp. 4–5; Ganslmayer & Nied, 2026b).

The positive effects of dictionary work in native language teaching are undisputed among experts, but in practice, neither has a robust dictionary culture developed (cf. Kühn, 1989, 1999a) nor have the suggestions of dictionary didactics (e. g. Kühn, 1994, 1999b, 2011; Merten, 2022a) been widely taken up within the teaching community. From a lexicographic perspective, this gap is particularly problematic because successful dictionary consultation depends not only on the availability of high-quality resources but also on users' ability to select, interpret and critically evaluate lexicographic information. In German language education, dictionary skills acquired in primary school are often tacitly assumed rather than systematically developed further in an age- and proficiency-appropriate manner. As a result, dictionary use in German lessons is still predominantly associated with orthographic purposes, while the broader potential of dictionaries for language reflection, vocabulary development, text production, and the promotion of textual literacy remains underused. This limited and largely instrumental approach also makes it difficult to establish a sustainable, lifelong relationship between learners and lexicographic resources. However, the current competence-oriented approach to teaching is bringing dictionaries back into focus (cf. Haß, 2019; Merten, 2022b; Ganslmayer & Nied, 2026a, pp. 4–5; Ganslmayer & Nied, 2026b).

With increasing digitisation, dictionaries as a medium have changed fundamentally: the traditional structure of dictionaries has been dissolved, and various hybrid formats have emerged, such as dictionary + grammar table, dictionary + grammar table + parallel text, or even dictionary + parallel text + automatic translator resources, which were physically separate before the digital age (cf. Müller-Spitzer et al., 2018). Moreover, digitisation has given rise to new lexicographic formats, including online glossaries, corpus-based dictionaries, dictionary portals, hybrid language tools, resources incorporating parallel corpora and translation functions, and, more recently, AI-based language resources. As a result, dictionaries have evolved beyond their traditional form into broader digital lexicographic information systems that provide integrated access to linguistic knowledge and language-support tools (Ganslmayer & Nied, 2026a, p. 3). Consequently, the object of dictionary didactics can no longer be limited to dictionaries in the traditional sense but must encompass the competent use of digital lexicographic information systems and related lexical language technologies. Rather than replacing dictionaries, AI-supported language tools complement existing lexicographic information systems by providing interactive, context-sensitive access to lexical information. This development raises new questions for lexicographic competence, as users must understand the strengths, limitations and appropriate use of different types of lexicographic and AI-supported resources.

Recent research therefore argues that that dictionary didactics should move beyond teaching dictionary consultation alone towards fostering learners' ability to select, compare, critically evaluate and purposefully use different digital lexicographic information systems, machine translation systems and AI-supported language tools (cf. Klosa-Kückelhaus & Nied Curcio, 2024). Despite these developments, research into dictionary use shows that users of digital lexicographic resources are not necessarily successful users (cf. Nied Curcio, 2014, 2015a, pp. 449–450): They often consult a few websites, they hardly know online dictionaries of high quality, they usually use the first dictionary/website they find via a search engine, they cannot really orient themselves in the dictionary article, they overlook important information, etc. Immediate, fast and free access to these tools is considered very important for users.

It is well known that with the language policy of the Council of Europe and the publication of the Common European Framework of Reference for Languages (CEFR 2001) and its Companion Volume (2020), with its emphasis on action and competence orientation, multilingualism and lifelong learning, dictionaries have regained importance in the field of GFL. With regard to GL1, current German language curricula are shaped by a competence-oriented approach to teaching and learning (cf. Ossner, 2019; Düsing, 2023; e. g. LehrplanPLUS in Bavaria). Within this framework, dictionaries are not regarded as ends in themselves but as tools that support learners in accomplishing authentic language-related tasks and solving linguistic problems independently. Their use is linked to the development of subject-specific competences, such as language awareness, reading and writing competence, and reflective language use. Moreover, digital lexicographic resources contribute to broader educational objectives by developing digital literacy, critical information skills, and the capacity for self-directed and lifelong learning. Across both contexts, the appropriate use of dictionaries and lexicographical resources is regarded as an integral part of language learning strategies and an important component of methodological and media competence development (cf. Nied Curcio, 2015b, pp. 303–305).

When learners are asked about their needs in relation to the use of digital lexicographic resources, they express a desire to learn more about the lexicographic tools available. They want to know which dictionaries are available for (foreign) language learning, how they are organised and structured, how to recognise reliable information and which ones are most suitable for a specific goal. They show great interest in improving their skills in using lexicographic tools in order to make fewer mistakes and become experienced and competent users (cf. Nied Curcio, 2015a, p. 462). In the classroom, however, teachers seem to be of little help. Many of them are not successful users of online lexicographic resources themselves, feel insecure with their skills in using online dictionaries and are at a loss as to how to integrate modern lexicographic resources into the classroom. They themselves express a desire to receive further training in this area (cf. Nied Curcio, 2024).

Recent studies have increasingly examined the relationship between curricular requirements, classroom practice and dictionary use. Research has documented discrepancies between educational guidelines and actual teaching practices (cf. Abel, 2024) as well as a declining pedagogical emphasis on dictionary skills despite the growing importance of digital language resources (cf. Dringó-Horváth & Márkus, 2024). However, considerably less attention has been paid to the extent to which these developments are reflected in language textbooks, although textbooks remain one of the central mediating instruments between curricula and classroom practice. To adress this gap, the present study addresses the following overarching research question:

To what extent is the digital transformation of lexicography, including the emergence of AI-supported language tools, reflected in contemporary German language textbooks for GFL and GL1?

To answer this question, the analysis focuses on three interrelated aspects:

- Which types of lexicographic resources (print, digital, hybrid and AI-supported) are represented in contemporary German language textbooks?
- Which pedagogical and lexicographic functions are assigned to these resources, and how are they integrated into learning tasks and competence development?

– To what extent do textbooks reflect current developments in lexicography, particularly regarding digital lexicographic information systems, lexicographic competence and AI-supported language consultation?

At present, these questions can only be answered tentatively, as there is still insufficient knowledge about the actual use of lexicographic resources in educational settings. Reliable empirical data remain scarce. Furthermore, investigating the use of digital lexicographic resources is particularly challenging because these resources are continuously updated, which may limit the long-term validity of research findings (Nied Curcio, 2022, p. 74). Consequently, current insights into classroom practice are largely based on isolated observations and informed assumptions. It appears that the extent to which lexicographic online resources are integrated into teaching often depends on individual teachers, provided that the use of tablets, computers, or mobile devices is permitted in the classroom. Moreover, teachers frequently assume that younger learners are capable of using these tools effectively and independently; however, research suggests that this assumption is not always justified.

Against this background, textbook analysis offers an appropriate point of departure for investigating how contemporary lexicographic developments are represented in language education. Although textbooks cannot fully capture actual classroom practice, they provide valuable evidence of the lexicographic concepts, consultation practices and competence models that are conveyed to learners. To address the research questions outlined above, this paper presents findings from two exemplary textbook analyses: (1) German as a Foreign Language (GFL) textbooks and (2) German as a First Language (GL1) textbooks, with particular attention to tasks involving dictionary use and the consultation of lexicographic resources. The comparison makes it possible to examine how contemporary lexicographic developments – including digital lexicographic information systems and AI-supported language tools – are represented across two educational contexts. While such analyses cannot capture actual classroom practices, they may nevertheless provide initial insights into the role assigned to lexicographic resources in teaching materials. The analysis draws on two corpora: GFL textbooks published between 2011 and 2024 covering CEFR levels A1–C1, and GL1 textbooks published between 2003 and 2021 and used across different educational levels in German schools. All textbooks were examined manually for dictionary-specific terminology and references to dictionaries or other lexicographic resources. Relevant terms, passages, and tasks were systematically identified, excerpted and coded for analysis. The study combines quantitative frequency analysis with qualitative content analysis, focusing on the frequency of references to lexicographic resources, the terminology used to refer to them, the types of tasks and instructions provided, the degree of digital integration, the representation of AI-supported language tools, and the didactic functions attributed to these resources.

2 Lexicographic (online) resources in textbooks for GFL

2.1 GFL textbook sample

The first part of the comparative textbook analysis focuses on GFL textbooks published between 2011 and 2024, with a particular focus on exercises involving dictionary use. It examines exercises involving dictionary use in order to investigate how contemporary lexicographic resources are represented in GFL teaching materials and whether recent developments in digital lexicography and AI-supported language tools are reflected in textbook design. Textbooks from various publishers were selected to represent different proficiency levels of the CEFR. The aim is to determine at which proficiency levels digital lexicographic resources

are used, whether they are predominantly employed in beginner-level instruction, and whether their use may decline at higher proficiency levels. In addition, it was important not only to select current textbooks published after 2020 but also to review older textbooks, particularly to examine whether they met the requirements of the CEFR from the outset and whether the situation has changed over time. The analysed GFL textbooks are:

- *DaF kompakt*. A1-A2-B1. 2011
- *Studio. Die Mittelstufe*. B2. 2010/2011
- *Studio. Die Mittelstufe*. C1. 2015
- *Daf kompakt neu*. A1-A2-B1. 2016
- *Studio express*. A1-A2-B1. 2018
- *Kompass DaF*. B2. 2020
- *Motive*. A1-A2-B1. 2023
- *Kurs DaF*. A1-A2. 2024.

As mentioned above, the main focus of the analysis was to identify the tasks involved in working with the dictionary. To this end, all instructions – both in the coursebook and in the workbook – were included in the corpus if they explicitly mentioned the word *Wörterbuch* ('dictionary') or related terms such as *Print-Wörterbuch* ('print dictionary'), *Online-Wörterbuch* ('online dictionary'), *einsprachiges Wörterbuch* ('monolingual dictionary'), *zweisprachiges Wörterbuch* ('bilingual dictionary'), as well as *Suchmaschine* ('search engine'), *automatischer Übersetzer* ('automatic translator') or *Übersetzungsprogramm* ('translation program') or tasks and exercises in which dictionary excerpts or excerpts from dictionary entries are included.

2.2 Findings of the GFL textbook analysis

The analysis reveals an extremely low presence of lexicographic resources. Out of approximately 10,000–12,000 task instructions, only 65 explicitly referred to dictionaries or other lexicographic resources. References to digital formats are marginal. These findings suggest that lexicographic resources have played only a minor role in GFL textbooks in recent years and continue to occupy a rather peripheral position in contemporary teaching materials.

In the task instructions, *Wörterbuch* appears by far the most frequently, namely 64 times. Only five instances of a more specific designation of the dictionary were identified: *einsprachiges Wörterbuch* ('monolingual dictionary'), (*digitales*) *Wörterbuch* ('digital dictionary') and *Online-Wörterbuch* ('online dictionary'). Not a single reference is made to a bilingual dictionary, even though its use is implicitly suggested in certain task formulations, such as *Arbeiten Sie mit dem Wörterbuch. Wie heißt das auf Deutsch?* ('Work with the dictionary. How do you say this in German?').

Overall, the terminology used in the textbooks reflects a predominantly traditional understanding of dictionaries and their use. Dictionaries are generally treated as generic reference works, without distinguishing between different types of lexicographic resources or consultation environments. Neither the digital transformation of lexicography nor empirical findings from dictionary-use research – showing that foreign language learners increasingly rely on search engines, online dictionaries, and machine translation tools – are reflected in the analysed textbooks.

TABLE 1 Instructions for working with dictionaries in GFL textbooks

instructions for working with dictionaries	frequency
Arbeit mit dem Wörterbuch. / Arbeiten Sie mit dem Wörterbuch. ('work with a dictionary')	21
Benutzen/ Nutzen/ Verwenden Sie ein/ Ihr Wörterbuch. ('use a dictionary')	8
Schlagen Sie (. . .) im Wörterbuch nach. ('look up . . . in a dictionary')	5
Lesen Sie die Auszüge aus dem Wörterbuch und (. . .) / Lesen Sie den Wörterbuchauszug/ -aufriss ('read dictionary entries/ extracts')	5
Kontrollieren Sie mit dem Wörterbuch. ('check or verify information using a dictionary')	3
Schauen Sie (. . .) im Wörterbuch nach. ('look up . . . in a dictionary')	2
Schauen Sie ins Wörterbuch. ('consult a dictionary')	2
Suchen Sie im Wörterbuch. / Suchen Sie (. . .) im Wörterbuch. ('Search for information in a dictionary')	2
Finden Sie im Wörterbuch (. . .) / (. . .) im Wörterbuch finden. ('find . . . in a dictionary')	2
Überprüfen Sie mit weiteren Wörtern aus dem Wörterbuch. ('verify your findings using additional words from the dictionary')	1
Übersetzen Sie (. . .) mit einem Wörterbuch (. . .) ('translate using a dictionary')	1
Schreiben Sie eine Definition, wie Sie im Wörterbuch stehen könnte. ('write dictionary-style definitions')	1
Arbeiten Sie ohne (digitales) Wörterbuch. ('work without a (digital) dictionary')	3

If one examines the sentences in the tasks in which the word *Wörterbuch* ('dictionary') appears, the following picture emerges (see table 1).

Table 2 presents the distribution of dictionary-related tasks across different linguistic domains in the analysed GFL textbooks.

The analysis of GFL textbooks shows that dictionary use is conceptualised in a rather narrow and traditional way. Only a limited number of tasks explicitly address the use of dictionaries or other lexicographic resources, and these are generally formulated in broad and unspecific terms. In most cases, neither the type of dictionary nor the intended consultation strategy is specified. Consequently, learners receive little guidance in selecting appropriate lexicographic resources or in developing effective consultation skills.

Additionally, the textbooks do not reflect the ongoing digital transformation of lexicography, as vocabulary work is still largely based on traditional print dictionaries while digital lexicographic resources are only marginally represented. Most activities focus on vocabulary acquisition rather than integrating dictionary use into specific language skills such as reading or writing. Furthermore, work with lexicographical resources is rarely embedded in meaningful, linguistically relevant usage scenarios or communicative contexts for learners. Most tasks related to dictionary use in textbooks are situated within the field of indirect dictionary didactics, which means that the dictionary is intended as a tool for completing specific tasks, rather than focusing on learning how to use the dictionary itself (cf. Zöfgen, 2019). There is also a lack of task formats that help students develop strategies for effectively using these resources or that promote (digital) media competence. Finally, the textbooks do not include activities that

TABLE 2 Distribution of dictionary-related tasks in GFL textbooks

linguistic domain	number of tasks	dictionary-related activity	frequency
vocabulary acquisition/ expansion	36	look up unknown words; infer word meaning	20
		understand entries or definitions in dictionary articles/ extracts	10
		find additional/ new words (vocabulary expansion)	4
		search for antonyms	1
		transfer linguistic expressions from a variety into standard language	1
Grammar	12	look up gender, singular and plural forms	8
		look up genitive endings in dictionary entries	1
		look up auxiliary verbs	1
		look up verb conjugation	1
		look up past participle forms	1
orthography	3	correct spelling	3
pronunciation	1	look up stressed vowel	1

encourage a critical examination of lexicographical tools, including online dictionaries, search engines and AI-supported language tools, limiting learners' ability to use them in a reflective and informed way and to develop lexicographic competence.

3 Lexicographic (online) resources in textbooks for GL1

3.1 GL1 textbook sample

The GL1 corpus consists of ten German language textbooks for the academic secondary school (Gymnasium) published between 2003 and 2021, comprising a total of 3,820 pages. The selection includes textbooks from different publishers for the same grade level (Grade 5), textbooks for different grade levels, and different editions of the same textbook series (Cornelsen):

- *Deutsch kompetent* 5 (2017, Klett)
- *Deutschbuch. Sprach- und Lesebuch* 5-9, revised edition (2017–2021, Cornelsen)
- *Deutschbuch. Sprach- und Lesebuch* 5-6, previous edition (2003, 2004, Cornelsen)
- *D.U.* (DeutschUnterricht) 5 (2017, C.C. Buchner)
- *P.A.U.L.* D.5 (2017, Schöningh Westermann)

All textbooks were approved for use in Bavarian schools and therefore comply with the legal and curricular requirements governing school textbooks in Bavaria. In Germany, textbooks undergo an official approval procedure before they may be used in schools. Consequently, they occupy a central position in classroom practice and are closely aligned with the curriculum. The textbooks analysed are largely based on the Bavarian curriculum (LehrplanPLUS), introduced

successively in secondary schools since 2017/18. The curriculum for German at Gymnasium level is competence-oriented and organised into four learning domains: (1) "Sprechen und Zuhören" ('Speaking and Listening'), (2) "Lesen – mit Texten und anderen Medien umgehen" ('reading and dealing with texts and other media'), (3) "Schreiben" ('writing'), (4) "Sprachgebrauch und Sprache untersuchen und reflektieren" ('investigating and reflecting on language use'). Methodological skills and working techniques are developed across all domains. The central aim of LehrplanPLUS is the development of sustainable competencies that enable students to apply knowledge, skills, and strategies independently in authentic situations both within and beyond the classroom.

It therefore seems appropriate to organise the analysis according to these learning domains. All textbooks were examined manually. Dictionary-related tasks and references to lexicographic resources were systematically identified, excerpted and analysed. Tasks involving dictionaries and other lexicographic resources were classified according to whether they primarily support language production (writing, speaking, and listening), language reception (reading and comprehension), language reflection (including vocabulary, word meaning, word fields, word formation, grammar, syntax, spelling, and style), or methodological competence. This categorisation allows examination not only of the frequency of dictionary-related activities but also of the pedagogical functions within the curriculum competence areas.

3.2 Findings of the GL1 textbook analysis

In contrast to the GFL textbooks, lexicographic resources are much more visible in the GL1 corpus. Across the ten textbooks analysed, 225 references to dictionaries and related reference works were identified. In line with the competence-oriented approach of the curriculum, references to dictionaries and other reference works are found in all textbooks. The number of references per textbook ranges from 11 to 40, with most containing around 20 references. Of these, 155 occurrences appear in task instructions. However, about one third occur in informational contexts rather than tasks, including mentions in competence boxes, basic knowledge sections, learning units, and appendices. These references primarily provide information about dictionaries and other reference works rather than offering explicit guidance on how to use them effectively. In most cases, the focus is on informing students about dictionaries and reference works rather than providing explicit guidance on how to use them effectively.

With regard to terminology, the textbooks rely primarily on the terms *Wörterbuch* ('dictionary') and *Lexikon* ('encyclopaedia'). Less frequently, the term *Nachschlagewerk* ('reference work') is used. A systematic distinction between different types of reference works is only occasionally made. Although specialised dictionaries are occasionally mentioned (*Rechtschreibwörterbuch* 'spelling dictionary', *Synonymwörterbuch* 'dictionary of synonyms', *etymologisches Wörterbuch* 'etymological dictionary', *Herkunftswörterbuch* 'dictionary of word origins', *Fremdwörterbuch* 'dictionary of foreign words'), references to specific lexicographic resources are rare. Only two explicit references to named works were identified (*Duden Rechtschreibwörterbuch, Fidibus)*. Within encyclopaedic reference works, distinctions are made between *Schülerlexikon* ('school encyclopaedia'), *Jugendlexikon* ('youth encyclopaedia'), and *Länderlexikon* ('country encyclopaedia'). Digital resources are mentioned only rarely and are usually associated with encyclopaedic reference works rather than with contemporary digital lexicographic tools. References include *Wikipedia, Online-Lexikon* ('online encyclopaedia'), *(Online-)Wörterbuch* ('(online) dictionary'), *Wörterbuch-App* ('dictionary app'), *elektronisches Lexikon* ('electronic encyclopaedia'), and *multimediale Enzyklopädie (CD-ROM)* ('multimedia encyclopaedia on CD-ROM'). In addition, the term *Thesaurus* appears in reference to the synonym function of a word-processing

program. A comparison of older and revised editions of the Cornelsen series suggests a certain updating of terminology, with references to electronic encyclopaedias and CD-ROM-based resources partly replaced by online dictionaries and dictionary apps. Nevertheless, the textbooks reflect only limited aspects of the ongoing digital transformation of lexicography. While contemporary forms of online lexicographic consultation remain largely absent, AI-supported language tools are not mentioned at all. Given that the analysed textbooks were published between 2003 and 2021, this finding is not unexpected. A closer examination of the 155 references occurring in task instructions (see table 3) reveals considerable variety in dictionary-related activities. While traditional dictionary consultation and verification tasks clearly dominate, the corpus also contains more creative and cognitively demanding activities, such as comparing dictionary entries, conducting research, participating in dictionary quizzes, and reflecting on the functions and structure of dictionary articles. At the same time, author- and publisher-specific differences become apparent. Certain instruction types are characteristic of particular textbook series; for example, tasks using the instruction verb *prüfen* ('check' or 'verify') occur exclusively in the revised *Deutschbuch* series published by Cornelsen. Furthermore, some textbook series concentrate dictionary-related activities in specific chapters or methodological sections, whereas others – particularly the Cornelsen textbooks – integrate them more evenly throughout the volume.

Table 4 presents dictionary-related tasks across different linguistic domains in the GL1 textbooks.

The distribution of tasks shows a clear emphasis on vocabulary-related activities, accounting for 49.7% of all dictionary-related tasks. Metalexicographic knowledge is the second-largest category (29.7%), followed by orthography (16.1%) and grammar (4.5%). As in the GFL corpus, vocabulary work clearly dominates, but – as expected – with a different orientation. Two main functions can be identified: supporting text comprehension and providing access to factual information, often via encyclopaedic resources. The dominant task type is the lookup of unfamiliar words. This focus is reinforced by informational textbook sections, where dictionaries are frequently associated with meaning retrieval. However, these discussions are often detached from authentic problem-solving situations and provide only limited guidance on transferring dictionary skills to broader learning contexts. Orthography-related tasks are less frequent than expected. It is possible, however, that actual classroom practice relies on spelling dictionaries more extensively than the textbooks suggest. Grammar-related activities are rare, and dictionaries are only marginally used for word formation, lexical relations, and semantic fields.

In contrast to the GFL corpus, the prominence of metalexicographic knowledge is striking. Tasks address dictionary structure, entry components, alphabetical ordering, and the evaluation of entries. This can be interpreted as evidence of elements of direct dictionary didactics in GL1 textbooks, providing learners with operational knowledge and promoting dictionaries as tools for independent learning. Regarding the curriculum competence areas, dictionary-related activities are mainly situated within the domain of reflecting on language use, often in connection with spelling, but are also linked to vocabulary development, loanword etymology, and grammar. Dictionary use is considerably more frequent in text reception and text revision than in text production.

Overall, GL1 textbooks assign dictionaries a more prominent role than GFL textbooks and place stronger emphasis on dictionary-related knowledge. Nevertheless, the use of lexicographic resources remains limited in scope. Dictionary consultation is mainly framed as a tool for meaning retrieval and spelling control, while other functions receive less attention. Their potential for language production, reflection, vocabulary development, and strategic learning

TABLE 3 Instruction verbs used in dictionary related tasks in GL1 textbooks

instruction verbs	task example	fre-quency
nachschlagen ('look up')	Look up unfamiliar loanwords in a dictionary.	20
w-Frage beantworten (answer wh-questions)	What function do the words in the upper left corner of the page have in the dictionary?	17
klären ('clarify')	Clarify unknown words from the context using a dictionary or online research.	12
prüfen ('check' / 'verify')	Check the spelling in the dictionary.	12
erklären ('explain')	Using the following entries from a dictionary of foreign words, explain what kinds of information such a dictionary provides.	11
suchen ('search')	Search a dictionary for words marked as obsolete.	8
recherchieren, Recherche durchführen ('conduct research')	Research the meanings of the terms you have noted as unfamiliar in an encyclopaedia or dictionary.	6
beschreiben, kontrollieren ('describe, check')	Check the spelling of your words using a dictionary.	10 (2x 5)
lesen, notieren, überprüfen ('read, take notes, verify')	Read the entry from an online encyclopaedia.	12 (3x 4)
begründen, formulieren, sich informieren, nutzen, verfassen, vergleichen ('justify, formulate, inform oneself, use, compose, compare')	Compare the entries in a spelling dictionary with those in a dictionary of foreign words.	18 (6x 3)
berichten, besprechen, bilden, ergänzen, erstellen, finden, zu Hilfe nehmen, zuordnen ('report, discuss, form, complete, create, find, consult, assign')	Find the correct article for the following words.	16 (8x 2)
ansehen, austauschen, auswählen, bearbeiten, benutzen, bestimmen, sich ein Bild machen, entscheiden, erläutern, feststellen, herausfinden, herausschreiben, lösen, nachprüfen, nachschauen, nennen, reagieren, sammeln, schreiben, Schreibkonferenz durchführen, überarbeiten, veranstalten, verwenden, wiedergeben ('consult, exchange ideas, select, edit, use, determine, form an opinion, decide, explain, identify, find out, extract, solve, verify, look up, respond, collect, write, conduct a writing conference, revise, organise, reproduce')	Organise a "dictionary quiz": work in pairs or groups of three and solve the following tasks using a dictionary.	25 (25x 1)

TABLE 4 Distribution of dictionary-related tasks in GL1 textbooks

linguistic domain	number of tasks	dictionary-related activity	frequency
Vocabulary	77	look up unknown words and infer meaning (esp. loanwords)	35
		understand encyclopaedia and dictionary entries and extract specific information	17
		identify and learn about the origins of loanwords	7
		look up sensitive words and reflect on language use	4
		find synonyms and compile semantic fields	3
		research word histories and trace semantic change	3
		look up and explain technical terms	3
		search for and explain obsolete words and dialect words	2
		look up and explain idiomatic expressions	1
		identify the base word of a derived word	1
		identify related word formations	1
meta-lexicographic knowledge	46	learn about dictionary structures (microstructure)	12
		compare and evaluate dictionary or encyclopaedia entries	12
		practise alphabetical ordering	7
		reflect on dictionary use	5
		formulate information as a dictionary or encyclopaedia entry	5
		learn about dictionary structures (macrostructure)	4
		use dictionary guide texts and user information	1
Orthography	25	check the spelling of (loan)words	10
		practise problematic areas of spelling (e.g. sound-letter correspondences, silent *h*, *s*-spelling, capitalisation)	6
		learn about spelling variants (especially loanwords)	4
		look up spelling rules	3
		practise hyphenation and word division	2
Grammar	7	look up gender, singular and plural forms	5
		identify infinitive forms	2

is only partly exploited, particularly in writing, where links to lexical variation, stylistic revision, semantic field development, and idiomatic or regional expression are rare. Dictionaries also play only a minor role in grammatical reflection. Moreover, many tasks could benefit from dictionary consultation but do not explicitly encourage it; instead, textbooks often provide internal lexical support rather than promoting external resources. References to contemporary digital lexicographic resources are absent. It also remains unclear to what extent textbooks reflect actual classroom practice, as dictionary use ultimately depends on individual teachers.

4 Summary, conclusions and future directions

4.1 Summary and conclusions

The analysis of school textbooks and learning materials shows that lexicographic resources continue to play only a marginal role in language education in the fields of GFL and GL1. Although all textbooks include tasks related to dictionary use, these remain limited in number and vary depending on publisher and competence area. Some differences also appear to reflect author- and publisher-specific editorial concepts, indicating that the treatment of lexicographic resources is shaped not only by curricular requirements but also by the didactic orientation of individual textbook series. Moreover, the quality of such tasks is often rather general and underdeveloped. The focus is mainly on vocabulary acquisition (GFL) or meaning clarification (GL1), while more advanced language activities such as writing, text production, or language reflection are rarely addressed. Extended lexicographic information (e.g. grammar, etymology, or usage) is also seldom systematically integrated. In addition, the digital transformation has not yet been sufficiently reflected in textbook design. Digital dictionaries and online lexicographic resources are only marginally represented, whereas AI-supported language tools are not mentioned at all. Furthermore, task formats that promote strategic, reflective, or critical use of lexicographic resources are largely missing in GFL textbooks and occur only sporadically in GL1 textbooks, often depending on the textbook series or publisher. As a result, important aspects of media and information literacy, as well as the critical handling of linguistic tools, are insufficiently developed and are not sustainably embedded. Overall, the potential of lexicographic work for authentic and action-oriented language learning processes remains largely untapped.

The findings of the present study underline the need to strengthen the role of lexicographic resources in (foreign) language education and to adapt dictionary didactics to the ongoing digital transformation of lexicography. In response to these challenges, the thematic issue on *Digitale Wörterbücher* ('Digital Dictionaries') (Ganslmayer & Nied, 2026b) presents current digital lexicographic resources, tools, and projects, demonstrates how they can be used in German language education, and highlights their didactic potential. It aims to strengthen awareness of their relevance for language education, show their contribution to media and information literacy as well as critical language awareness, and promote the meaningful integration of digital lexicographic resources into classroom practice. In addition, it provides guidance for teachers by presenting current digital lexicographic resources and illustrating their functional use in concrete learning contexts, thereby contributing to the re-establishment of dictionaries and related lexicographic information systems as an integral component of language teaching.

In this context, AI-based applications, particularly generative AI chatbots, can further extend the scope of lexicographic support in language learning (cf. Ganslmayer, 2026). They provide a useful complement to traditional dictionaries and other digital lexicographic resources by offering contextualised explanations, immediate feedback, and adaptive support. While dictionaries ensure reliable and standardised information on vocabulary, grammar, and usage, AI tools may additionally assist learners with meaning-making in context, language production, individualised learning processes, and interactive, dialogic exchange. Through adaptive prompting, AI outputs can be personalised and adjusted to learners' proficiency levels or age. However, as AI-generated content may be inaccurate, critical evaluation remains essential. Used in combination with scientifically developed lexicographic resources, AI-based applications can support linguistic accuracy, research skills, learner autonomy, and digital literacy.

Preliminary classroom observations suggest that the strengths of AI-based tools vary depending on the task. Generative AI chatbots appear less reliable when dealing with idiomatic expressions and fixed phrases, where meanings cannot easily be inferred from the individual

components. This applies particularly to the distinction between literal and figurative meanings and to the context-appropriate use of phraseological vocabulary. At the same time, they can provide particularly useful support for identifying and explaining collocations, in some cases even outperforming conventional machine-translation tools. Such differences offer valuable opportunities for exploratory classroom activities in which learners compare dictionaries, machine-translation systems, and AI chatbots in authentic language tasks. By reflecting on the strengths and limitations of each resource, learners can develop a more critical and informed approach to language-related digital tools while strengthening both lexicographic and digital literacy.

4.2 Future directions and desiderata

For a sustainable integration of dictionary didactics into language education, several further measures are required. These include systematic teacher training and the integration of relevant content into teacher education programs. In addition, appropriate teaching materials should be developed and made accessible via digital platforms and Open Educational Resources (OER).

The findings of the present study also point to the need for a reconceptualisation of dictionary didactics. Rather than focusing exclusively on dictionaries as individual reference works, it should address the informed and critical use of the full spectrum of contemporary lexicographic information systems, including online dictionary portals, corpus-based resources, integrated lexical tools, machine translation systems and AI-supported language applications. Accordingly, the central educational objective is no longer dictionary consultation alone but enabling learners to select, use, compare and critically evaluate lexical information across different digital environments.

A stronger learner-centered approach is also necessary to promote active and reflective dictionary use. Assessment formats should be adapted accordingly to reflect dictionary skills. Furthermore, curricular embedding in language education programs and continued research on the development and evaluation of lexicographic didactics are essential. Against this background, for the practical implementation of dictionary work in the classroom, two complementary approaches should be taken into account more systematically: direct and indirect dictionary didactics (cf. Zöfgen, 2019), building on successful existing task designs. Direct dictionary didactics focuses on the explicit development of dictionary skills. Learners systematically acquire knowledge about the structure, design, and use of lexicographic resources. Typical task formats include understanding dictionary entries, identifying specific information (e.g. meaning, register, collocations), comparing different types of dictionaries (print and digital), and developing strategies for efficient and critical use. Indirect dictionary didactics, by contrast, integrates dictionaries as tools within communicative and task-based language activities. The goal is functional and autonomous dictionary use in authentic language situations. This includes, for example, inferring word meaning from context and verifying it with a dictionary, using dictionaries in writing processes for greater precision and variation, error correction, text revision, and project-based lexical research.

These considerations also have important implications for the development of future textbooks. Textbook authors should design concrete, meaningful, engaging and motivating tasks with clearly defined pedagogical objectives. At the same time, dictionary-related activities should be aligned with learners' actual consultation needs and reflect authentic usage scenarios in which lexicographic resources support the completion of communicative and language-related tasks. Moreover, many of the activities already included in contemporary textbooks could readily be extended to incorporate dictionary work, both directly and indirectly, thereby

strengthening learners' lexicographic competence without requiring entirely new task formats. Future research should therefore investigate more systematically the extent to which the representation of lexicographic resources in language textbooks is shaped by curricular requirements, authorial decisions, or publisher-specific editorial policies. Such research would help to distinguish between broader trends in language education and differences resulting from individual textbook concepts, thereby providing a more differentiated basis for the further development of textbook design and dictionary didactics. In addition, closer cooperation with publishers and educational technology providers is needed, as well as the establishment of networks and communities of practice to share experiences and best practices. Any future integration of AI-supported language tools into teaching materials will need to be aligned with the respective ministerial guidelines governing the use of AI in schools. Overall, these measures contribute to securing dictionary work as a permanent component of language education and to fully exploiting its potential to promote language learning, informed and critical dictionary use, media literacy, and critical language awareness.

References

Abel, A. (2024). The use of (online) dictionaries at the interface of curricular requirements and practice. In A. Klosa-Kückelhaus & M. Nied Curcio (Eds.), *Dictionary Use and Dictionary Teaching. New Challenges in a Multilingual, Digital and Global World* (pp. 171–191). De Gruyter. https://doi.org/10.1515/9783111373294-202

Bayerisches Staatsministerium für Unterricht und Kultus (o. J.). *LehrplanPLUS*. https://www.lehrplanplus.bayern.de

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge University Press. https://www.coe.int/en/web/common-european-framework-reference-languages

Council of Europe (2020). *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion volume with new descriptors*. Council of Europe Publishing.

Dringó-Horváth, I. & Márkus, K. P. (2024). Using dictionaries in teaching (and learning) English as a Foreign Language – the beginning of a longitudinal research project. In A. Klosa-Kückelhaus & M. Nied Curcio (Eds.), *Dictionary Use and Dictionary Teaching. New Challenges in a Multilingual, Digital and Global World* (pp. 9–28). De Gruyter. https://doi.org/10.1515/9783111373294-202

Düsing, E. (2023). Bildungsstandards und Kompetenzen für das Fach Deutsch. In C. Goer & K. Köller (Eds.), *Fachdidaktik Deutsch. Grundzüge der Sprach- und Literaturdidaktik* (pp. 51–61). Fink (UTB).

Ganslmayer, C. (2026). Von traditioneller Recherche zu KI-unterstützter Wortschatzarbeit. Chancen und Herausforderungen für den Deutschunterricht. In *Der Deutschunterricht*, 78(1), 76–87.

Ganslmayer, C. & Nied Curcio, M. (2026a). Wörterbücher im digitalen Wandel. Kompetenzen, Möglichkeiten und didaktische Implikationen. In *Der Deutschunterricht*, 78(1), 2–6.

Ganslmayer, C. & Nied Curcio, M. (Eds.) (2026b). *Digitale Wörterbücher*. Themenheft *Der Deutschunterricht*, 78(1). Friedrich.

Haß, U. (2019). Digitale Nachschlagewerke: Wozu sind welche Wörterbücher und Lexika beim Schreiben und Lesen in den Sekundarstufen I und II nützlich? In K. Kuhs & S. Merten (Eds.), *Arbeiten am Wortschatz. Schreiben* (pp. 253–268). Wissenschaftlicher Verlag Trier.

Klosa-Kückelhaus, A. & Nied Curcio, M. (2024). Introduction. In A. Klosa-Kückelhaus & M. Nied Curcio (Eds.), *Dictionary Use and Dictionary Teaching. New Challenges in a Multilingual, Digital and Global World* (pp. 1–7). De Gruyter. https://doi.org/10.1515/9783111373294-202

Kühn, P. (1989). Schulwörterbücher sind lexikographische Langweil-Bestseller oder warum eine Wörterbuchkultur in Deutschland nicht in Gang kommt. In H. E. Wiegand (Ed.), *Wörterbücher in der Diskussion. Vorträge zum Heidelberger Lexikographischen Kolloquium* (pp. 111–132). Niemeyer.

Kühn, P. (1994). *Mein Schulwörterbuch. Kleine Didaktik und Methodik der Wörterbucharbeit*. Dümmler.

Kühn, P. (1999a). Positionen und Perspektiven der Wörterbuchdidaktik und Wörterbucharbeit im Deutschen. *Lexicographica*, 14, 1–13.

Kühn, P. (1999b). *Wörterbucharbeit in der Schule. Anregungen für den Umgang mit dem Wörterbuch im Bereich der Primar- und Sekundarstufe I*. Trier Universitätsdruck.

Kühn, P. (2011). Die Wörterbuch-Werkstatt. Wörterbucharbeit zur Wortschatzvertiefung und -erweiterung. *Praxis Deutsch*, 165, 21–24.

Merten, S. (2022a). Arbeit mit Wörterbüchern. In I. Pohl & W. Ulrich (Eds.), *Wortschatzarbeit* (pp. 380–394). Schneider.

Merten, S. (2022b). Sichtung von Wörterbüchern für die Schule. In I. Pohl & W. Ulrich (Eds.), *Wortschatzarbeit* (pp. 457–473). Schneider.

Müller-Spitzer, C., Domínguez Vazquez, M. J., Silva Dias, I. M., Nied Curcio, M. & Wolfer, S. (2018). Correct Hypotheses and Careful Reading Are Essential: Results of an Observational Study on Learners. Using Online Language Resources. *Lexikos*, 28, 287–315. https://doi.org/10.5788/28-1-1466

Nied Curcio, M. (2014). Die Benutzung von Smartphones im Fremdsprachenerwerb und -unterricht. In A. Abel, C. Vettori & N. Ralli (Eds.), *The User in Focus. Proceedings of the XVI EURALEX International Congress* (pp. 263–280). Institute for Specialised Communication and Multilingualism, Bolzano/Bozen.

Nied Curcio, M. (2015a). Wörterbuchbenutzung und Wortschatzerwerb. Werden im Zeitalter des Smartphones überhaupt noch Vokabeln gelernt? *Info DaF* 42(5), 445–468.

Nied Curcio, M. (2015b). Spielen Wörterbücher bei der Sprachmittlung noch eine Rolle? In M. Nied Curcio, P. Katelhön & I. Bašić (Eds.), *Sprachmittlung – Mediation – Mediazione linguistica. Ein deutsch-italienischer Dialog* (pp. 291–317). Frank & Timme.

Nied Curcio, M. (2022). Dictionaries, foreign language learners and teachers. New challenges in the digital era. Die Benutzung von Smartphones im Fremdsprachenerwerb und -unterricht. In A. Klosa-Kückelhaus, S. Engelberg, C. Möhrs & P. Storjohann (Eds.), *Dictionaries and Society. Proceedings of the XX EURALEX International Congress* (pp. 71–84). Mannheim, IDS-Verlag.

Nied Curcio, M. (2024). The use of lexicographic online resources by foreign language teachers and the effectiveness of teacher training. In A. Klosa-Kückelhaus & M. Nied Curcio (Eds.), *Dictionary Use and Dictionary Teaching. New Challenges in a Multilingual, Digital and Global World* (pp. 193–219). De Gruyter. https://doi.org/10.1515/9783111373294-202

Ossner, J. (2019). Kompetenz. In B. Rothstein & C. Müller-Brauers (Eds.), *Kernbegriffe der Sprachdidaktik Deutsch. Ein Handbuch* (pp. 195–198). Schneider.

Wiegand, H. E., Beißwenger, M., Gouws, R. H., Kammerer, M., Mann, M., Storrer, A. & Wolski, W. (Eds.) (2010). *Wörterbuch zur Lexikographie und Wörterbuchforschung. Dictionary of Lexicography and Dictionary Research*. Vol. 1 (A–C). De Gruyter.

Zöfgen, E. (2019). Wörterbuchdidaktik. In W. Hallet & F. G. Königs (Eds.), *Handbuch Fremdsprachendidaktik* (pp. 107–110). Klett.

Contact information

Christine Ganslmayer
Friedrich-Alexander-Universität Erlangen-Nürnberg
christine.ganslmayer@fau.de

Martina Nied Curcio
Università degli Studi Roma Tre
martina.nied@uniroma3.it

Zoe Gavriilidou & Evanthia Konstantinidou

PEDAGOGICAL LEXICOGRAPHY IN PRACTICE

Raising Dictionary Use Strategies Through Card-Based Instruction

Abstract This paper presents the pedagogical rationale, theoretical foundations, and instructional design of the book *100 Practice Cards for Dictionary Use and Vocabulary Development* (Gavriilidou & Konstantinidou, 2025), a classroom-oriented educational resource developed to promote strategic dictionary use among upper-primary and lower-secondary learners. Grounded in contemporary pedagogical lexicography, strategy-based instruction, and vocabulary acquisition research, the material aims to transform dictionary consultation from a mechanical reference activity into a meaningful learning process that supports vocabulary development, metacognitive awareness, learner autonomy, and language proficiency. The study highlights the broader educational value of the material, especially its contribution to inclusive education, strategic dictionary use, soft skill cultivation and autonomous language learning.

Keywords Pedagogical lexicography; dictionary use strategies; vocabulary development; learner autonomy; differentiated instruction; dictionary culture

1 Introduction

Strategic dictionary use is increasingly conceptualized as a strategy-based learning activity involving cognitive, metacognitive, and decision-making processes that contribute to language learning and competence development (Gavriilidou, 2013). By requiring learners to analyse lexical problems, evaluate alternatives, justify lexical choices, and monitor their learning, it also fosters the soft skills emphasized in competence-based education.

Despite their importance as language-learning tools, dictionaries are often used in limited and ineffective ways, with learners relying mainly on meaning retrieval instead of exploiting their potential for vocabulary development, reading comprehension, writing, and learner autonomy (Gavriilidou, 2014; Lew, 2013). Explicit instruction in dictionary use strategies enhances strategic dictionary use while fostering dictionary culture (Gouws, 2013) and metacognitive awareness (Gavriilidou, Markos, & Konstantinidou, 2024).

Against this background, *100 Practice Cards for Dictionary Use and Vocabulary Development* (Gavriilidou & Konstantinidou, 2025) was developed as a strategy-based instructional resource translating pedagogical lexicography and strategy instruction into contextualized, collaborative, and reflective classroom activities. This paper outlines its pedagogical rationale, theoretical foundations, and instructional design; evidence of its implementation and effectiveness has been reported elsewhere (Konstantinidou, 2023; Gavriilidou, Markos & Konstantinidou, 2024).

2 Theoretical and Research Background

Research in pedagogical lexicography views dictionaries as dynamic learning tools supporting language development, vocabulary learning, and learner autonomy (Béjoint, 2010; Lew, 2013). Effective dictionary consultation extends beyond meaning retrieval, requiring learners to

interpret lexicographic information, select appropriate meanings, and evaluate lexical choices in context (Campoy-Cubillo, 2015; Tono, 2001).

Within this framework, the concepts of dictionary culture (Gouws, 2013) and Dictionary Use Strategies (DUS) (Gavriilidou, 2013) are central. DUS include dictionary awareness, dictionary selection, lemmatization, and look-up strategies, highlighting dictionary use as a complex learning activity involving problem-solving, self-regulation, metacognitive awareness, and informed lexical decision-making (Chadjipapa et al., 2020).

The material was informed by empirical research on Greek upper-primary and lower-secondary learners. Previous studies showed that students relied mainly on basic look-up procedures while underusing dictionaries for grammatical information, collocations, semantic relations, and contextual variation (Chadjipapa et al., 2020; Gavriilidou, 2014). Intervention studies further demonstrated that explicit instruction in dictionary use strategies enhances strategic behaviour, metacognitive awareness, dictionary culture, and learner autonomy (Gavriilidou & Konstantinidou, 2021; Konstantinidou, 2023; Gavriilidou, Markos & Konstantinidou, 2024).

While these findings continue to underpin strategy-based dictionary instruction, recent technological developments have transformed learners' access to lexical information. The emergence of generative artificial intelligence (GenAI) has introduced new conditions for vocabulary learning, as large language models increasingly provide lexical information, definitions, usage examples, and language support traditionally sought through dictionaries (Lew, 2024; Rees & Lew, 2024). However, access to lexical information alone does not ensure informed lexical decision-making. Recent research shows that learners should critically evaluate AI-generated lexical information by assessing alternative lexical choices, interpreting meanings in context, verifying the appropriateness and reliability of AI-generated information, and making informed linguistic decisions (Lew, Ptasznik & Wolfer, 2024; Ptasznik, Wolfer & Lew, 2024). These competences closely correspond to the strategic and metacognitive skills fostered through explicit dictionary instruction. Accordingly, *100 Practice Cards for Dictionary Use and Vocabulary Development* translates these theoretical, empirical, and technological developments into a classroom resource designed to cultivate strategic, critical, and autonomous dictionary users. The DUS promoted through the material therefore extend beyond traditional dictionary consultation and support the critical use of AI-generated lexical information.

3 Description of the Material

3.1 Target users and structure

The book *100 Practice Cards for Dictionary Use and Vocabulary Development* consists of 100 instructional cards designed for upper elementary lower and secondary pupils.

The cards are organized into three proficiency levels corresponding approximately to Grades 4, 5, and 6/lower secondary education. This graded organization supports developmental progression, differentiated instruction, and age-appropriate language learning practices (Tomlinson, 2011, 2014). The progression additionally reflects the gradual development of dictionary literacy, beginning with foundational dictionary awareness and moving toward more strategic and autonomous dictionary use.

The structure of the material is informed by learner-centred and strategy-based approaches to instruction. Rather than treating dictionary consultation as an occasional reference activity, the

material conceptualizes dictionary use as a systematic learning practice that can be explicitly taught, practised, and gradually internalized by learners.

3.2 Organization of the cards

Each instructional card extends across two pages and follows a consistent pedagogical structure intended to promote learner familiarity, instructional coherence, and gradual strategy development. More specifically, the cards incorporate:

- explicit learning objectives,
- thematic vocabulary linked to the school curriculum,
- dictionary-based activities,
- vocabulary-development tasks,
- collaborative learning activities,
- visual scaffolding and illustrations.

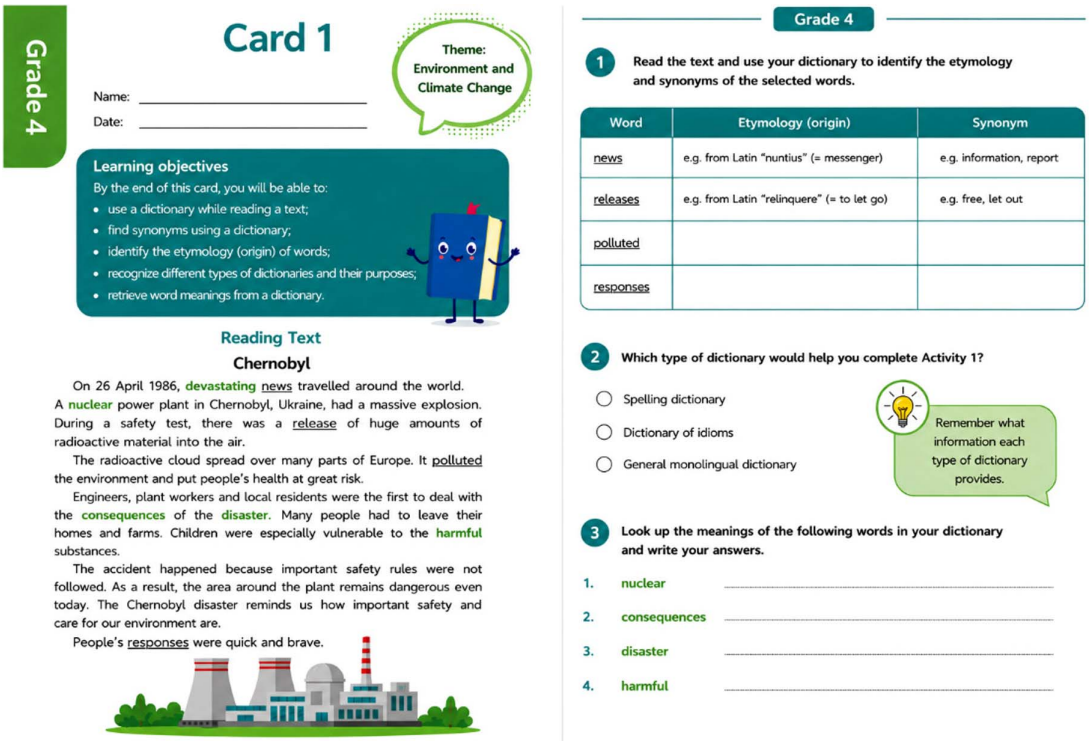


FIGURE 1 Adapted sample instructional card (Grade 4) illustrating the integration of dictionary use strategies, vocabulary development, and strategy-based instruction (adapted from Gavriilidou & Konstantinidou, 2025)

Figure 1 illustrates a representative instructional card from the material. Each card combines a short contextualized reading text or a multimodal text with a sequence of explicit dictionary consultation activities designed to promote the gradual development of Dictionary Use Strategies (DUS). In the present example, learners are guided to identify the appropriate dictionary type, explore etymological information, locate synonyms, and retrieve word meanings through purposeful dictionary consultation. Rather than practising isolated look-up

procedures, the activities encourage learners to interpret lexicographic information, compare lexical alternatives, and justify their lexical choices within context. This integrated design operationalizes the principles of strategy-based instruction by simultaneously supporting vocabulary development, metacognitive awareness, and learner autonomy.

As illustrated in Figure 1, the inclusion of explicit learning objectives reflects strategy-based approaches to instruction according to which clearly defined goals enhance learner awareness, engagement, and self-regulation. Learners are therefore encouraged not only to complete activities, but also to understand the purpose of dictionary consultation and the strategies involved in successful lexical problem-solving.

The thematic organization of vocabulary allows learners to process lexical items within semantic and communicative networks rather than as isolated units, thereby facilitating contextualized vocabulary learning and deeper lexical retention (Webb & Nation, 2017). At the same time, the material incorporates collaborative and strategy-oriented activities, including pair and group work, guided discovery tasks, problem-solving activities, reflective questions, and vocabulary-development exercises. These activities seek to promote interaction, negotiation of meaning, active learner engagement, and learner autonomy, while simultaneously supporting flexible classroom implementation across different educational contexts (Gavriilidou, Markos & Konstantinidou, 2024; Mercer & Howe, 2012).

3.3 *Visual and multimodal design*

Particular emphasis is placed on visual scaffolding and multimodal support throughout the instructional material. Illustrations, structured layouts, and graphic organization are integrated into the cards to facilitate comprehension, reduce cognitive load, and sustain learner motivation, particularly among younger learners and developing dictionary users (Mayer, 2009). This design reflects contemporary perspectives on multimodal literacy and multimedia learning, according to which meaning is constructed through the interaction of multiple semiotic modes (Kress & van Leeuwen, 2006; Serafini, 2014). The visual design also supports dictionary consultation by presenting lexical and lexicographic information in a clear and accessible format, integrating textual, visual, and collaborative elements into meaningful classroom activities that promote learner engagement, vocabulary development, and strategic dictionary use.

4 Pedagogical and Lexicographic Contribution

A central aim of the material is to promote strategic dictionary use by helping learners develop dictionary consultation strategies, including identifying lemmas, distinguishing among meanings, interpreting grammatical information, using contextual clues, and navigating printed and digital dictionaries. Through repeated engagement with dictionary-based tasks, learners progressively develop dictionary awareness and become more capable of selecting relevant lexical information according to communicative context and learning purpose (Gavriilidou, 2013; Lew, 2013). Introductory cards further familiarize learners with dictionary structure, alphabetical ordering, abbreviations, and lexicographic conventions, supporting a gradual progression towards strategic and autonomous dictionary use (Gavriilidou, 2013; Lew, 2013).

Particular emphasis is placed on contextual meaning selection, encouraging learners to recognize that dictionary consultation involves interpretive judgement rather than mechanical meaning retrieval. By comparing meanings, analysing contextual clues, and justifying lexical choices, learners engage in deeper lexical processing and more flexible vocabulary

use (Tono, 2001). Beyond vocabulary development and dictionary culture, the material promotes broader soft skills that are increasingly regarded as essential for academic success, lifelong learning, and active participation in contemporary knowledge societies. Strategic dictionary use requires learners to identify lexical problems, evaluate contextual information, compare meanings, justify lexical choices, and monitor their searches, thereby fostering critical thinking, problem-solving, decision-making, self-regulation, and metacognitive awareness (Tono, 2001). Collaborative dictionary activities further promote communication and interpersonal skills through negotiation of meaning, explanation of lexical choices, and joint problem-solving, while supporting the transfer of dictionary use strategies to reading comprehension, text production, information literacy, and independent learning.

Differentiated and collaborative instructional techniques, including pair and group work, think–pair–share activities, jigsaw tasks, Frayer models, and game-based vocabulary activities, are systematically integrated into the instructional design to promote learner participation, collaborative meaning-making, and flexible implementation across diverse learning profiles and classroom contexts, in line with the principles of differentiated instruction and inclusive pedagogy (Florian & Beaton, 2018; Tomlinson, 2014).

Finally, learner autonomy, metacognitive awareness, and dictionary culture are further strengthened by encouraging learners to monitor search effectiveness, compare meanings, justify lexical choices, and reflect on their dictionary consultation strategies. Dictionary use is therefore approached as a reflective process of lexical analysis, contextual interpretation, and informed decision-making rather than as a mechanical reference activity (Lew & De Schryver, 2014; Müller-Spitzer, Koplenig & Töpel, 2015).

5 Conclusion

The book *100 Practice Cards for Dictionary Use and Vocabulary Development* demonstrates how pedagogical lexicography can be transformed into meaningful classroom practice through strategy-oriented and learner-centered instructional material. Grounded in research on dictionary use, vocabulary acquisition, and strategy-based instruction, the material aims to cultivate strategic and autonomous dictionary users while promoting soft skills, vocabulary development, metacognitive awareness, critical thinking, and learner autonomy. More broadly, it highlights the educational potential of dictionaries as dynamic learning tools that support language awareness and lifelong learning.

References

Béjoint, H. (2010). *The Lexicography of English: From Origins to Present*. Oxford University Press.

Campoy-Cubillo, M. C. (2015). Assessing dictionary skills. *Lexicography ASIALEX*, 2(1), 119–141.

Chadjipapa, E., Gavriilidou, Z., Markos, A. & Mylonopoulos, A. (2020). The effect of gender and educational level on dictionary use strategies adopted by upper-elementary and lower-secondary students attending Greek schools. *International Journal of Lexicography*, 33(4), 443–462. https://doi.org/10.1093/ijl/ecaa021

Florian, L. & Beaton, M. (2018). Inclusive pedagogy in action: Getting it right for every child. *International Journal of Inclusive Education*, 22(8), 870–884. https://doi.org/10.1080/13603116.2017.1412513

Gavriilidou, Z. (2013). Development and validation of a questionnaire for the measurement of dictionary use strategies. *Research Papers in Language Teaching and Learning, 4*(1), 549–568.

Gavriilidou, Z. (2014). User's abilities and performance in dictionary look-up. In N. Lavidas, T. Alexiou & A. Sougari (Eds.), *Major Trends on Theoretical and Applied Linguistics 2* (pp. 41–52). De Gruyter Open.

Gavriilidou, Z. & Konstantinidou, E. (2021). The design of an explicit and integrated intervention program for pupils aged 10–12 with the aim to promote dictionary culture and strategies. In Z. Gavriilidou, L. Mitits & S. Kiosses (Eds.), *Proceedings of the XIX EURALEX International Congress: Lexicography for Inclusion* (pp. 735–745). SynMorPhoSe Lab, Democritus University of Thrace.

Gavriilidou, Z. & Konstantinidou, E. (2025). *100 καρτέλες εξάσκησης στη χρήση του λεξικού και ανάπτυξης λεξιλογίου.* Δίσιγμα.

Gavriilidou, Z., Markos, A. & Konstantinidou, E. (2024). How can we raise strategic dictionary use in the classroom: The effect of a dictionary awareness program on dictionary use strategies. *Lexikos,* 34(1), 99–122. https://doi.org/10.5788/34-1-1882

Gouws, R. H. (2013). Dictionary use and dictionary culture. In R. H. Gouws, U. Heid, W. Schweickard & H. E. Wiegand (Eds.), *Dictionaries: An International Encyclopedia of Lexicography* (pp. 1273–1281). De Gruyter.

Konstantinidou, E. (2023). *Research on the effect of an explicit and integrated dictionary awareness intervention program on primary school students' dictionary use strategies* [Doctoral dissertation, Democritus University of Thrace]. National Archive of PhD Theses.

Kress, G. & van Leeuwen, T. (2006). *Reading Images: The Grammar of Visual Design* (2nd ed.). Routledge.

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, 11(1), 1-8.

Lew, R. (2013). Online dictionaries of English. In R. H. Gouws, U. Heid, W. Schweickard & H. E. Wiegand (Eds.), *Dictionaries: An International Encyclopedia of Lexicography* (pp. 2306–2314). De Gruyter.

Lew, R., Ptasznik, B., & Wolfer, S. (2024). The effectiveness of ChatGPT as a lexical tool for English, compared with a bilingual dictionary and a monolingual learner's dictionary. *Humanities and Social Sciences Communications*, *11*(1), 1-10.

Lew, R. & De Schryver, G.-M. (2014). Dictionary users in the digital revolution. *International Journal of Lexicography, 27*(4), 341–359. https://doi.org/10.1093/ijl/ecu011

Mercer, N. & Howe, C. (2012). Explaining the dialogic processes of teaching and learning: The value and potential of sociocultural theory. *Learning, Culture and Social Interaction, 1*(1), 12–21. https://doi.org/10.1016/j.lcsi.2012.03.001

Müller-Spitzer, C., Koplenig, A. & Töpel, A. (2015). Online dictionary use: Key findings from an empirical research project. In I. Kosem, M. Jakubíček, J. Kallas & S. Krek (Eds.), *Electronic Lexicography in the 21st Century: Linking Lexical Data in the Digital Age* (pp. 425–457). Bloomsbury.

Ptasznik, B., Wolfer, S., & Lew, R. (2024). A learners' dictionary versus ChatGPT in receptive and productive lexical tasks. *International Journal of Lexicography*, *37*(3), 322-336.

Rees, G. P., & Lew, R. (2024). The effectiveness of OpenAI GPT-generated definitions versus definitions from an English learners' dictionary in a lexically orientated reading task. *International Journal of Lexicography*, *37*(1), 50-74.

Serafini, F. (2014). *Reading the Visual: An Introduction to Teaching Multimodal Literacy*. Teachers College Press.

Tomlinson, B. (2011). *Materials Development in Language Teaching* (2nd ed.). Cambridge University Press.

Tomlinson, C. A. (2014). *The Differentiated Classroom: Responding to the Needs of All Learners* (2nd ed.). ASCD.

Tono, Y. (2001). *Research on Dictionary Use in the Context of Foreign Language Learning*. Max Niemeyer Verlag.

Webb, S. & Nation, I. S. P. (2017). *How Vocabulary Is Learned*. Oxford University Press.

Contact information

Zoe Gavriilidou
Greece Democritus University of Thrace
zoegab@otenet.gr

Evanthia Konstantinidou
Greece Democritus University of Thrace
evikon1990@gmail.com

Laura Rebosio

OSTENSIVE AIDS IN LEARNER'S DICTIONARIES

A Diachronic, Comparative Analysis

Abstract This paper proposes a diachronic analysis of the use of ostensive aids in 35 print editions and the current online versions of six major English monolingual learners' dictionaries (Oxford, Longman, Collins, Cambridge, Macmillan, Merriam-Webster). Alongside macrostructural features of illustrations (density, types, placement, labeling), I analyze their content, performance, and consistency with respect to three questions: (1) whether principles for selecting lemmas to illustrate are stated or inferable; (2) how polysemy and semantic fields are represented; and (3) how culture-specific concepts are handled. The findings show a persistent lack of explicit, operational selection criteria or principles (with LDOCE2 as a partial exception), a substantial reduction in the use of illustrations in the microstructure in print editions, and an unsystematic implementation of ostensive aids in online editions despite the absence of space constraints. The treatment of polysemy, prototypes, and culture-bound items remains uneven. The present analysis is a preliminary step toward outlining a theoretically grounded methodology for the implementation of ostensive aids in the *Phrase-based Active Dictionary* (see Giacomini, 2025; DiMuccio-Failla, 2025).

Keywords Illustrations; english learners dictionaries; prototypes; cultural connotations; polysemy

1 Introduction

In this paper, I present the results of a comprehensive diachronic analysis of the use of illustrations in all the print and online versions of English monolingual learner's dictionaries of the six major publishing houses in the anglophone countries—Oxford University Press, Pearson Longman, Collins, Cambridge University Press, Macmillan Education, and Merriam-Webster Incorporated. The analysis spans the entire range from their first print editions until their latest ones, covering every print edition (35 dictionaries in total). Online versions, if still accessible, are analyzed too. The aim of this analysis is, on the one hand, to identify good practice in older editions but later abandoned, on the other hand, to highlight methodological gaps that are still an issue in modern practical lexicography.

Ten years ago, Klosa (2016, pp. 530–531) stated that "the question of which factual information and how many and which illustrations belong in the dictionary has not yet been definitively answered and must be decided on again each time depending on the intended user group and function of the dictionary" and that "sound criticism of practice up until now as well as thorough research into recently developed types of information and forms of presentation in both printed and electronic dictionaries are essential." This statement still holds today, since in theoretical lexicography, no systematic and operational criteria have yet been formulated that regulate when the use of ostensive aids is warranted (for which kind of entries/senses), which type of aid is most suitable for a specific meaning disambiguation, where to place the visual aid in the microstructure, and how to manage cross-references among visual representations across entries.

In practical lexicography, no specific information about the criteria for the use of illustrations is given in any of the User Guides of the analyzed editions or online versions of the dictionaries: why were some entries chosen to be visually represented and others not? Print editions show a clear contraction in the use of illustrations (OALD, for example, went from 66 illustrations

across a section of 41 pages in 1948 to 6 for the corresponding section in the 10th edition in 2020). This can be sensibly explained through space and cost constraints, but even in online versions, where space is not a constraint, the use of illustrations has not been systematically implemented yet, and the use of more modern types of visual aids (videos, animations) or, lately, of AI-generated visuals, remains largely underexplored.

In Section 2, I will briefly summarize the advantages and disadvantages of the use of illustrations in dictionaries, and the results of important user studies conducted on the consultations of entries with and without ostensive aids (and different types of them). I also list the problems and the main questions that come up in practical lexicography concerning their use. In Section 3, I introduce the dictionaries used for my analysis, the research questions, and the outcomes. In Section 4, I discuss the outcomes and present the next steps to take for a systematic and theoretically motivated implementation of ostensive aids in the *Phrase-based Active Dictionary* (PAD), a monolingual dictionary model developed by Giacomini and DiMuccio-Failla (cf. DiMuccio-Failla & Giacomini, 2017, 2022; DiMuccio-Failla, 2025; Giacomini, 2025) as one of the three components of PhraseBase—a model of a phraseologically and cognitively oriented Linguistic Information System for different languages. Section 5 gives concluding remarks.

2 The use of illustrations in dictionaries

The history of the use in illustrations in dictionaries begins in 1850, when the first illustrated dictionary—the *Imperial Dictionary*, edited by J. Ogilvie—was brought on the market in Scotland;[1] France followed in 1852, Spain in 1857, the U.S. in 1860 (*A dictionary of the English Language*, edited by J.E. Worcester, was published in Boston, MA, and in London, UK); Italy in 1891, and Germany in 1931 (cf. Hupka, 1989, pp. 120–122, 124, 135, 138, 141). The illustrations contained in the first dictionaries were realized using woodcut and wood engraving techniques; line drawings by artists and halftone photographs were slowly integrated around the beginning of the new century and in a different pace from country to country (cf. Hupka, 1989, Chapter 6.6). Before 1850, illustrations were a feature typical of encyclopedias.

Hupka's 1989 monograph offers the first comprehensive survey, historical account, and critical review of the use of illustrations in the printed lexicographic works available to him at the time. His research was of great relevance, because it brought into focus a topic that had been largely neglected in lexicography until that moment (Stein, 1991, p. 99, speaks of "a curious oversight in lexicographic research"). The research in the field of illustrations (or more generally, of ostensive aids) in lexicography can be subdivided in two groups: the first aims at the classification of illustrations and the creation of illustration typologies. Within this line of research three major works are still relevant today: Ilson (1987), Hupka (1989), and Stein (1991), and I will come back to these in Section 4. The second group comprises more recent studies that aim to test the efficacy of ostensive aids in meaning reception and retention, and the user's interaction with them. I will give a brief overview of these in Section 2.1, and in Section 2.2 I will delineate a range of concrete issues in practical lexicography.

2.1 Studies on the use of illustrations

Two main names emerge consistently connected with user studies concerning ostensive aids in online dictionaries: Anna Dziemianko and Robert Lew (cf. also the latest brief overview by Rees,

[1] The basis of the *Imperial Dictionary* was Webster's 1841 edition, and it was seen as plagiarism by the publishers of the Webster's in 1859 (cf. Hupka, 1989, pp. 120–121).

2025). They dealt extensively with this topic seeking answers about, firstly, the fundamental question whether the integration of ostensive aids in dictionaries is indeed beneficial to the users (backing or rejecting existing theories with empirical data). Secondly, they sought evidence in the users' behavior to provide practical clues for the design of online dictionaries that aim to integrate ostensive aids: what kind of typologies work better than others? What are the user's main gaze patterns if a given entry contains visual elements? When and how should visual elements be displayed to be most effective? In the following paragraphs I summarize the main findings.

The advantages of dictionary illustrations adduced in lexicographic literature and cognitive science are outlined in detail in Dziemianko (2022) and corroborated with the relevant sources which will not be repeated here (an older, more concise overview is in Svensén, 2009, pp. 298–300). Dictionary illustrations support meaning comprehension and retention, enhance information retrieval, support disambiguation, elucidate and complement verbal definitions; they can be used to associate unfamiliar words with concepts, illustrate semantic fields, deliver cultural information, and to add an aesthetic value to the medium itself (Dziemianko, 2022, pp. 218–222).

Paivio's (2010, p.211) *Dual Coding Theory*—developed since the 1970s—holds that, when information is processed through both the verbal and nonverbal systems simultaneously, the memory trace is stronger than if it is processed through only one. Deductively, an extensive use of visual aids in dictionaries would be particularly conducive to concept & word retention. Nesi (1998) showed with a user test that illustrations in dictionary entries help the immediate recognition of everyday objects; Gumkowska (2008) repeated Nesi's test but also added L1 equivalents, and showed that this constellation promotes vocabulary acquisition. A cautious enthusiasm regarding the use of animations was expressed by Lew and Doroszewska (2009, p. 254) on the base of the results of a user test showing that the best predictor of vocabulary retention was not the consultation of animations, but of the combination of both the L1 translation and the L2 definition of the tested words, followed in second position by the consultation of the L1 translation only.

Dziemianko (2022, p. 220, 2024, pp. 33, 35) outlines critical positions in non-lexicographic research and studies that underline the disadvantages of a multimodal approach, such as cognitive load, distraction, and the lack of added value in comparison to single modes alone (verbal or visual). Against the background of the dissonant positions about multimodality fostering language learning, Dziemianko (2022) conducts a user test to assess how to design or use illustrations in dictionaries that reduce extraneous cognitive load. The results show that black-and-white drawings are more conducive to meaning reception and immediate and delayed retention than color and grey-scale illustration (in 2nd and 3rd position respectively).

Lew et al. (2018)—following up on and extending a study done by Kemmer (2014)—conduct an eye-tracking test to assess the competition for users' attention between verbal definition and illustration, and conclude that, when consulting entries with visual aids, 40% of dwelling time is spent on the illustration, that this gaze pattern (i.e. whether looking at illustrations or verbal definition first) is a matter of personal preference, and that visual aids do not "steal away" the focus from the definitions. Basically, it is confirmed that verbal and pictorial information complement each other. Dziemianko (2024) investigates the best access path to illustrations, concluding that instantly visible pictures are most recommendable—backing in this way the superior effects of multimodal processing in information comprehension and retention.

Even though vocabulary retention and language learning are welcome effects of dictionary consultation, they are "often not the primary, or immediate, goal" of the users (Lew & Doroszewska 2009, p. 254). Therefore, the purposes of learners using monolingual learner's dictionaries may shift mainly on meaning disambiguation, or language production. Another point that emerges from these recent studies is that, when it comes to meaning retention, the highest scores are achieved when an L1 translation is accessible. This piece of evidence leads implicitly to the broader question of the role of L1 in SL learning in adults—obviously different from language acquisition.

Considering the empirical data delivered for lexicography (which doubtless help steer dictionary design in the right direction), a major suspended question is still open, i.e. what kind of entries/senses need visual support. In the above-mentioned studies, the lemmas that had been chosen for the testing were associated with concepts of everyday items that are better shown visually than described verbally (cf. Nesi, 1998); or those with a low frequency of use; or they were unknown to the target subjects of the tests. We could deduce that physical appearance of objects and frequency of a word use could be criteria for the choice of lemmas to illustrate in learner's dictionaries, but these criteria need to be extended and additional ones are currently being tested (cf. Rebosio, in press). Depending on the type of target users and dictionary type, a different use of illustrations is expected.

2.2 Issues in practical lexicography

The use of ostensive aids raises broader theoretical issues related to polysemy, prototypicality, and cultural specificity that have then concrete repercussions in practical lexicography. The disambiguation of senses of a polysemous word is challenging if the word is taken in isolation and not as a phraseological expression—i.e. embedded in its typical syntactic-semantic context (cf. DiMuccio-Failla & Giacomini, 2017). In their pilot study designed to evaluate AI image-to sense alignment for polysemous noun entries, Perdih et al. (2025, p. 862) found that not only AI had trouble to assign the right picture to the entry of SI *akvarel* (EN 'aquarelle'/'watercolor'), but also the human annotators could not decide whether the selected picture represented *akvarel* as a painting technique or the physical painting itself. In this case indeed, depending on the syntactic-semantic environment in which the lexical unit is embedded, two distinct senses—and therefore two distinct illustrations or ostensive aids—should be distinguished and presented. However, the questions then would be, firstly, what is the most appropriate ostensive aid for 'watercolor painting technique' (a picture or a video?), and secondly, how to select the best candidate within the chosen typology of ostensive aid: is a box of watercolors representative for the painting technique? How much information can be presupposed and inferred by the users when symbolic illustrations are used varies from concept to concept, and the lexicographers should consider their choice.

The presentation of polysemy and the depth of fine-grained semantic distinctions depend on the type of dictionary, the compilation methodology, and the theoretical orientation. If the senses of a given entry are formulated as syntactic-semantic patterns (containing the word used in one, specific, and univocal sense, as in the PAD model, cf. DiMuccio-Failla, 2025, p. 19), then the image-to-sense alignment should be univocal too. Ideally, an ostensive aid is added every time it is relevant to meaning disambiguation, that is, when it conveys additional information that is not contained in the verbal definition, when it displays information which is implicit in it, or when it shows many more typical details than those normally included in a definition. For some concepts, even long definitions are not necessarily exhaustive descriptions. For example, even a long, detailed description of *cartwheel (GYMN.)* cannot convey the various intermediate phases of the movement or the differentiated posture of the body during

it (which, is, to some extent, straightened and to some extent not) as effectively as a sequential illustration can do immediately and without verbal dispenses (cf. Hupka, 2003, p. 366).

Prototypes are another challenging issue to deal with. They come along with well-known theoretical questions, such as how to determine the relationship of an item (or a concept) to the prototype itself, that is, how to select the attributes of centrality and typicality necessary for belonging to the category (cf. Labov's vagueness measurement of the concept *cup*, 2004). The fuzziness of prototypes has, in fact, boundaries too.

Additionally, prototypes that do not refer to, for example, taxonomies, are culturally entrenched. In line with Wierzbicka (2014, p. 352): "cultural models are important indeed, but they are not 'another important factor', in addition to meaning. Cultural models are reflected in the meanings of words." Cultural information—which often finds no place in verbal definitions but is essential for sense disambiguation and conceptual understanding—can be conveyed through visual aids. For example, the picture of a prototypical car in the U.S. differs from that in Italy. In the U.S., *car* may be associated with a Chevrolet or a GMC SUV; in Italy, *automobile/macchina* is more likely to be associated with other brands and more compact models. This difference is not explicitly stated in a verbal definition, but it can be communicated through culturally adapted illustrations. Speaking about cars, also the position of the steering wheel, on the right or on the left, could be displayed accordingly to the country regulations of the dictionary language, thus delivering additional information that is not, strictly speaking, necessarily constitutive of the prototype itself, but stays true to the culture the language is associated with. A consistent consideration of cultural specifics currently lacks in English learner's dictionaries (cf. Section 3.2.3).

A general issue connected with the visual representation of prototypes is that they can trigger connotations that are unintentional. When choosing a picture featuring human beings, stereotypes, gender, unintended body language, unintended power dynamics, and even the perspective of the general composition should be carefully considered. For example, the picture of *body* in OALD10 and OALDo is identical and unchanged, depicting a young white male body. Additionally, unintentional cultural connotations related to clothing, styling, architecture, religious and historical (national) symbols, myths, etc. must be considered too.

The more variables to consider, the slower does the implementation of policies for ostensive aids in a dictionary get. Systematic and operational criteria that are theoretically motivated are therefore needed to formalize the workflow.

3 Analysis of print and online English learner's dictionaries

In this section, I will describe the points of analysis and present the results elicited by the research questions. Due to space constraints, it is impossible to discuss the peculiarities of each single dictionary, but the practices that stand out positively or negatively will be highlighted. Also, I will focus on the evolution of the use of illustrations across the editions and the differences between publishing houses. In Section 3.1, I list the dictionaries I chose for the analysis; in Section 3.2 the fundamental questions that guided it.

3.1 Source selection and motivation

The analysis that I present focuses on print and online monolingual English Learner's dictionaries of the six major publishing houses in Great Britain and the U.S.:

1. *Oxford Advanced Learner's Dictionary* (OALD): all ten print editions starting from the very first one in 1948 until the 10th in 2020, and the current online version of the dictionary;[2]
2. *Longman Dictionary of Contemporary English* (LDOCE): all six print editions starting from 1978 to 2014, and the current online version of the dictionary;
3. *Collins Cobuild Learner's Dictionary* (COBUILD): all ten print editions starting from the first one in 1987 until the 10th in 2023, and the current online version of the dictionary;
4. *Cambridge Advanced Learner's Dictionary* (CALD): the very first 1995 edition (not numbered), the following four print editions until the latest one in 2013, and the current online version of the dictionary;
5. *Macmillan English Dictionary for Advanced Learners* (MEDAL): the two print editions of 2002 and 2007 (the online version was shut down in 2023);
6. *Merriam-Webster's Advanced Learner's English Dictionary* (MWALED): the two print editions of 2008 and 2016, and the current online version of the dictionary.

The analysis of the dictionaries was twofold: on the one hand, macrostructural characteristics of illustrations such as density, types, position in the microstructure, and the use of labels were recorded for each edition; on the other, the content, performance, and consistency of use of illustrations was assessed.

3.2 Research questions

The main research questions concern the content, performance, and consistency of use of ostensive aids, which in the above-mentioned print and online editions consist solely of illustrations. Three questions were formulated:

(1) Is there a description (e.g. in the introductory section or in the User Guides) of the practical conventions applied for the use of illustrations? If not, can plausible underlying principles be inferred on the basis of the actual use?

(2) How is the representation of semantic fields and polysemy dealt with?

(3) How is the representation of cultural elements that can be relevant for learners dealt with?

In order to assess (3), the following lemmas were looked up: *carnival*, *coffee*, *tea*, *football* as *soccer*, *football* as *American football*, *cricket*, *scone*, *Super Bowl*, and *punk* (n.). These lemmas stand for concepts that vary from culture to culture to a different degree. Some of them are specific concepts of British culture (*tea, cricket, football/soccer*); some of American culture (*American football*, *Super Bowl*); some of both (*scone*). In these cases, it is interesting to see how the dictionaries present these concepts to the target users (who may have a different cultural background), and how much cultural knowledge has been presupposed in the users.

Coffee is culture-specific in the way it is consumed. If the way of consumption is not contained in the verbal definition, it can be implicitly delivered through an illustration, i.e. by depicting a filter coffee in English learner's dictionaries (and not an espresso or a Turkish coffee).

[2] The 1948 edition is not, technically speaking, numbered as "first". The 1952 reprint bears a new title—*The Advanced Learner's Dictionary of Current English*—and is numbered as the first in the series (OALD1).

Carnival is neither a British nor an American cultural-specific concept: in this case, it is interesting to see how it is dealt with a foreign cultural-specific concept. How is it presented to the users? For *carnival*, there can be no unified visual representation, since it strongly varies in at least three cultures: Brazilian (*carneval do Rio de Janeiro*), Italian (*carnevale di Venezia*), and German *(Kölner Karneval*).

Finally, *punk* (n.) as a music genre, is a culture-specific concept that originated simultaneously in London and NYC in 1976 and has since spread worldwide, acquiring distinctive features in every culture where the genre has taken root. In this case, we deal with a culture-specific concept from an historical perspective. The punk artists' appearance and the music itself make them particularly apt for a visual representation (e.g. through a picture in print editions, or through a video in online ones).

Cultural information can be implied by examples; or explicitly stated in the form of an encyclopedic short text in a separate box placed in the microstructure; or, more subtly, through regional labels ("UK"/"US"), or through cross-references to other entries (e.g., in OALDo 2026: *coffee*, we find cross-references to *dandelion coffee*, *Irish coffee*, and *Turkish coffee*). In my analysis, I looked solely at the use of ostensive aids for the representation of cultural specificities.

3.2.1 Motivation for the illustrated lemmas

None of the dictionaries states explicitly how the selection of lemmas to illustrate was made, nor is it possible to infer the principles that may have regulated this selection. LDOCE2 gives more details about the guidelines for including illustrations, but there is no systematicity in their application (cf. the illustrations of animals). In the introductory text (LDOCE2, 1987, p. F49), we read that, besides common animals, plant and objects, illustrations are used for "things that are not easily explained in words, such as shapes, complex actions, or small differences between words that are similar but not the same", groups of related objects (to display their similarities or differences in shapes and forms), and, finally, for showing "the basic or physical meaning of words that are commonly used in an abstract or figurative way, with the aim of making these abstract uses easier to understand." In the other editions of LDOCE, illustrations are not mentioned in the preface.

In COBUILD6 to COBUILD10, the introduction contains descriptions of the types of illustration used—frames/scenes and partonomies (termed "Picture Dictionary Boxes" here), and semantic fields (termed "Word Webs" here, i.e. a group of tokens/items of the same type, e.g. COBUILD9: *clocks*, or a group of items that belong to a thematic macro-category, e.g. COBUILD8: *jewellery*) which are accompanied by a short text with encyclopedic information—and their purposes, but provides no information about the selection of lemmas.[3] In the first three editions of OALD, only the fact that the dictionary contains illustrations is mentioned (in subsequent editions, illustrations are not mentioned). CIDE contains a full list of illustrations (160 in total), later editions do not mention them in the introduction, or they motivate their presence superficially (cf. CALD3, p. ix), "illustrations [concentrate] on small items that are more easily understood in pictures than in words").[4] MEDAL1 and MEDAL2 do not mention illustrations in the

[3] In COBUILD1 to COBUILD5 there are no illustrations in the microstructure, possibly due to the presence of the *extra column* (and only COBUILD4 and COBUILD5 have an end section of eight pages of color illustrations).

[4] The first learner's dictionary published by Cambridge is CIDE, the *Cambridge international dictionary of English* (1995). As for Oxford, this first edition is not, technically speaking, numbered as "first". The following edition, published in 2003, bears a new title—*Cambridge Advanced Learner's Dictionary*—and is numbered as the first in the series (CALD1).

introduction, and MWALED1 and MWALED2 only explicitly mention the people involved in the drawing process or selection of illustrations. In all online editions of the dictionaries there is no information about the use of illustrations.

The principles behind the selection of lemmas to illustrate is not inferable. I looked at the entries of *camel*, *dromedary*, *elk*, and *moose*: their similarities can make them apt to representation for meaning disambiguation. Additionally, what is called *elk* in BE is called *moose* in AE (referent: Alces Alces); and what is called *elk* in AE is called *moose* in BE (referent: Cervus canadensis,) (cf. Encyclopædia Britannica). *Wapiti* is sometimes used to refer to Cervus canadensis. An illustration of the referents is in this case particularly relevant for learners. The entry in OALDo (2026: *elk*) mirrors the entries in OALD8, OALD9, and OALD10, and neither the definition nor the illustration helps the user disambiguate the meaning. For LDOCE, only LDOCE1 has pictures for *camel*, *dromedary* (and *deer*) under *ruminant* (cross-referenced); LDOCE2 has no illustrations for these lemmas, but has, for example, illustrations for *crocodile* and *elephant*. LDOCE5 depicts only *moose*. LDOCEo depicts only *camel*. The print editions of COBUILD do not illustrate these lemmas (*dromedary* is not even lemmatized). COBUILDo illustrates all of them (however, the picture of *camel* is identical to the one for *dromedary*). CALDo is the most consistent dictionary: it has pictures for all four lemmas.[5] In its print editions, only *camel* is consistently depicted. MEDAL1 and MEDAL2 do not illustrate these lemmas. MWALED1 and MWALED2 illustrate the semantic field *deer*, where *deer*, *moose* and *elk/wapiti* are depicted.

The overall impression is that, in practical lexicography, the selection of lemmas to be illustrated has been subject solely to the lexicographer's discretion. Space and cost constraints limit the range of freedom in print editions and online editions bear other challenges (the costs for the access to image databases are still high), but the full potential of their medium has not been exploited yet.

3.2.2 Semantic fields and polysemy

In my analysis, I looked for the use of illustrated semantic fields simply as representations of groups of conceptually or semantically related items. A finer classification of the types of semantic fields entails ontological questions regarding the illustrated concepts associated with the lexical items (cf. Wierzbicka, 1984; Giacomini, 2024). For the purpose of the present paper, it is only important to notice that semantic fields can be of a different nature (in the lexicographic literature, this differentiation is not explicit, see e.g. Stein, 1991). The precise theoretical delimitations between the following groups are currently under investigation and development, but some constitutive differences in the nature of illustrated semantic fields are already evident and warrant the following preliminary macro-distinctions:

a) semantic fields of taxonomic/partonomic nature;

b) semantic fields that display selected instances of a class;

c) semantic fields that display purely semantic differences (peculiar to a language system);

In (a), semantic fields are visual representations of taxonomic relations, where different "types of X" are illustrated and where X is the concept associated with the lexical item (here written as a lemma and repeated a legend). For example, the illustrated semantic field of *headset* in MWALED1 shows two different types of headsets, one as on-ear headphone, and one as a

[5] A qualification should be made: in CALDo, one can select the type of dictionary to look up. If one selects "English" pictures are displayed. If one selects "Learner's Dictionary" no pictures are displayed. It is not clear if this is on purpose or a technical error.

single-ear headset with microphone. Another example is *cabbage* in OALD5, where *cabbage* is depicted next to *cauliflower*, *broccoli*, and *Brussel sprouts*.[6] This relation holds not only for concepts associated with concrete items, but also for abstract concepts that can only be depicted with a symbolic illustration.

In (b), semantic fields are the visual representation of a selection of different concepts belonging to a class (in the common sense of group, collection). For example, for *cereal* in OALD6 there is a semantic field with drawings of *wheat*, *rye*, *barley*, *millet*, *oats*, *maize/corn*, and *rice*. Here, the meaning of *cereal* is not depictable with a single exponent, since it is, semantically speaking, a collective concept which comprises different edible grains. Another example is LDOCE5: *fruit* and LDOCE5: *vegetable(s)*.

In (c), the different senses of a polysemous word are depicted (the polysemy of a lexical item varies from language to language, and the concepts that are visually represented may have an evident conceptual common ground or not). For example, for *blade* in CIDE a *razor blade*, a *blade of grass*, a *blade of a propeller* etc. are depicted.

Almost all dictionaries make use of (a–b) (realized both with drawings or photos), the use of (c) is rarer. The Longman print editions stand out for the careful approach to the representation of all semantic fields, with LDOCE2 being the most homogenous, clear, and detailed dictionary in the print series; in LDOCEo, no semantic fields can be found for the tested lemmas (in general, the use of illustrations in LDOCEo is minimal—the lowest among all online editions).

CALD2 (see, e.g., CALD2: *drum*, *neck*, or *net*) stands out for the use of (c). This practice is already observable in CIDE (see, e.g. CIDE: *blade*, *bill*, *club*) and CALD1 (see *match*), but was abandoned in CALD3 and CALD4.

3.2.3 Culture-specific concepts

Among the lemmas linked to culture-specific concepts, sport games (*football/soccer*, *American football*, and *cricket*) are the most frequently illustrated. *Cricket* is depicted in all ten print editions of OALD, mostly through an illustrated partonomy of the players, the pitch, and the equipment. In COBUILD6 to COBUILD10 only *football/soccer* is depicted. All print editions of LDOCE except for LDOCE4 depict *American football* (*cricket* was depicted only in the first two editions, *football/soccer* only in LDOCE3, LDOCE5, and LDOCE6). *Cricket* is not depicted in MEDAL1 and MEDAL2. MWALED1 and MWALED2 contain cross-references to the equipment (e.g. *ball*, *gloves*—which are depicted) in the entries of these sport games. CIDE/CALD is the most consistent dictionary to show all lemmas in the topic section *sport* (cross-referenced). Among online editions, CALDo proves again to be the most consistent dictionary in dealing with visual representations: all lemmas are depicted (through pictures placed right next to the corresponding sense, and similar in quality and picture composition). In OALDo, there is a link to an illustrated semantic field termed "team sports" with a selection of nine team sports. COBUILDo displays only pictures

[6] The fact that, technically speaking (in the botanic domain), the denomination *cabbage* refers to both the species (referent: Brassica oleracea) and a specific variety of it (referent: Capitata) should not concern us here, since we are not dealing with specialized language (the definition of OALD5: *cabbage* describes solely the referent Capitata) and we focus on the visual representation of the semantic field (here: one referent, *cabbage*, drawn next to other similar referents). It is true that, technically speaking, *cauliflower*, *broccoli*, and *Brussel sprouts* belong to the same taxonomic class *cabbage* (Brassica oleracea), but, again, there is no reference to this information in the definition, and the taxonomic qualification of this type of semantic field refers to the illustration of the type *cabbage* (Capitata).

for *American football* and *football/soccer*: the selected pictures are neither harmonious in quality nor in subject composition. LDOCEo and MWo do not have any pictures for these lemmas.[7]

Moving to the other lemmas, in none of the print editions of the dictionaries, *carnival*, *Super Bowl*, and *punk* (n.) are illustrated. CALDo and COBUILDo each have a picture for *carnival* (the first is of a general carnival mask—very appropriate and most neutral; the second is of a carnival float, supposedly of the carnival of Rio, but the picture has no caption and is too pixelated to be recognized unambiguously).

OALD6, MWALED1 and MWALED2 depict *scone* as part of larger semantic fields in the separate (middle/end) section of (color) illustrations subdivided into different topics (*scone* is cross-referenced to one of these topics—*bread, cake & dessert* for OALD6, and *baking* for MWALED1 and MWALED2). Among online editions, only CALDo and COBUILDo have a picture of *scone*.

As to *coffee* and *tea*, only COBUILD6 to COBUILD9 have boxes with a conceptually-related picture and a short encyclopedic text explaining the gathering process of the beans or leaves from the plant, and the drink preparation.[8] The focus of the descriptions is neutral, neither British nor U.S.-oriented, and the pictures do not reveal any cultural influence. In online editions, CALDo shows a picture of a filter coffee, and of a cup of tea of fine old porcelain with tea in it (under the sense "a cup of tea"). COBUILDo and LDOCEo display similar pictures (i.e. a cup with coffee or tea in it). OALDo has no illustrations for these lemmas; MWo displays a black-and-white drawing of a section of a coffee plant under *coffee*, and one of a tea plant under *tea*.

Overall, the visual treatment of cultural specificities is uneven across publishing houses and editions—ranging from a more careful coverage (notably in the earliest editions of CALD and LDOCE, and in CALDo) to an irregular or absent one. Image quality is variable, ad hoc cross-referencing to middle/end sections seems to be the preferred variant in the recent print editions. Given the prominence of culture-bound concepts in everyday life, dictionaries should adopt a more systematic policy: consistent inclusion in online editions, harmonized visual style and quality, clear captions, and a balanced coverage of regional variants. Such standards would better support users' cross-cultural comprehension and reduce bias.

4 Discussion

The analysis reveals some persistent gaps in dealing with the use of illustrations. Even if for the moment we set aside the lack of systematicity and traceability of criteria for lemma selection, the progressive abandonment of older practices that gave the first, oldest dictionaries a general sense of homogeneity and harmony is the most evident loss. For example, OALD1, LDOCE1, and CIDE used only black-and-white drawings, whose use has been recently proven to be particularly conducive to meaning disambiguation and retention (cf. Dziemianko, 2022). The drawings in OALD1 are very small (2x3 cm) but very neat, precise, and effective because they are placed in near proximity to the entry. The results delivered by Dziemianko (2024) on the advantages of instantly visible pictures in online dictionaries would be positively transferrable into print lexicography as well: earlier dictionaries have—if they do—a very reduced middle/end section with

[7] The *Merriam-Webster.com Dictionary* (MWo) is originally based on *Merriam-Webster's Collegiate® Dictionary* (11th ed.), but "has since been significantly updated and expanded with new entries and revised definitions" (cf. https://www.merriam-webster.com/about-us/faq).

[8] *Coffee* has no illustration in COBUILD9. The short encyclopedic text presents topic related vocabulary (highlighted in bold). This text is framed together with a symbolic picture or an illustrated scene/frame and encourages language exploration or "cognitive framing" (this practice reminds one, at least theoretically, of the work of Ostermann, 2015, on cognitive lexicography).

illustrations, and rely heavily on illustrations in the microstructure; newer editions, instead, prefer middle/end sections and opt more for cross-reference, which is, as we have seen, not optimal (but saves costs). The introduction of pictures/photographs not only diminishes the general harmony and appeal of print dictionaries (if black-and-white photos are mixed with color ones, for example), but also yields questions about their suitability concerning the representation of prototypicality: a drawing has more potential for abstraction which makes it more symbolic, whereas photos suggest always a 1:1 correspondence with the referent, making them less apt for evoking the flexibility of concepts, as argued in Rebosio (in press). If a dictionary is well structured, harmonious in its layout/macrostructure, and attentive and consistent in its ostensive content, it is also more enjoyable to use (cf. the websites of MWo vs. COBUILDo).

Labels (*legend*, *differentiating* labels, *identifying* labels, and *captions*, cf. the terminology used by Stein, 1991), are not used consistently. When illustrating a prototype, the use of a legend with an indefinite article (cf. OALD4: "a church") is better than using no article at all, because the indefinite article suggests that the content of the illustration is only a possible instantiation of the concept—and not the only possible one. In COBUILD9 and COBUILD10, clear legends for semantic fields and partonomies (treated as distinct categories in these editions) are used ("types of X"; "parts of X"), but not in COBUILDo. Finally, if such explicit legends cannot be adopted, the use of plurals is preferable to singular for semantic fields that illustrate types or classes (e.g. "flowers", "vegetables"). In online editions (with the exception of OALDo which uses legends and differentiating labels for its linked illustrations, and of MWo, and COBUILDo) illustrations do not have any legend.

In the lexicographic literature, the most influential typological classification of illustrations was provided by Hupka (1989).[9] Before him, Ilson (1987) suggested another typology by adopting, more or less wholesale, Fillmore's (1978, pp. 163–165) structures of semantic domains and coupling them with various types of dictionary illustrations (without arguing why this classification is exhaustive or otherwise most appropriate).

Hupka's classification is more fine-grained than Ilson's, because it specifies in greater detail Ilson's tripartite distinction of *single*, *collective* and *composite* illustrations. However, Hupka's bipartition of *structural* and *functional* illustrations is very subtle and supposedly superfluous (functional illustration can be seen as a subcategory of structural illustrations). Likewise, it is unclear why *exemplary/encyclopedic* illustration should constitute a standalone category when their realizations draw on types from the other stipulated categories. Therefore, both Ilson's and Hupka's classification come with problematic issues.

Stein's (1991) approach differs from that of Ilson (1987) and Hupka (1989) in that it does not begin with an inventory of illustrations—a naturally open-ended set—but with the types of lexical items that can be visually represented. Stein (1991, p. 104) argues that starting from illustrations precludes any systematic refining and complementing of the typology, such that typologies produced by this approach "could only be expanded by randomly discovered illustrations that do not fit a particular typology set up on exactly such a randomly established basis." I agree with Stein (1991, p. 105) that the starting point should be the lexical items and that the function of an illustration is to complement information not contained in the definition of the lemma. In analyzing existing dictionaries, however, we must contend with numerous types of illustrations that are not used consistently and still require a classification.

[9] The English translation of the chapter about the classification (Hupka, 1989, pp. 235–248) was produced by Reinhard Hartmann and published a decade later in Hupka (2003, pp. 363–390). The classification was not updated in its typological depth (contrary to what is stated in the introduction).

The wide range of cases presented by Stein (1991) should have started, as she states, from the types of lexical items and not those of illustrations. However, the list she provides seems to be tailored on selected examples of illustrated lemmas that she could collect from LDOCE2 and OALD4 (i.e. the two sources she analyzed), whose illustrations she reproduces as typical examples. Adverbs, prepositions, and pronouns appear only in the last collective category she describes (bound with the use of captions to describe the illustration they denote). The inventory she presents is a valid starting point but needs to be refined (considering also the ontological nature of the concepts to represent) and updated with more modern ostensive aids too.

In order to extend present classifications and incorporate them into a systematic workflow for the addition of ostensive aids in dictionaries, three fundamental questions need to be answered and theoretically motivated. The first question is: what is the function of ostensive aids in lexicographic resources? The answer depends on dictionary type and target user, but this question can be also asked at a higher level, i.e. questioning *when* adding an ostensive aid is purposeful and relevant. If ostensive aids should complement the definition, then only ostensive aids that deliver additional information not contained in the definition, or is implicit, or would be too verbose, or has a pragmatic value (that is usually not part of a purely lexicographic definition) should be added to the lemmas. The second question is, how the senses of a lemma that should be candidates for visual representation can be systematically selected. The third question is, how the most appropriate ostensive aid can be selected among not only different typologies (single item/concept illustration, semantic field, partonomy, scene/frame, diagram) but also different representation techniques (static/animated, line drawing, picture/photograph, digital artwork, AI-generated; from which perspective, in what quality). The formalization of a workflow concentrating on these questions (and the connected criteria and conditions) is currently under development for the PAD (cf. Giacomini, 2025; DiMuccio-Failla, 2025).

5 Conclusion

Over seven decades of publication, one pattern is clear: the use of illustrations in learners' dictionaries has been ad hoc, sparsely documented, and inconsistent across print and online formats. A single partial exception—LDOCE2—does not offset the general nebulosity around the principles governing lemma selection. Taking into consideration general semiotic aspects regarding visual representations (e.g. Kress & van Leeuwen, 2021) should not be a peripheral concern of something secondary (i.e. illustration use) in lexicography, but become part of the compiling methodology of modern, reliable lexicographic resources that attract users at first, and keep them wanting to continue using them. In response to the gaps identified here, the PAD will implement a theoretically grounded workflow for ostensive aids: operational criteria for when a sense should include visual support; image-to-sense alignment at the level of syntactic-semantic patterns; guidance on the choice of visual representation (single-item illustration, collective illustration, composite illustration) and medium (line drawing, photo, animation, video, etc.), with a preference for visual aids that are instantly visible; consistent and systematic labeling; and cultural sensitivity for prototypes and regionally bound concepts. This project aims to make the use of ostensive aids principled and transparent.

References

DiMuccio-Failla, P. (2025). A theory for a usage-based cognitive lexicography. In L. Giacomini & V. Piunno (Eds.), *Patterns of meaning in lexicography and lexicology* (pp. 19–90). De Gruyter. https://doi.org/10.1515/9783111545943

DiMuccio-Failla, P. & Giacomini, L. (2017). Designing a learner's dictionary based on Sinclair's lexical units by means of corpus pattern analysis and the Sketch Engine. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek, & V. Baisa (Eds.), *Electronic lexicography in the 21st century: Proceedings of the eLex 2017 conference* (pp. 437–457). Lexical Computing CZ s.r.o.

DiMuccio-Failla, P., & Giacomini, L. (2022). A proposed microstructure for a new kind of active learner's dictionary. *Lexicographica*, *38*(1), 475–499. https://doi.org/10.1515/lex-2022-0016

Dziemianko, A. (2022). The usefulness of graphic illustrations in online dictionaries. *ReCALL*, *34*(2), 218–234. https://doi.org/10.1017/S0958344021000264

Dziemianko, A. (2024). Pictures in online dictionaries: Shall we see them? *GEMA Online® Journal of Language Studies*, *24*(2), 31–53. http://doi.org/10.17576/gema-2024-2402-03

Encyclopædia Britannica. Retrieved May 12, 2026, from https://www.britannica.com/

Fillmore, C. J. (1978). On the organization of semantic information in the lexicon. In Farkan, D., Jacobsen, W. M., & Todrys, K. W. (Eds.), *Papers from the parasession on the lexicon* (pp. 148–173). Chicago Linguistic Society.

Giacomini, L. (2024). Ontologies and knowledge representation in terminology: Present and future perspectives. *Applied Ontology*, *19*(1), 7–21. https://doi.org/10.3233/AO-230042

Giacomini, L. (2025). Introduction to the PhraseBase project. In L. Giacomini & V. Piunno (Eds.), *Patterns of meaning in lexicography and lexicology* (pp. 15–17). De Gruyter. https://doi.org/10.1515/9783111545943

Gumkowska, A. (2008). *The role of dictionary illustrations in the acquisition of concrete nouns by primary school learners and college students of English* [BA thesis, Collegium Balticum; Supervisor: Robert Lew].

Hupka, W. (1989). *Wort und Bild: Die Illustrationen in Wörterbüchern in Enzyklopädien*. De Gruyter. https://doi.org/10.1515/9783111347776

Hupka, W. (2003). How pictorial illustrations interact with verbal information in the dictionary entry: A case study (transl. by R. Hartmann). In R. Hartmann (Ed.), *Lexicography: Critical concepts* (Vol. 3) (pp. 363–390). Routledge. https://doi.org/10.7146/ln.v0i12.18675

Ilson, R. (1987). Illustrations in dictionaries. In A. Cowie (Ed.), *The dictionary and the language learner* (pp. 193–212). De Gruyter. https://doi.org/10.1515/9783111340500

Kemmer, K. (2014). Rezeption der Illustration, jedoch Vernachlässigung der Paraphrase? Ergebnisse einer Benutzerbefragung und Blickbewegungsstudie. In C. Müller-Spitzer (Ed.), *Using online dictionaries* (pp. 251–278). De Gruyter. https://doi.org/10.1515/9783110341287.251

Klosa, A. (2016). Illustrations in dictionaries: Encyclopaedic and cultural information in dictionaries. In P. Durkin (Ed.), *The Oxford handbook of lexicography* (pp. 515–531). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780199691630.013.37

Kress, G., & van Leeuwen, T. (2021). *Reading images: The grammar of visual design* (3rd edition, revised and updated). Routledge. (Original work published 1996). https://doi.org/10.4324/9781003099857

Labov, W. (2004). The boundaries of words and their meanings. In Aarts, B., Denison, D., Keizer, E., & Popova, G. (Eds.) *Fuzzy Grammar: A reader* (pp-67–89). Oxford. (Original work published 1973)

Lew, R. (2010). Multimodal lexicography: The representation of meaning in electronic dictionaries. *Lexikos*, *20*, 290–306. https://doi.org/10.5788/20-0-144

Lew, R., & Doroszewska, J. (2009). Electronic dictionary entries with animated pictures: Lookup preferences and word retention. *International Journal of Lexicography*, *22*(3), 239–257. https://doi.org/10.1093/ijl/ecp022

Lew, R., Kaźmierczak, R., Tomczak, E., & Leszkowicz, M. (2018). Competition of definition and pictorial illustration for dictionary users' attention: An eye-tracking study. *International Journal of Lexicography*, *31*(1), 53–77. https://doi.org/10.1093/ijl/ecx002

Nesi, H. (1998). Defining a shoehorn: The success of learners' dictionary entries for concrete nouns. In B. T. S. Atkins (Ed.), *Using dictionaries. Studies of dictionary use by language learners and translators* (pp. 159–178). Max Niemeyer Verlag. https://doi.org/10.1515/9783110929997.159

Ostermann, C. (2015). *Cognitive lexicography: A new approach to lexicography making use of Cognitive Semantics*. De Gruyter. https://doi.org/10.1515/9783110424164-004

Paivio, A. (2010). Dual coding theory and the mental lexicon. *The Mental Lexicon*, *5*(2), 205–230. https://doi.org/10.1075/ml.5.2.04pai

Perdih, A., Gabrovšek, D., & Ježovnik, J. (2025). Image-to-sense alignment using AI tools. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda, & A. Salgado (Eds.), *Intelligent lexicography: Proceedings of the eLex 2025 conference* (pp. 852–865). Lexical Computing CZ s.r.o.

Rebosio, L. (in press). The presentation of word meaning through ostensive aids in (electronic) dictionaries. *Studia Translatorica*.

Stein, G. (1991). Illustrations in dictionaries. *International Journal of Lexicography*, *4*(2), 99–127. https://doi.org/10.1093/ijl/4.2.99

Svensén, B. (2009). *A handbook of lexicography: The theory and practice of dictionary-making*. Cambridge University Press.

Wierzbicka, A. (1984). 'Apples' are not a 'kind of fruit'. *American Ethnologist*, *11*(2), 313–328. https://www.jstor.org/stable/643853

Wierzbicka, A. (2014). 'Prototypes save': on the uses and abuses of the notion of 'prototype' in linguistics and related fields. In S. L. Tsohatzidis (Ed.), *Meanings and prototypes: Studies in linguistic categorization* (pp. 347–367). Routledge. https://doi.org/10.4324/9781315857398

Dictionaries

CALD: *Cambridge advanced learner's dictionary*. Editions *1*: 2003; *2*: 2005; *3*: 2008; *4*: 2013. Cambridge University Press.

CALDo: CALD online (1999–2026). *Cambridge advanced learner's dictionary*. https://dictionary.cambridge.org/

CIDE: Procter, P. (Ed.) (1995). *Cambridge international dictionary of English*. Cambridge University Press.

COBUILD: *Collins Cobuild advanced learner's dictionary*. Editions *1*: 1987; *2*: 1995; *3*: 2001; *4*: 2003; *5*: 2006; *6*: 2009; *7*: 2012; *8*: 2014; *9*: 2018; *10*: 2023. HarperCollins Publishers.

COBUILDo: COBUILD online (2007–2026). *Collins Cobuild advanced learner's dictionary*. Collins. https://www.collinsdictionary.com/

Hornby, A. S., Gatenby, E. V., & Wakefield, H. (Eds.) (1948). *A learner's dictionary of current English*. Oxford University Press.

LDOCE: *Longman dictionary of contemporary English*. Editions *1*: 1978; *2*: 1987; *3*: 1995; *4*: 2003; *5*: 2009; *6*: 2014. Pearson Education Limited.

LDOCEo: LDOCE online (2008–2026). *Longman dictionary of contemporary English*. https://www.ldoceonline.com/

MEDAL: *Macmillan English dictionary for advanced learners*. Editions *1*: 2002; *2*: 2007. Macmillan Education.

MWALED: *Merriam-Webster's advanced learner's English dictionary*. Editions *1*: 2008; *2*: 2017. Merriam-Webster Incorporated.

MWo: MW online (1996–2026). *Merriam-Webster's.com Dictionary*. Merriam-Webster Incorporated. https://www.merriam-webster.com/

OALD: *Oxford advanced learner's dictionary*. Editions *1*: 1952; *2*: 1963; *3*: 1974; *4*: 1989; *5*: 1995; *6*: 2000; *7*: 2005; *8*: 2010; *9*: 2015; *10*: 2020. Oxford University Press.

OALDo: OALD online (2000–2026). *Oxford advanced learner dictionary*. https://www.oxfordlearnersdictionaries.com/definition/english/

Contact information

Laura Rebosio
University of Innsbruck
laura.rebosio@uibk.ac.at

Shigeru Yamada

PEDAGOGICAL VALUE AND INSIGHTS FROM COMPARING EFL DICTIONARIES

Evidence from *Apparently*, *Be on Speaking Terms*, and *Suffer (From)*

Abstract Certain aspects of word meaning and usage become visible only when dictionaries are compared directly. This paper examines three cases—*apparently*, *be on speaking terms*, and *suffer (from)*—to show how cross-dictionary analysis reveals patterns that remain hidden in single-dictionary consultation. By comparing definitions, sense divisions, and examples across major EFL dictionaries (CALD, COBUILD, LDOCE, MWALED, and OALD), the study demonstrates the pedagogical value of such comparison. The analysis of *apparently* shows notable variation: CALD lists three senses, whereas others merge or reduce them. Example sentences further reveal mismatches, such as an OALD example illustrating a sense not explicitly defined there but found in CALD. The expression *be on speaking terms* illustrates how usage tendencies emerge only through comparison. Dictionaries show that the phrase appears predominantly in the negative, and positive uses typically describe the restoration of communication after conflict. The verb *suffer* shows distinct transitivity and collocational patterns. Cross-dictionary analysis reveals contrasts among intransitive use, *suffer from*, and transitive constructions, clarifying semantic tendencies and demonstrating the value of comparative consultation for more precise lexical understanding. Overall, dictionary comparison reveals usage patterns invisible in single sources, offering clearer insights for interpretation and learning.

Keywords Dictionary comparison; dictionary instruction; EFL dictionary

1 Introduction

Despite the concerns about cost-effectiveness, a comparative use of multiple dictionaries brings clear benefits. Yamada (2025) proposes four informational dimensions: more information, better information, different information, and the same or similar information. More information is plainly beneficial. Better information is characterized by explicitness and, in some cases, exclusivity. Different information arises from different approaches—for example, in the treatment and description of headwords and set phrases, or in the division of senses. Even the same or similar information can contribute to confirming what one has learned and consolidating vocabulary knowledge. Only through side-by-side comparison of dictionaries do certain aspects of word meaning and usage—otherwise obscured when a single reference work is consulted—become visible. By systematically analysing entries in major online EFL dictionaries (CALD, COBUILD, LDOCE, MWALED, and OALD), this paper highlights the pedagogical value of comparative dictionary use and, building on Yamada (2025) and subsequent research, reports on three illustrative cases—*apparently*, *be on speaking terms*, and *suffer/suffer from*. An analysis of *apparently* presents a case in which the meaning conveyed by an example is not sufficiently explained in the dictionary in which it appears but is clarified in another. An examination of examples of *be on speaking terms* reveals recurring patterns in its usage. A comparison of the definitions and examples of *suffer* further clarifies the collocational and semantic differences between the transitive and intransitive uses of *suffer* and *suffer from*.

2 Comparison

2.1 *Apparently*[1]

As far as *apparently* is concerned, CALD uncharacteristically divides the adverb into these three senses (more than any other): 'uncertainty about reported information,' 'recognition of assumption-reality discrepancy'[2], and 'uncertainty about appearance-based inference.' COBUILD and LDOCE each provide two senses whereas MWALED and OALD give only one. However, OALD separates the definition into two parts by a semicolon. There will be no problem in deciding that the two senses (definitions) in COBUILD, LDOCE, and OALD correspond to CALD's first and third senses (see Table 1 for the correspondence of definitions across the dictionaries). MWALED's only definition is taken from the same publisher's unabridged dictionary: it seems apparent → used to describe something that appears to be true based on what is known[3]. The definition can be regarded as broad enough to cover CALD's first and third senses. This comparison suggests that CALD's second sense is likely to be of low frequency.

TABLE 1 Correspondence between dictionary definitions

	CALD	COBUILD	LDOCE	MWALED	OALD
'uncertainty about reported information'	[1]* used to say you have read or been told something although you are not certain it is true	1. You use **apparently** to indicate that the information you are giving is something that you have heard, but you are not certain that it is true.	1 used to say that you have heard that something is true, although you are not completely sure about it	— used to describe something that appears to be true based on what is known	according to what you have heard or read;
'recognition of assumption-reality discrepancy'	[2] used when the real situation is different from what you thought it was				
'uncertainty about appearance-based inference'	[3] used to say that something seems to be true, although it is not certain	2. You use **apparently** to refer to something that seems to be true, although you are not sure whether it is or not.	2 according to the way someone looks or a situation appears, although you cannot be sure	— used to describe something that appears to be true based on what is known	according to the way something appears

* The definitions in CALD are numbered in the order in which they appear.

1 This section builds on Yamada (2025).

2 This sense is included in CALD's first edition, *Cambridge International Dictionary of English* (1995): People also use apparently to mean 'in fact' when the real situation is different from what they thought it was.

3 https://www.merriam-webster.com/dictionary/apparently (accessed May 5, 2026)

Next, let us examine the examples. Focusing on the examples in MWALED and OALD – which do not distinguish senses by numbering – we will try to match each example (numbered in order of appearance) to the corresponding senses in CALD (See Table 2). The judgment is sometimes difficult because dictionary examples are extracted from their original contexts, but in principle, we assume that examples are intended to illustrate the definitions provided by each dictionary. An interesting observation can be made for each of the dictionaries.

TABLE 2 Correspondence between CALD's definitions and MWALED and OALD's examples

	Definitions in CALD	Examples in MWALED (MW) and OALD (O)
'uncertainty about reported information'	[1] used to say you have read or been told something although you are not certain it is true	[MW3] Apparently, he died of a heart attack. = He died of a heart attack, apparently. = He apparently died of a heart attack. [=it appears that a heart attack caused his death] [O1] Apparently they are getting divorced soon. [O4] The government was apparently unaware of the magnitude of the crisis. [O5] The security forces were apparently taken by surprise.
'recognition of assumption-reality discrepancy'	[2] used when the real situation is different from what you thought it was	[O2] I thought she had retired, but apparently she hasn't.
'uncertainty about appearance-based inference'	[3] used to say that something seems to be true, although it is not certain	[MW1] The window had apparently been forced open. [MW2] We were surprised when their apparently [=seemingly, ostensibly] happy marriage ended after only two years. [O3] He paused, apparently lost in thought.

Importantly, we are compelled to conclude that OALD's second example corresponds to CALD's second sense, namely the recognition of an assumption–reality discrepancy. Strictly speaking, this means that the use of *apparently* in that example is not actually covered by OALD's own definition. Without this knowledge, a user who relies solely on the dictionary will have only a shallow understanding. Caught between the two definitions, the user would attribute the resulting ambiguity to the lack of sufficient context.

Whereas all the other dictionaries distinguish between the senses of 'uncertainty about reported information' and 'uncertainty about appearance-based inference,' MWALED offers a broad definition that subsumes both. Nevertheless, the dictionary's third example is better accounted for by 'uncertainty about reported information,' while the first and second examples are more appropriately explained by 'uncertainty about appearance-based inference.'

In dictionary compilation, editors sometimes need to take a bold axe to the material; this is simply part of their craft. The point here, therefore, is not to argue about how much information a dictionary should or should not contain. What matters is that certain facts do not become visible when one consults only a single dictionary; they emerge only when several are compared side by side. A different dictionary may offer information that helps interpret a particular example or usage, and that foothold can lead to deeper understanding and, ultimately, to more effective language production and vocabulary development. These observations still have implications for reception, and especially for production and vocabulary expansion.

2.2 Be on speaking terms

Many learners probably remember *be on ... terms* as meaning "to be in a certain kind of relationship." From that, *be on speaking terms* is understood as 'to be in a relationship where you speak to someone.' Let us now consider what the EFL dictionaries say. OALD lists *be speaking (to somebody)* alongside this expression. The note following "especially" ("especially after an argument") is particularly informative. This initially led me to think that *be on speaking terms (with somebody)* might mean something like 'you had a quarrel, but you made up and are on speaking terms again.' However, when I look at the first example, it says that they "haven't spoken for years":

> **be on speaking terms** (**with somebody**) | **be speaking** (**to somebody**)
> to be willing to be polite or friendly towards somebody, especially after an argument
> *She's not been on speaking terms with her uncle for years.*
> *Are they speaking to each other again yet?* (OALD, **speak**, idioms)

Since "after an argument" appears as a note introduced by "especially", I wondered whether it really applies in all cases. But when I checked LDOCE, I found that the negative *not* is actually included in the headword item itself. This indicates that *be on speaking terms/be speaking* tends to be used in the negative:

> **be not speaking/not be on speaking terms**
> if two people are not speaking, they do not talk to each other, usually because they have argued
> He was not on speaking terms with his brother or sisters. (LDOCE, speak, sense 8)

Looking at other dictionaries, CALD provides two separate headwords: **be on speaking terms** (with no example) and **not be on speaking terms**. COBUILD treats *on speaking terms* under sense 17 of **speak** and sense 21 of **term**, while MWALED includes an example with *on speaking terms* under sense 5 of *term*[4]. All of the examples are in the negative:

> **not be on speaking terms** (**with**)
> *They had an argument last night and now they're not on speaking terms (with each other)*. (CALD)
> *For a long time her mother and her grandmother had hardly been on speaking terms*. (the same example; COBUILD, **speak** and **term**)
> *They are no longer on speaking terms*. [=they are no longer speaking to each other] (MWALED)

However, note that OALD's second example expresses the idea of 'becoming able to speak to someone again.' *Cambridge Academic Content Dictionary*, available on the same site as CALD, provides the following 'make up' example:

> *We had an argument, but we're back on speaking terms now* (**on speaking terms**, idiom).

In this way, it is only through comparison that we can see that *be on speaking terms/be speaking* is used mainly in the negative, and that when it is used affirmatively, it often carries the meaning of 'having repaired a relationship after an argument and being able to speak to each other again.'

[4] An example containing *on speaking terms* is also given under sense 3 of **speak**: They're not on speaking terms. [= they're not friendly and do not speak to each other.]

2.3 Intransitive and transitive *suffer*

Suffer is a troublesome verb both in meaning and in usage. The verb has transitive and intransitive uses, and in the intransitive use, the verb is used by itself and with various prepositions. Here we would like to focus on *suffer* with a human subject – specifically, the intransitive use with no following prepositional phrase and with *from* (seemingly the most frequent preposition) and the transitive use – in order to find out how the syntactic/collocational patterns correlate with semantic distinctions and what factors determine the choice between these competing constructions. While different dictionaries treat the verb in different ways, OALD provides a good starting point, separating the intransitive and transitive uses. Here are the definitions:

> **1** [intransitive] to be badly affected by a disease, pain, sad feelings, a lack of something, etc.
> **2** [transitive] **suffer something** to experience something unpleasant, such as injury, defeat or loss

Here, I will analyse the example sentences in the entries for *suffer* in EFL dictionaries – corresponding to OALD's senses 1 and 2 – by dividing them into three categories: the intransitive use (with no following prepositional phrase), *suffer from*, and the transitive use. For collocations, I will focus on those listed by multiple dictionaries.

- Intransitive use (with no following prepositional phrase)

  The intransitive use of *suffer* – with no prepositional phrase following it – expresses physical or mental pain, or suffering caused by circumstances. Most of the examples illustrate the meaning of 'physical pain' (see the last two paragraphs except the final one of this section for detailed analysis):

  > He died instantly and did not *suffer*. (MWALED, sense 1)

- *Suffer from*

  Collocations are given that include nouns referring to illnesses and symptoms such as the following:

  > *She's been **suffering from** (= been ill with) cancer for two years.* (CALD, (FEEL PAIN))
  > *Johnny suffers from (= is often ill with) asthma.* (CALD, (FEEL PAIN))
  > This patient is clearly *suffering from* shock. (MWALED, sense 1)

- Transitive verb

  Collocations are given with nouns such as *an injury*, *defeat*, *a heart attack*, *pain*, and *loss*:

  > *25 policemen suffered minor injuries during the riots.* (CALD, (EXPERIENCE))
  > *The Democrats suffered a crushing defeat*[5] *in the last election.* (CALD, (EXPERIENCE))
  > He *suffered* a heart attack and died instantly. (MWALED, sense 1)
  > *Within a few days she had become seriously ill, suffering great pain and discomfort. [VERB noun]* (COBUILD, sense 1)
  > *The company **suffered huge losses** in the last financial year.* (OALD, sense 2)

Based on these results, I would like to suggest the following revisions to OALD's definitions: supplementing the definition for sense 1 with *a medical condition* (and an example involving *shock*), and supplementing the definition for sense 2 with *a heart attack* and *pain*:

[5] LDOCE gives an example illustrating the meaning 'defeat in war' under sense 3: In June 1667, England **suffered** a humiliating **defeat** by the Dutch.

> **1** [intransitive] to be badly affected by a disease, [a medical condition,] pain, sad feelings, a lack of something, etc.
> **2** [transitive] **suffer something** to experience something unpleasant, such as [a heart attack, pain,] injury, defeat or loss

My revision is based on the fact that the definition for sense 1 functions as metalanguage, whereas the definition for sense 2 provides actual collocates introduced by "such as." Accordingly, the two instances of *pain* must be distinguished and treated as such.

As OALD does in the definition for sense 1, CALD (FEEL PAIN) and MWALED (sense 1) use *pain* as metalanguage. However, COBUILD and LDOCE appear to adopt a different approach. COBUILD characteristically frames its definition in a way that signals the *suffer* + *pain* collocation and illustrates it in one of the examples. LDOCE likewise seems to indicate *pain* as a collocate of *suffer*, both through the wording of its definition and through one of its examples:

> **1. verb B1+**
>
> If you **suffer** pain, you feel it in your body or in your mind.
> *Within a few days she had become seriously ill, suffering great pain and discomfort.* [***VERB noun***]
> *Can you assure me that my father is not suffering?* [*VERB*] *(COBUILD,* ***suffer****, sense 1)*
> **1** PAIN **[intransitive, transitive] to experience physical or mental pain**
> At least he died suddenly and didn't suffer.
> She's suffering a lot of pain. (LDOCE, **suffer**, sense 1)

Among all the examples in the entries for *suffer* in the EFL dictionaries surveyed, only the first example in COBUILD and the second example in LDOCE take *pain* as the direct object of *suffer*. The noun in the former refers to 'physical pain,' and the latter to either 'physical' or 'mental.' In the intransitive examples above, the type of 'pain' can either be inferred from the context ('physical' in LDOCE's example 1) or prove indeterminate (COBUILD's example 2). The other intransitive examples can be categorized in the following way:

– Physical pain:

> Before the surgery it was clear that she was really *suffering*. [=was in pain] (MWALED, sense 1)
> *She suffers in the winter when it's cold and her joints get stiff.* (CALD, (FEEL PAIN))

– Mental pain:

> I think he suffered a lot when his wife left him. (CALD, (FEEL PAIN))

To sum up, *pain* in the definition of *suffer* can function either as metalanguage or as language. The idea of 'suffer pain (physical or mental)' may be expressed either through the collocation *suffer* + *pain* or through the intransitive use of *suffer*. In the latter case, information about the nature of the pain can be supplied by, or inferred from, the context. As far as dictionary examples are concerned, nearly all of them refer to physical pain. These findings emerge only when multiple dictionaries are consulted.

As for the distinction between *suffer from* and the transitive *suffer*, the collocations are helpful, but it will be impossible for dictionaries to exhaust all collocations or for students to remember them. LDOCE's definitions (the underlined parts) serve as useful clues for understanding the tendencies in the choice between *suffer from* and *suffer* (transitive). *Suffer from* tends to refer to something long-lasting (sense 2), whereas transitive *suffer* conveys the nuance of something 'momentary' or 'a single event' (sense 3):

2 **DISEASE/MEDICAL CONDITION** **[intransitive]** to have a particular disease or medical condition, especially for a long time

3 **BAD EXPERIENCE/SITUATION** **[intransitive, transitive]** if someone suffers an unpleasant or difficult experience, or is in a difficult situation, it happens to them or they experience it

3 Conclusion

This paper has argued that dictionary comparison is not merely a way of accumulating more information, but a method for uncovering usage patterns, sense frequency, and semantic nuances that individual dictionaries cannot fully capture on their own. No dictionary is perfect, and each operates under conventional restraints of space, editorial policy, and descriptive approach. When dictionaries are consulted side by side, however, exposure to different descriptions and approaches can be eye-opening, bringing to light aspects of usage that may remain unnoticed in a single source. Such comparative use provides learners and teachers with a stronger interpretive foothold, leading to deeper lexical understanding and more effective language learning. It also has important implications for language education, as it helps cultivate an eye for critical dictionary use and encourages learners to engage with dictionaries more analytically, fostering a more informed and reflective use of reference tools.

Teachers, in particular, have a crucial role to play. They are expected to stay ahead of their students, and this includes taking on the responsibility of consulting multiple dictionaries and other reference resources when necessary. To cater to the needs of foreign students, EFL dictionaries have developed in the direction of "easier access and more lucid presentation" (McArthur, 1992, p. 594), to the extent that Rundell (1999, p. 48) could remark that "there is no question that MLDs are, in every department, very much easier to use than they were 20 years ago." However, there are still many points which the students need to be made aware of and for which they require explicit guidance. They cannot be expected to infer all the conventions and assumptions embedded in lexicographic practice, and without such guidance they may not be able to make full and truly effective use of EFL dictionaries.

In this respect, the teacher remains "one of the protagonists on the dictionary scene and [a] potential provider of guidance and instruction to users about how to select, access and exploit reference works appropriate to (language) learning" (Hartmann, 2000, p. 179). Taken together, these perspectives highlight the importance of equipping learners not only with dictionaries but also with the skills to use them critically and effectively. Comparative dictionary consultation—supported by informed pedagogical guidance—can play a central role in cultivating such skills.

References

Online dictionaries

CALD: *Cambridge Advanced Learner's Dictionary*. Retrieved May 24, 2026, from https://dictionary.cambridge.org/

COBUILD: *Collins Cobuild Advanced Learner's Dictionary*. Retrieved May 24, 2026, from https://www.collinsdictionary.com/

LDOCE: *Longman Dictionary of Contemporary English*. Retrieved May 24, 2026, from https://www.ldoceonline.com/

MWALED: *Merriam-Webster's Advanced Learner's English Dictionary*. Retrieved May 24, 2026, from https://www.britannica.com/dictionary/schematic

OALD: *Oxford Advanced Learner's Dictionary of Current English*. Retrieved May 24, 2026, from https://www.oxfordlearnersdictionaries.com/

Other works

Hartmann, R. R. K. (2001). *Teaching and Researching Lexicography*. Harlow: Pearson Education.

McArthur, T. (1992). *The Oxford Companion to the English Language*. Oxford: Oxford University Press.

Rundell, M. (1999). Dictionary Use in Production. *International Journal of Lexicography*, 12(1), 35-53.

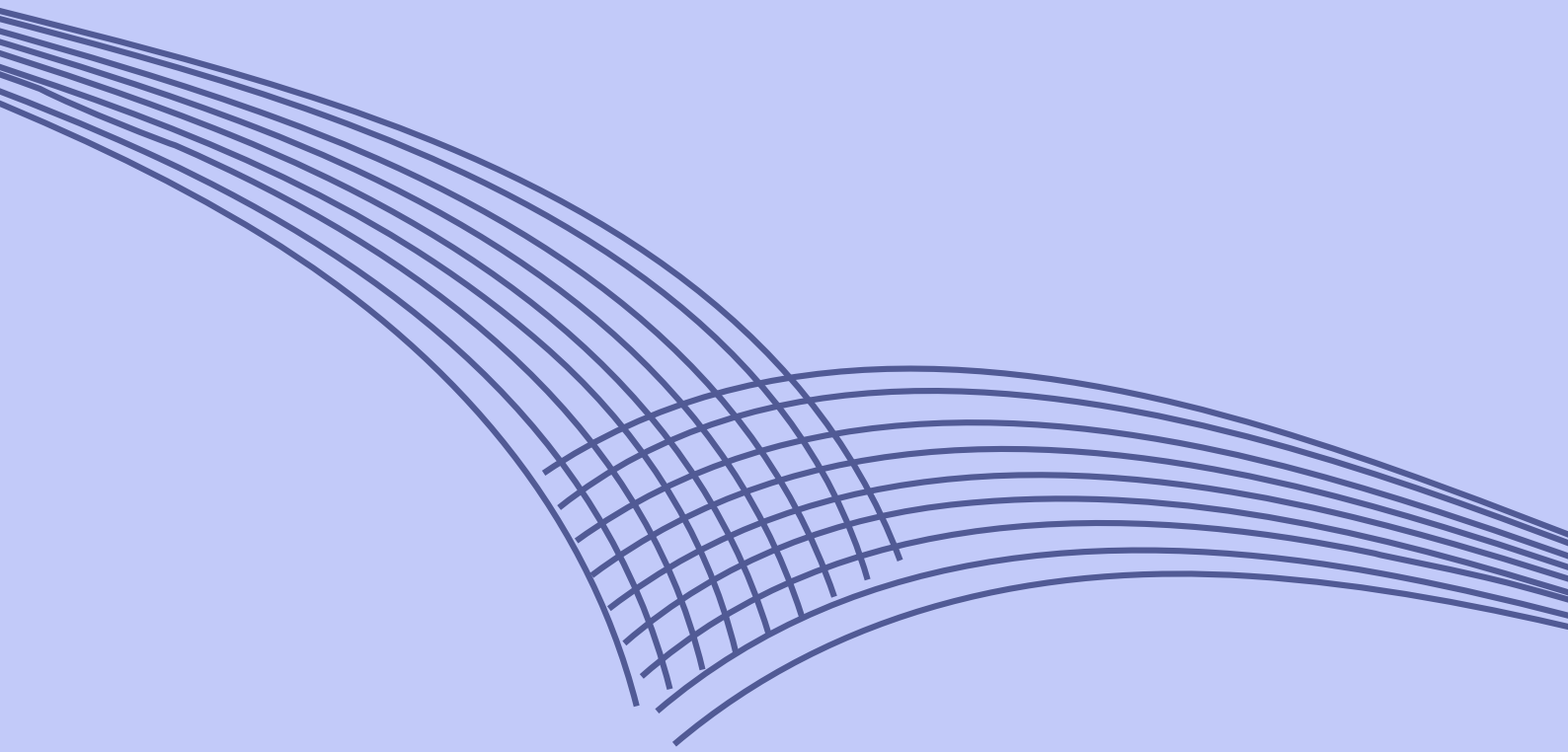
Yamada, S. (2025). *Benefits of consulting multiple EFL dictionaries*. The 25th Biennial Dictionary Society of North America Conference. Buffalo State University, New York. May 29-31, 2025.

Contact information

Shigeru Yamada
Waseda University
shyamada@waseda.jp

LEXICAL SEMANTICS, NEOLOGY & PHRASEOLOGY

EURALEX 2026



Elena Berthemet

DESIGNING AN IDIOMATIC DICTIONARY

Abstract This paper presents the principles and structure of *Colidioms*, an online idiomatic dictionary designed to provide a fine-grained description of idioms for their full understanding and production across typologically different language families, including French, Japanese, Korean, and Chinese. Grounded in cognitive and corpus linguistics and in recent work in phraseology and phraseography, the tool offers a unified model for both source and target languages, describing each idiom on at least three levels: formal structure, meaning(s), and use. The methodology combines corpus-based analysis, introspection, and linguistic experiments with native and non-native speakers. The paper focuses on the microstructure of entries, namely the canonical form and combinatorial properties, corpus-based definition(s), motivation, lexical, semantic, and syntactic co-occurrence constraints, register labels, variants, semantic links to related idioms (synonyms, equivalents, conversives, antonyms, false friends, hypernyms), notions, and pedagogical examples. The resource is both source- and target-oriented and enables multidirectional search for equivalents. We argue that *Colidioms* serves both practical aims (decoding and encoding idioms) and theoretical ones (contributing to the description of phraseological systems in typologically different languages and clarifying the theoretical status of idioms).

Keywords Equivalence; microstructure; motivation; production; register; semantic relations; understanding; variation

1 Main objective of the project

The paper presents the results of the research project of an online piece of software aimed at a detailed and accurate description of idioms. The main objective of the project is to build a system that allows one to compare idioms in typologically different languages (including Japanese, Korean and Chinese). It provides a unified model for both source and target languages and describes idioms at three levels of representation (formal structure, meaning(s), and use).

The idea of this tool initially grew out of my own frustrations as a learner of additional languages: often it is time-consuming to find information about idioms in dictionaries, and the offered information is fragmented or lacks coherence. The project has been considerably refined during my 20 years of experience as a French as a foreign language (FFL) teacher. Whilst designed with the foreign language learner and teacher in mind, it is also appropriate for translators, linguists and can serve the aims of linguistic theory. The main research question I address is the following: How can we improve theory by a systematic comparison of idioms across languages?

An exhaustive account of this project, regrettably, exceeds the scope of this paper, that is why, after explaining general features of the project (Section 1), we will briefly go into an overview of the microstructure (Section 2). The main point is illustrated by the French idiom written *in italics* followed by an English translation and a definition inside single quotation marks: *raconter des salades (à qqn.)* (lit. to tell salads) 'to tell well-argued lies to the other person, as if we were making a mixed salad with various ingredients'.

1.1 Comparison with similar approaches

Before we discuss the content and structure of entries, we need to take a brief look at other lexicographic and computational initiatives dealing with idioms and phraseology addressing similar design questions. A number of lexicographical projects exist: Colman & Tiberius, 2018; Gizatova, 2016; Moussallem et al., 2018; Villavicencio et al., 2004; and *Expressio* among others. Since these resources were developed with different objectives and for different user groups, their coverage and functionality are different too. However, in order to clarify the specific added value of *Colidioms*, a comparison with these initiatives is presented in Table 1 below.

As shown in Table 1, we addressed those issues, that, in our judgment, require special attention:

- Is the resource corpus-driven?
- What kinds of users are possible (humans or machines)?
- What type of dictionary is it (mono- or multilingual)?
- What connecting strategies are used:
    - Descriptors (the search is made through semantically motivated links such as ontologies, thesauri or markers);
    - Translations (equivalents in L2);
- What type of information is provided for each idiom:
    - Definitions;
    - Motivation (the explanation of *the inner form*, i.e. "the image fixed in its content plane as well as the motivation of its lexicalized meaning, which the speakers derive from the meaning of its constituents" (Dobrovol'skij, 2016, p. 23));
    - Variants (the focus is made on formal, i.e. lexical variation, whereas semantic variation is left aside);
    - Register;
    - Examples.

The ✓ symbol is used to indicate the presence of something, the 0 symbol to indicate either its absence or its uncertain and non-explicit status. In order to avoid misunderstandings, it is important to emphasize that this table is maximally simplified for two reasons: first, because of space restrictions, secondly, because of my lack of expertise.

As can be seen, idioms are described from various angles, therefore the breadth and depth of description differ among the above-mentioned resources (for example, some of them focus primarily on the translation, most of them contain information only for understanding and not production, register labels are generally disregarded). Furthermore, upon closer scrutiny it appears that none of the above-mentioned resources fully accounts for all of the points, which, in our judgment, are important for describing idioms. The desire to describe idioms for a full understanding and production in different language families dictated the design of *Colidioms*. As it is not possible to give a full description of our application, we will confine ourselves to listing some of its features:

- It is suitable for different language families;

TABLE 1 Comparison of lexicographic and computational initiatives dealing with idioms

Ressource	Corpus-driven	Users		Languages		Connecting strategies		Type of information				
		Humans	Machines	Monolingual	Multilingual	Descriptors	Translation	Definitions	Motivation	Variation	Register labels	Examples
Colman & Tiberius, 2018	✓	✓	0	✓	0	✓	0	✓	0	✓	0	✓
Gizatova, 2016	✓	✓	0	✓	0	✓	0	✓	✓	✓	0	✓
Moussallem et al., 2018	0	✓	✓	✓	✓	✓	✓	✓	✓	0	✓	0
Villavicencio et al., 2004	0	✓	0	✓	✓	0	✓	✓	0	✓	0	✓
Expressio	0	✓	✓	✓	✓	0	✓	✓	✓	✓	✓	✓
Colidioms	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓

- It is both source-oriented and target-oriented (idioms are uniformly described independently of the language);
- It enables multi-directional search of equivalents;
- It provides semantic relations to other related idioms;
- It contains information both for understanding (definition, inner form) and production (lexical, semantic and syntactic co-occurrence constraints, register labels, variants, examples).

Other features of *Colidioms* that are particularly useful for the description of idioms include:

- Possibility to filter and group idioms;
- Connection to Google Books API (in addition to the pedagogical examples, the user has additional illustrative material: link to Google Books);
- Possibility to
  - compare all the versions of the same article and to revert to the previous one;
  - improve and enlarge entries very quickly;
  - integrate other search engines in order to automatically explore online corpora;
- Finally, descriptions are tested on and improved by native and international students.

It is worth noting that work continues on defining the optimal content and structure of entries.

1.2 Theoretical framework

This project derives its theoretical and methodological sources from cognitive and corpus linguistics, it takes into account recent developments in phraseology (Granger & Meunier, 2008; Burger and al., 2007; Cowie, 1998; Moon, 1998) and phraseography (IJL, mono- and bilingual paper resources such as Baranov & Dobrovol'skij, 2007; Apresjan, 2008, 2010; Lubenski 1995). Description of idioms is based on corpus-based data (mainly on https://www.sketchengine.eu/), introspection and linguistic experiments. As the testing phase is being conducted in French, monolingual French corpora are used to extract information. However, it is planned to use other monolingual corpora for languages other than French, as well as parallel corpora to refine the description of idioms.

1.3 Methodology

The method that we employ consists of three main steps: corpus-based analysis, and two types of linguistic experiments (with native and non-native speakers). Since the project is needs-driven, we start with idioms that students encounter from the beginner level, for example *raconter des salades (à qqn.)* (lit. to tell salads) (see Section 2). Our aim is to explain why, where, and how to use it. Our first step is *Sketchengine*. To describe lexico-semantic and syntactic combinatorics as well as definitions, a sample of 200 concordance lines is checked and analysed manually for each idiom. This number of contexts seems to be sufficient and necessary for determining its canonical form, syntactic patterns and meaning(s). This task is performed manually, since human analysis allows to reliably identify and refine semantic and pragmatic descriptions of idioms. For example, depending on lexico-semantic contexts, the noun *salades*

(lit. salads) may be polysemous: '(lit.) a mixture of vegetables' or '(fig.) lies'. When it is preceded by the speech act verb *raconter* (lit. *tell*), it is an idiom. However, since the verb in the idiom is actually optional, as in the third pedagogical example *Arrête tes salades* (lit. Stop your salads), the idiomatic status of *salade* (lit. salad) could be put into question. Two of the most important criteria for inclusion of lexical units in our database are semantic opacity and dependency on the context in which they occur, we therefore decided to include one-word expressions in the database. The user will find this information in *Commentary* (see Section 2.5). Then, we retrieve three or four examples providing the best illustrations of idioms in prototypical situations (see Section 2.4).

In order to validate the obtained definitions, the second step is to present them to native speakers and to ask them to find the corresponding idiom. If the idiom is recognised by the majority of natives, the definition is presented to FFL learners, otherwise, it is revised. The third step consists in improving descriptions with the help of international students. In fact, FFL learners may shed light on the aspects that could have remained unnoticed by natives. For example, definitions can be satisfactory for natives, but insufficient for foreign learners. If international students fully understand and produce idioms, in other words encode meaning in a native-like way, we consider that the description and explanation of idioms encompass all of their particularities. Thus, the first corpus-based version of description undergoes a number of changes as a result of further testing on native and international students.

1.4 Technical aspects

On the technical side, it is a Web application that can be used with any web browser. The application is written in Java, using Vaadin, Spring, Apache Lucene, Tomcat, and Microsoft SQL Server.

The relational database comprises seven entities: Discussion, Expression, Language, Link, Notion, Revision, and Tag. Each Expression links to a Revision entity, which preserves the full change history and allows reverting to earlier versions. An Expression can have unlimited Links to any expression in the same or another language, each defined by "from"/"to" expressions and a "linktype". Notion and Tag entities categorize expressions. The relational model manages links, whereas Lucene handles search: all Expression fields are indexed including its content.

Because the software supports Unicode, adding a language only requires a new Language entry and the relevant Lucene "analyzer". Analyzers know how the text's language must be split and indexed, space and punctuation-based for most languages, but n-gramming (grouping characters in bi- or tri-grams) for Chinese, Japanese, and Korean (CJK). Finally, Colidioms integrates the Google Books API: clicking the Google button lists titles and authors containing the expression, and the Preview button opens a book at the relevant page where the author permits it. Colidioms contains 280 expressions, 500 revisions, 198 notions and 110 links.

Now that we have seen general features of *Colidioms*, we can turn our attention to its microstructure.

TABLE 2 Example of the entry *raconter des salades (à qqn.)*

Raconter des salades (à qqn.) 'dire des mensonges argumentés afin de convaincre l'interlocuteur, comme si on faisait une salade mélangée avec des ingrédients divers' (X raconte des salades à Y, X et Y sont des noms d'humains).
Notions: lie, rumour, exaggeration, argument, manipulation
Links:
(Fr.)
(1) *raconter des histoires*
(2) *dire <raconter> tout et n'importe quoi*
(3) *X [mentir] comme X respire*
(4) *tomber dans le panneau*
(5) *n'avoir rien vu venir*
(6) (*se faire*) *mener en bateau*
(7) (*se faire*) *avoir* (*en* (*toute*) *beauté*)
(8) (*se faire*, *se laisser*) *mener <tenir> par le bout du nez*
(9) (*se faire*) *rouler dans la farine*
(10) (*se faire*) *avoir comme un bleu*
(11) (*faire*) *tourner en rond*
(12) (*faire*) tourner en bourrique
(13) (*se faire*, se laisser) pigeonner
(Eng.)
(14) *spin a yarn*
(It.)
(15) *raccontare balle*
(16) *raccontare fandonie*
Pedagogical examples:
1. Il ne faut pas le croire, il raconte toujours des salades.
2. Maxime a raconté des salades pour ne pas se faire gronder par ses parents.
3. Arrête tes salades!

2 Microstructure

Since idioms have an extremely complicated structure involving many interacting phenomena (lexical, semantic, syntactical, and pragmatic), their formal description is no easy task. The development of the tool involves a number of theoretical issues (identification and capturing of the salient components of meaning, metalanguage, idioms' motivation, formal and semantic variations, to name some of the most substantial). In view of the limitations of space, we will discuss the microstructure only superficially. What the dictionary entry for the French idiom *raconter des salades (à qqn.)* (lit. to tell salads) would look like is demonstrated in Table 2. This example, considerably simplified, illustrates the content and structure of entries: canonical form, combinatorial properties, stylistic labels, definition(s), notions, links, and examples.

2.1 Entry

2.1.1 Headword

As can be seen, each idiom is identified by its canonical form: *raconter des salades (à qqn.)* (lit. to tell salads), where parentheses specify combinatorial properties. Next to the headword, corpora-based definition(s) can be found, their function is to explain the actual meaning: the traditional explanation in the first part ('to tell well-argued lies to the other person'), and the inner form in the second part ('as if we were making a mixed salad with various

ingredients'). The semantic structure of idioms is also provided: slots X and Y are filled with human-like nouns.

Since variation in idioms is more frequent than in words (Baranov, A. N., Dobrovol'skij D. O. (2014, p. 17)), we distinguish two types of variation: *formal* or in other words *structural*, and *functional* or *pragmatic* variation. The list of variants is not exhaustive.

2.1.2 Formal variation

Every phraseological unit has its own description, since idioms which are identical at all levels are extremely rare. Thus, variants sharing meaning, syntactic, combinatorial and stylistic properties but differing in images such as *raconter des salades (à qqn.)* (lit. to tell salads) *(à qqn.)* and *raconter des histoires* (lit. to tell tales) are accessed via a hyperlink in an autonomous entry.

Only largely interchangeable idioms (sharing inner form, meaning, syntactic, combinatorial and stylistic properties) such as *avoir quelqu'un dans le nez* < *le pif* > (lit. to have someone in the nose < the schnoz >) 'to hold a grudge against someone' and *passer un mauvais* < *sale* > *quart d'heure* (lit. to spend a bad < nasty > quarter of an hour) 'to find oneself temporarily in an unpleasant situation' appear in the same headword. As can be seen, angle brackets indicate mutually exclusive components.

2.1.3 Functional variation

Idioms with a specific area of use are followed by labels. Following the broad outlines of the framework set out in (Baranov & Dobrovol'skij, 2007), we distinguish four mutually complementary levels of *functional* or *pragmatic* variation:

(1) *temporal* or *diachronic* (the historical period during which the idiom was used, for example *acheter chat en poche* (lit. to buy a cat in a bag) (obs.) 'to buy something without first inspecting it');

(2) *territorial* or *diatopic* (the geographical area in which the idiom is used, for example in France, when talking about 'physical effort', one can say *huile de coude* (lit. elbow oil) whereas the equivalent Quebec idiom is *jus de bras* (lit. arm juice);

(3) *stylistic* or *diastratic* (the speakers' individual choice, for example *ça gaze* (low coll.) (lit. it gases) 'I'm fine');

(4) *discursive* or *diaphasic* (the type of discourse, thus, the *idiom faire le grand voyage* (high) (lit. set off on a long journey) is a euphemism for 'die' whereas *bouffer les pissenlits par la racine* (rude) (lit. eat dandelion by the roots) is a dysphemism).

2.2 Notions

For each idiom, the user is presented with *notions* (in English) that function as key words, an index to search in the database: for example, by clicking on the notion 'lie', the user will be provided with idioms associated with lie. They express rudimentary, elementary, immediate, intuitive knowledge about something and are written in simple, easy-to-understand unambiguous words: lie, rumour, exaggeration, argument, manipulation. Definitions can be viewed as

an analysis, whereas notions, as a synthesis of the meaning. There is no finite, pre-established inventory of notions. The inventory of notions is controlled rather than open. In order to maintain consistency of notions, users won't be able to freely create new notions. They will instead open a discussion on the addition of the notion and a debate will decide if existing notions already encompasses it or if it's useful to add it. Even if decomposing idioms' meaning is a risky task (notions are neither well defined (theoretically or methodologically) nor universally accepted), these semantic labels provide an onomasiological bridge that enables intra- and cross-lingual linking between lexical units.

2.3 Links

The third field is devoted to links – in other words, paradigmatic semantic relations to other intra- and cross-lingual related idioms, such as synonyms, equivalents, conversives, antonyms, false friends and hypernyms. In Table 2, one can see synonyms, equivalents, and conversives:

1) intralingual synonyms (idioms 1-3) and equivalents (idioms 6-13) like *raconter des histoires (à qqn.)* (lit. to tell tales) and *mener en bateau* (lit. 'to lead on a boat') 'mislead, trick someone';

2) conversives (idioms 4-5) such as *tomber dans le panneau* (lit. to fall into the board) 'believe the stories that one is told' and idioms that can function as conversives (idioms 6-13), for instance *se faire rouler dans la farine* (lit. 'to be rolled in the flour') 'be manipulated';

3) cross-lingual equivalents (idioms 14-16), e.g. (Eng.) *spin a yarn*, (It.) *raccontare balle* and *raccontare fandonie*.

The user is provided with an integral description by clicking on the hyperlink of each idiom, whether it is source idiom or a target idiom. For example, s/he is explicitly aware of the divergences with respect to their semantic subtleties and syntactic structures: synonyms are not simply listed, but receive a singular definition; actants are encoded for each idiom (*X raconte des salades à Y and Y est tombé dans le panneau*).

2.4 Pedagogical examples

In the last section, pedagogical examples are retrieved from corpora and are modified to suit pedagogical needs. They satisfy two requirements: first, they offer enlightening contexts and secondly, they help learners know how to use an idiom by showing its combinatorial properties. Thus, the above examples were modified to suit pedagogical needs (as far as possible, they were shaped to be accessible to learners at the B2 level: long sentences have been shortened, preferred syntactic and semantic environment have been incorporated, difficult vocabulary has been removed, only words that can be understood literally were kept).

2.5 Commentary

We plan to create a section with special notes for helping learners avoid common errors. At the present time, this field is not normalised in any way, and contains information about additional peculiarities which are not indicated in the previous fields such as:

1) Grammar comments (for example, *faire un malheur* with an indefinite article, *faire mouche* with zero article, *boire la tasse* with a definite article);

2) Prosodic features for idioms like *Ah, ça* and *Tu parles*, where a specific intonation contributes to the meaning;

3) Pragmatic properties with the description of the prototypical situation where the idiom is used;

4) Information about polylexical units that can be reduced to a node, i.e. a single word, as in *salades* (lit. salads) 'lies, exaggeration' used in a specific way.

3 Conclusion

Although a number of issues still need to be addressed, we hope that the presented project will be useful from the theoretical point of view and for practical purposes. From the applied point of view, its broad goal is to take into account all the linguistically and culturally relevant properties of idioms and to provide the basis for a full decoding and encoding. From the theoretical point of view, it can serve the objectives of linguistic theory: as an instrument for linguistic analysis, it contributes to describing phraseological systems in typologically different languages and to clarifying the theoretical status of idioms.

References

Apresjan, J.D. (2008). *Systematic Lexicography*. Oxford University Press.

Apresjan, J. D. (Ed.) (2010). *Prospekt aktivnogo slovarâ russkogo âzyka*. Âzyki slavânskih kul'tur.

Baranov, A. N., Dobrovol'skij, D. O. (Eds.) (2007). *Slovar'-tezaurus sovremennoj russkoj idiomatiki*. Mir entsiklopedij Avanta+.

Baranov, A. N. & Dobrovol'skij, D. O. (2014). *Osnovy frazeologii*. Flinta.

Burger, H., Dobrovol'skij, D., Kühn, P. & Norrick, N. R. (2007). *Phraseology. An International Handbook of Contemporary Research*. Walter de Gruyter.

Dobrovol'skij, D. (2016). "The notion of "inner form" and idiom semantics". In Viellard, S. (Ed.), *Études et travaux d'Eur'ORBEM, 2016, Proverbes et stéréotypes: forme, formes et contextes*, No. 1 (pp. 21-35). Eur'ORBEM.

Colman, L. & Tiberius, C. (2018). A good match: A Dutch collocation, idiom and pattern dictionary combined. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (Eds.), *Lexicography in Global Contexts: Proceedings of the XVIII EURALEX International Congress* (pp. 233-246). Ljubljana University Press.

Cowie, A. P. (Ed.) (1998). *Phraseology. Theory, Analysis, and Applications*. Clarendon Press.

Gizatova, G. (2016), A corpus-based approach to lexicography: Towards a Thesaurus of English idioms. In T. Margalitadze & G. Meladze (Eds.), *Lexicography and Linguistic Diversity: Proceedings of the 17th EURALEX International Congress* (pp. 348-354). Ivane Javakhishvili Tbilisi State University.

Granger, S. & Meunier, F. (2008). *Phraseology. An interdisciplinary perspective*. John Benjamins Publishing Company.

Lubenski, S. (1995). *Random House Russian-English dictionary of idioms*. Random House.

Moon, R. (1998). *Fixed expressions and idioms in English*. Clarendon press.

Moussallem, D., Sherif, M. A., Esteves, D., Zampieri, M. & Ngonga Ngomo, A. (2018), LIdioms: A multilingual linked idioms data set. In N. Calzolari et al. (Eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation LREC* (pp. 2468-2474).

Villavicencio, A., Baldwin, T., and Benjamin Waldron, B. (2004), A Multilingual Database of Idioms. In Proceedings of the Fourth International Conference on Language Resources and Evaluation LREC, (pp. 1127-1130). European Language Resources Association (ELRA).

Contact information

Elena Berthemet
Centre de Linguistique en Sorbonne
elenaberthemet@free.fr

Sophia Fiedler

THE INTERACTIONAL LEXICON OF SPOKEN GERMAN

A Functional, Audio-Based Lexicographic Resource

Abstract This paper introduces *The interactional lexicon*, an online lexicographic resource for lexical units in German talk-in-interaction. The resource's core contribution is a function-first entry model that complements form-based access. Grounded in Interactional Linguistics, all entries are derived from empirical, peer-reviewed studies and document units that are frequent or distinctively used in spoken German, with transcripts from naturally occurring conversations and access to audio/video recordings (FOLK-corpus, IDS Mannheim). To accommodate the organizational specificities of interaction, we propose a compact entry schema: headword and function; position in the turn; sequential position; a corpus-linked example and an analysis thereof. Further categories for prosody and activity framework are optional. A dual access structure supports lookup by form (alphabetical headwords) and by function (e.g., request, response, repair-initiation), based on a controlled functional taxonomy aligned with core concepts of Conversation Analysis. The online resource responds to a documentation gap for German. Lexical units in spoken language, particularly in social interaction, have not yet been brought together in a single resource that serves users with diverse linguistic backgrounds – from linguists to L2-German teachers and advanced students at universities worldwide. The resource will be published online with a restricted number of entries in early 2027, following a user study.

Keywords Online lexicography; interactional linguistics; spoken language; dual access structure

1 Introduction

The present paper introduces an innovative project at the interface of lexicography and interactional linguistics: The interactional lexicon. Within the project, we are building a corpus-based online resource of lexical units in German talk-in-interaction. To do so, we recur to published research in Conversation Analysis (CA) (Sacks et al., 1974) and Interactional Linguistics (IL) (Couper-Kuhlen & Selting, 2018). Both disciplines have produced a substantial body of knowledge on the organization of spontaneous spoken interaction in German, describing linguistic forms from both form-based and function-based perspectives. From a functional viewpoint, particularly IL investigates which linguistic resources speakers recurrently use to accomplish specific interactional tasks, for instance, recruitments, assessments, or account solicitations. A crucial finding in IL is that the same lexical unit can take over a range of functions in everyday talk, including interaction-organizational tasks like the management of turn-taking.

Despite the richness of IL research, there is currently no comprehensive information system that systematically documents findings on interactional features of lexical units. Information on those interactional functions is scattered across a large number of empirical studies. The project *Interaktionale Lexik* ('interactional lexicon') addresses this gap by developing an online dictionary that documents lexical units that are specific to, or distinctively used in, German talk-in-interaction. The aim of the project is to restructure existing findings into specialized entries and make them accessible in a centralized online resource together with audio/video excerpts from a publicly[1] accessible corpus of spoken German. Through new descriptive categories and infrastructural innovation, the project takes a first prototype – the LeGeDe Project (Meliss & Möhrs, 2019) – one step further. The new dictionary is unprecedented in both its

[1] Upon registration, users can access the entire corpus, including audio-/video-aligned transcripts.

format and presentation, containing information that many users would not typically expect to find in dictionary entries, as the description of spoken language necessitates forms of information that are absent from written language.

The paper introduces the project and its contribution to lexicography. I first describe the conceptual foundation of the online resource (Section 2). Secondly, I introduce the methodological and technical-infrastructural framework of the project (Section 3). Thirdly, Section 4 presents next steps and discusses the (potential) use of AI and finally, Section 5 concludes the article.

2 Conceptual Foundations

The main idea behind the project *Interaktionale Lexik* is to document lexical units or combinations thereof in a way that does justice to their interactional use. A crucial first step was to determine the features that set interaction apart from written language and from non-interactional spoken language like speeches or audio messages. A large body of research in CA and IL has documented that talk-in-interaction is organized in turns and sequences and that interlocutors follow certain conversational "rules" reflecting the pro-social character of interaction. These concern the general tendency of 'one speaker at a time', the regularities of turn-taking (Sacks et al., 1974) and preference structure (Bilmes, 1988; Pomerantz, 1984), as well as the opening and closing of sequences (Hopper, 1989; Sacks & Schegloff, 1973). Although a detailed explanation of these rules is beyond the scope of this paper, it is important to briefly introduce the concepts of turn-taking and sequence structure.

Turns are organized in so-called adjacency pairs. That means that a turn in first position, such as a question, is expected to be followed by a response in second position. These turns are called first and second pair part of an adjacency pair, for instance:

A: What time is it?	=> first pair part (first position)
B: Five o'clock	=> second pair part (second position)

The organization principle behind this order is that some actions in first position – for instance, questions, offers, or requests – make a response in second position relevant. In other words, interlocutors expect a response after a question. Together, first and second pair part form an adjacency pair. Often, this adjacency pair is followed by another turn in third position, which accomplishes sequence-closing work, for instance with *thank you* or *lovely*. These three parts form a conversational sequence. Interlocutors create this order through turn-taking and the opening and closing of conversational sequences or topics.

The temporal emergence of conversation as turns and sequences has one crucial impact on language use in interaction: the position of lexical items creates functional differences (Schegloff, 1996). For the project, it was therefore decided that all entries in the online resource should, whenever relevant, include information on turn and sequence position, making these features crucial descriptive categories for each entry. The following three excerpts, which all figure in the resource, illustrate the importance of these features. In Excerpts 1 to 3[2], speakers use *nein/nee* 'no' in different positions within a sequence and within a turn. These positional differences are associated with distinct functions.

[2] All excerpts are transcribed according to the GAT2 transcription conventions. These conventions include symbols indicating final intonation (e.g., a semi-colon for semi-falling final intonation or a dash for continuative final intonation), measured pauses (e.g., 1.2 seconds), overlap (in square brackets) and prosodic stress (in capital letters) or changes in volume. For a detailed list of all symbols, see Selting et al. (2009).

Excerpt 1: brötchen 'roll' (FOLK_E_00355_SE_01_T_02_c101)

01 MOT:	magst du noch n halbes brötchen?
	do you want another half of a roll?
02 DAU:	**nei:n;**
	no
03	ich hab keinen hunger mehr;
	I'm not hungry anymore

In Excerpt 1, a family is having breakfast and the mother (MOT) asks her daughter (DAU) whether she wants another half of a roll (= question in first position). The daughter responds in second position with a slightly lengthened *nein* 'no'. She thereby rejects the mother's offer. At line 03, the daughter explains her refusal: she is not hungry anymore. Giving a reason for a refusal is a recurrent practice and falls under the organization principle of 'preference' (Pillet-Shore, 2017). According to this principle, negative responses, refusals or disagreements are *dispreferred* (Pomerantz, 1984) and therefore typically accompanied by 'more work' like an account. This information is relevant for dispreferred *nein* 'no' in second position and is therefore included in the respective entry.

Excerpt 2: phonologische repräsentation 'phonological representation' (FOLK_E_00003_SE_01_T_01_DF_01_c66)

01 DEN:	°hh also bei\ sie nimmt an dass ähm (.) also es immer zwischenstufen gibt-
	well with she assumes that uhm well there are always intermediate levels
02	erst hat man phonologische (0.21) repräsentation-
	first there is phonological representation
03	dann geht_s zur (0.59) °h sem:antischen-
	then we go to semantic
04	**nee** erst zur syntaktischen und dann zur se[mantischen.]
	no first to syntactic and then to semantic (representation)
05 JAS:	[hmhm-

In Excerpt 2, a student takes an oral exam in phonology. In lines 01 to 03, the student, Denisa (DEN), responds to the examiner's question asking her to explain how bottom-up and top-down processes in articulation work. After introducing the interactive model more generally (line 01), Denisa makes a list of different representation levels (lines 02 to 04). When listing the semantic level (line 02), she initiates repair with *nee* 'no' and replaces 'semantic' by 'syntactic', thereby self-repairing the order of the list. *Nee* occurs mid-turn and does not respond to a prior turn by another speaker like in Excerpt 1. *Nee* allows Denisa to delete the prior item in the list ('semantic') and replace it with the correct solution, 'syntactic' (Lerner & Kitzinger, 2019). Her repaired list is validated by the examiner (JAS) in overlap (line 05). In these positions in the turn and in the sequence, *nein* functions as repair-initiator.

Excerpt 3: cupcakes (FOLK_E_00331_SE_01_T_03_DF_01_c444)

```
01 RAH:  wie lang kann ma denn diese cupcakes noch aufheben dass ma
         se essen kann;
         how long can one PART keep storing these cupcakes that one
         can eat them
02       (1.1)
03 CHA:  ähm ne woche-
         uhm a week
04       (0.6)
05 RAH:  <<aspirated>nei:n;>
                      no
06       (0.7)
07 CHA:  wieso  nich-
         why    not
```

Excerpt 3 demonstrates yet another use of *nein* 'no'. Charlene (CHA) is baking cupcakes while chatting with her mother, Rahel (RAH). In response to Rahel's question at line 01 and a pause (line 02), Charlene gives a time indication: *ne woche* 'a week' (line 03). After another pause (line 04), Rahel responds with an aspirated and slightly lengthened *nein*, which expresses her surprise or even doubt concerning the given time span (Imo, 2016). By doing so, Rahel makes Charelene's response (e.g., a concession and a repair or a reconfirmation) relevant in next position. Charlene, however, utters a counter question by asking *wieso nich* 'why not' (line 08), thereby averting an explanation or repair of her time indication. *Nein* occurs in first position after another adjacency pair as a standalone unit. *Nein* does not only challenge Charlene's response, it also assesses the time indication as unlikely.

The three excerpts illustrate that position is indeed a key feature for determining a lexical unit's function in talk-in-interaction. Though unprecedented as descriptive categories in lexicography, a unit's position in the turn and its position in the sequence are essential for accurately describing its use in talk-in-interaction.

A second conceptual decision was the integration of data excerpts from a corpus of naturally occurring talk. While empirical research results on functional characteristics are drawn from published empirical research articles, corpus examples illustrating each function have been selected from the *Forschungs- und Lehrkorpus Gesprochenes Deutsch* (FOLK) (see Section 3). As in dictionary entries that incorporate examples from literary works and newspapers, transcripts from audio or video recordings have been integrated alongside a detailed analysis of the function described in the entry (see Section 2 for an illustration thereof).[3] The inclusion of data excerpts serves both as evidence of the phenomenon and as an illustration of it.

The challenging conceptual innovation of the resource lies in its dual access structure. Users will be able to access single entries by lexical form (e.g., *nein* 'no' or *oh nein* 'oh no') and by function (e.g., *nein* to initiate repair). This means, entries are also organized according to interactional activity types (such as response, request, or repair), which serve as overarching functional categories. Individual functions of lexical units are systematically embedded in those categories. For the three excerpts with *nein* 'no' above, this means that Excerpt 1 figures in the entry "*nein* to negate a first pair part" within the function category "response," Excerpt 2 appears in the entry "*nein* to initiate repair" in the function category "repair," and Excerpt 3 is included in the entry "*nein* to reenact a question" which is classified under the functional

[3] After registration, users can also access the FOLK-corpus to compare transcripts and consult the original recording.

category "newsmarks." On the one hand, this function-based architecture reflects the idea that speakers use lexical units in talk-in-interaction to implement specific actions. On the other hand, it makes functional relationships between entries transparent and navigable. Users could, for instance, look up which lexical units speakers use to respond to an offer or how they mark a change of epistemic state in spoken German. Particularly for advanced L2-speakers of German, this feature may be beneficial. [4]

3 Workflow and Infrastructure

3.1 Candidate selection of headwords

Three factors influence the decision as to which functional categories and lexical units are included in the resource: recurrence of the functional category, frequency and distinctiveness of a lexical unit in talk-in-interaction, and the existence of empirical research on the unit.

To select the initial functional category for inclusion in the online resource, I drew on findings from CA research. According to the methodology of IL applied here, which is based on CA, a distinctive feature of talk-in-interaction is the "turn-taking machinery" (Sacks & Schegloff, 1973, p. 293), i.e., the back and forth of speakers taking turns in real time. Units in second position in adjacency pairs are extremely relevant in interaction because they demonstrate how a turn in first position has been understood. For instance, the turn "*I have to work tomorrow morning*" responding to "*Why are you leaving so early?*" demonstrates that the question was understood as 'seeking information.' The response "*What? Usually I'm the last one to leave*", however, demonstrates that the question in first position has been understood as a reproach. Because turns in second position reflect distinct features of talk-in-interaction, the functional category *response* is being built first. Other functional categories in the dictionary will be – among others – *repair, request, newsmarks,* and *account.*

Each functional category yields a large number of headword candidates. The category *response*, for instance, includes typical response particles like *nein* 'no', units related to epistemics like *ich weiß nicht* 'I don't know' and *ach so* 'oh', but also confirmation particles like *(das) stimmt* '(that's) right'. The second factor to choose candidates among this group of lexical units is their higher frequency in FOLK compared to a corpus of written language and/or their specific use in interaction compared to written language. Therefore frequency classes between DeReKo (Kupietz et al., 2023) and FOLK (Reineke et al., 2023) have been compared. Only items with a frequency difference higher than 3 were further considered.

The third factor is the existence of research on the lexical unit. A published empirical study on a lexical unit in the field of CA or IL generally treats its distinct use in talk-in-interaction. The candidate list containing frequency comparisons is therefore filtered according to whether relevant literature is available or not, and completed by a bibliography of relevant literature.

3.2 Data base of entries

The research results that underlie each dictionary entry are published, peer-reviewed, empirical studies from the field of CA and IL that typically investigate lexical units or collocations in

[4] In research on German as a foreign language, there is now frequent criticism that language teaching places too much emphasis on written norms, and several articles have argued for the significance of medial diversity in L2-teaching contexts. It is to be expected that reference resources like this online resource will play a more important role in the future due to changes in L2 didactics (see, for instance, Wittner & Ender, 2025).

everyday German interactions. In that sense, the online resource can also be seen as a repository of research on lexical units in German talk-in-interaction.

These qualitative empirical studies typically comprise information on the position of the unit within the turn and the sequence, on the function related to those positions, and on additional features like prosody or activity type. Transcribed data excerpts used in these publications follow established conventions (for instance, Selting et al., 2009 for German) which include pauses, prosodic information, and overlaps. Because the normalization[5] of such transcripts is time-consuming, relative token frequencies are usually not calculated.

Only some of those empirical studies are based on recordings from FOLK[6]. To ensure internal consistency, data excerpts within the online resource are not quoted from the underlying published study. Instead, corresponding excerpts are manually selected within FOLK and re-formatted for the resource. That means, transcripts from FOLK are segmented according to intonation units and completed by prosodic information.

As of May 2026, within the version 2.26 of the DGD, the FOLK corpus comprises a total of 515 hours of audio recordings and 324 hours of video recordings recorded between 2003 and 2025. All data are recordings of naturally occurring talk, in other words, all interactions would have taken place in a comparable way without the presence of a researcher. The corpus includes non-institutional and institutional settings: interactions between family and friends like having coffee or cooking together (= non-institutional) as well as classroom-interactions or TV debates (= institutional).

3.3 Technical-infrastructural framework

All entries use XML as a device- and system-independent data format for storing and processing texts. The online resource will be integrated in an established framework at the IDS Mannheim that ensures long-term accessibility. Furthermore, all data will be incorporated into a long-term repository to ensure data preservation and integration into research data infrastructures like *NFDI-Text+*.

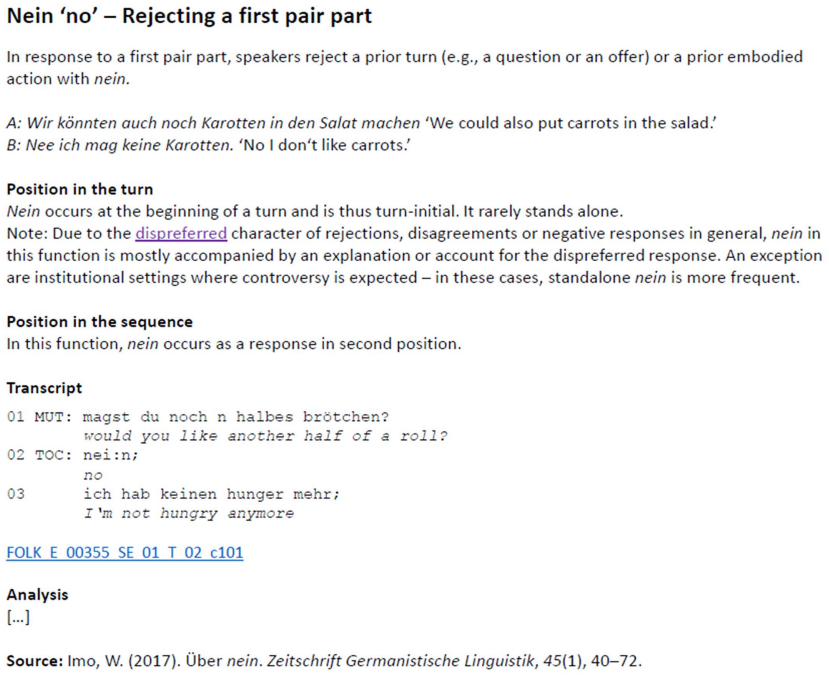
All transcripts have a backlink to the DGD, where transcripts, audio and video files of excerpts from FOLK are stored. This backlink is not only for technical reasons; it also enables comparison of transcripts and further research within the FOLK-corpus.

3.3.1 Entry structure

The entry structure and description style balances precise interactional linguistic terminology and understandability for a broader linguistic public outside IL. For the untrained eye, it may pose particular challenges, especially with regard to terminology; however, issues of terminological understanding are addressed in an accompanying glossary. This was implemented to enable both non-experts and L2 teachers/learners to make use of this resource as well.

[5] Normalization means to delete all transcription symbols and change morpho-phonological reductions like *dunno* (= I do not know) into units recognizable for automated analysis of collocations or statistical analysis.

[6] The *Forschungs- und Lehrkorpus gesprochenes Deutsch* 'Research and Teaching corpus of spoken German' (FOLK) (Reineke et al., 2023) is one of numerous corpora within the *Datenbank für Gesprochenes Deutsch* 'Data base for spoken German' (DGD) hosted by the *Archiv für Gesprochenes Deutsch* 'Archive for spoken German' of the IDS Mannheim.



Nein 'no' – Rejecting a first pair part

In response to a first pair part, speakers reject a prior turn (e.g., a question or an offer) or a prior embodied action with *nein*.

A: Wir könnten auch noch Karotten in den Salat machen 'We could also put carrots in the salad.'
B: Nee ich mag keine Karotten. 'No I don't like carrots.'

Position in the turn
Nein occurs at the beginning of a turn and is thus turn-initial. It rarely stands alone.
Note: Due to the dispreferred character of rejections, disagreements or negative responses in general, *nein* in this function is mostly accompanied by an explanation or account for the dispreferred response. An exception are institutional settings where controversy is expected – in these cases, standalone *nein* is more frequent.

Position in the sequence
In this function, *nein* occurs as a response in second position.

Transcript

```
01 MUT: magst du noch n halbes brötchen?
        would you like another half of a roll?
02 TOC: nei:n;
        no
03      ich hab keinen hunger mehr;
        I'm not hungry anymore
```

FOLK_E_00355_SE_01_T_02_c101

Analysis
[...]

Source: Imo, W. (2017). Über *nein*. *Zeitschrift Germanistische Linguistik*, *45*(1), 40–72.

FIGURE 1 Entry structure of function-based entries[7]

Figure 1 illustrates the entry structure and the extent of information (here translated) for a function-based entry (*nein* 'no'). The exact layout structure as it will be published in the online resource later is yet to be determined.

After a heading – composed of the target unit and its specific function – there is a box with IL-specific basic terms that readers need to be familiar with in order to understand the article (e.g., turn, sequence, . . .). This feature was included to reduce internal links to the glossary, ensuring fluent readability while accommodating a heterogeneous readership that may or may not be familiar with basic IL terminology. Less recurrent concepts, for instance *dispreferred* in Figure 1, contain a link to a glossary entry with an explanation and a corpus excerpt for illustration. A brief functional description is followed that may be accompanied by a short, possibly invented, example for illustration.

The next two descriptive categories are *Position in the turn* and *Position in the sequence*. Both categories are explained in the glossary because they are specific to the applied methodology of IL. As it can be seen under *Position in the sequence* in Figure 1, descriptions under these two categories are, with some exceptions (see Figure 1), rather short.

The core of each entry is the transcribed data excerpt and an analysis thereof. Excerpts for each function come from targeted searches in FOLK. Beneath each transcript, the source of the excerpt is indicated. This indication is backlinked to the original transcript and audio or video file stored on the DGD-platform. The analysis of the excerpt focuses on the target lexical unit and the related function. In line with the IL approach, the turns that lead up to the target unit and the turns that follow it are essential part of the analysis.

[7] All entries are in German. This example is a translation.

An optional category with necessary prosodic information, specific activity types (e.g., German *genau* 'exactly' in Power Point Presentations to mark a topic shift, cf. Auer, 2020) or diatopic variety may be included under a separate heading just above the transcript. The entry closes with the reference to the underlying empirical study.

3.3.2 Navigation

Through a dual access structure, users are able to access the online resource either through function or through a lexical form. Figure 2 illustrates the page flow resulting from this dual access structure. When clicking on *Access through function* users will see an overview of different functional categories, for instance, *assessment, newsmark,* or *response.*[8] Under each functional category, a list of lexical units or collocations that exhibit this function is provided. Clicking on one of these units opens a corresponding functional entry (see Section 3.3.1 for the detailed entry structure). By clicking on the functional category *response*, users can access a general explanation of the function. Terms that are not part of recurrent IL-terminology figure as internal links to the glossary (e.g., *positioning* in Figure 2).

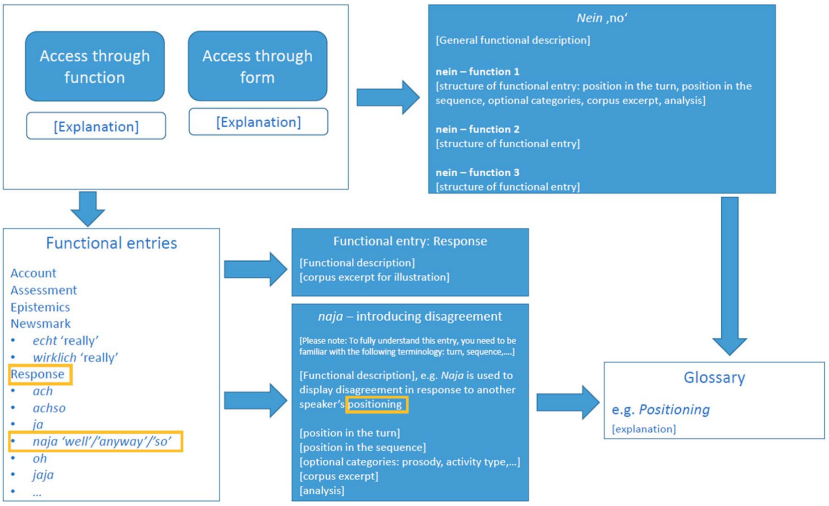


FIGURE 2 Conceptual page flow of the online resource

When users access the resource by form (rather than by function), they are presented with an alphabetical list of all units, as is traditional in dictionaries. In this view, all functions of the same lexical unit are listed consecutively.[9] The structure and content of these functions correspond to the functional entries described above.

[8] The online resource will be in German; all categories and descriptors have been translated into English here.

[9] Due to readability, functions are numbered in Figure 2. This will not be the case in actual form-based entries to avoid an association of the number with frequency or importance of certain form-function pairings.



FIGURE 3 Draft page design for functional entry access

Accessing a functional entry leads to the view shown in Figure 3.

In the left bar, users see the structure of their current choice of access (function or lexical form) and where the displayed entry figures within this structure. The entry itself figures in the middle of the page. In the right bar, additional information is provided like the glossary entry users clicked on, related functions in the same functional category or other functions of the same lexical unit.

Altogether, the user interface will be designed to make functional and formal relations between lexical units and functions visible. At the same time, information should be easy to find and clearly structured. Additional steps will be taken to enhance the design and information flow, some of which are outlined in the following section.

4 Next Steps and the Use of AI

In preparation for the planned online publication of an initial set of entries in early 2027, the page design needs to be finalized,[10] readability and searchability need further improvement, and a user study will be conducted. Integrating results of former studies on dictionary use[11], the aim is to design a resource with a user-friendly information structure and a clear navigation system. The searchability for forms, functions, and headwords is a major tool that will be implemented through the addition of tags. When the technical-infrastructural setup is finalized and a demo version of the online resource is generated, a larger user study will be carried out towards the end of 2026. In this study, different target groups of the resource are examined: IL linguists, linguists outside the field of IL/CA, students, advanced L2 German students at university level, and L2 German teachers at university level. Questions in this study will concern the understandability of entries, data excerpts, and analyses, the access of audio- and video recordings through the DGD platform, the ease of navigation and searchability of entries, and the overall design of the resource.

10 I thank Frank Michaelis for the technical implementation.

11 For an overview of research into dictionary usage of online dictionaries, see https://www.ids-mannheim.de/lexik/abgeschlosseneprojekte/bzvelexiko/benutzungsforschung/usingdictionaries/ (Klosa et al., 2014; Müller-Spitzer, 2014)

Given recent advances in large language models and AI, the resource will also be optimized so that AI systems can recognize it as a reliable reference for lexical units in spoken language. In order to apply principles of Generative Engine Optimization, the aim is to revise entries accordingly. Three points are particularly important to consider for revision: providing structured, in-depth content, tailoring entries to prospective users by enhancing readability, diversifying keywords, and integrating featured snippets (Xponent21, 2025; see also Rejón-Guardia, Molinillo & Anaya-Sánchez, 2026). The planned user study will be one crucial step to optimize readability, structure and keyword diversity. As for featured snippets, a special focus will lie on the optimization of heading formulations and the reformulation of descriptive paragraphs so that they "integrate the question into the answer itself" (Xponent21, 2025). Regular entry updates and the continuous growth of the online resource will signal relevance to AI systems, which will potentially increase citations. In general, all these planned steps need to be implemented with caution due to concerns about data protection – particularly regarding transcripts – and intellectual property rights. Search results may be misattributed because LLMs have not been sufficiently trained on this type of resource. An integration of AI into the workflow of writing entries is also impossible due to protection of corpus data and publishing rights of empirical studies.

5 Conclusion

The present paper introduced the project *Interaktionale Lexik* 'Interactional Lexicon' in its different conceptual, methodological, and technical-infrastructural features. The innovative online resource accommodates typical features of talk-in-interaction: a function-oriented use of talk, its organization in turns and sequences, and prosodic features. Transcripts of corpus excerpts of naturally occurring talk provide the possibility of accessing data directly and opens up opportunities for further individual research.

The resource will be valuable not only for linguistically trained users. Advanced L2 learners of German and their teachers may appreciate it for providing reliable information on units typical of spoken interaction. Reliable, accessible, and condensed information on spoken language remains scarce – a gap this project hopes to address. While it cannot yet accommodate learner-specific needs, it aims to offer a first accessible resource for this audience.

Going forward, the number of entries in the online resource will be continuously extended and searchability and entry variety will be increased, thus creating a resource that will be widely used by professional linguists as well as students worldwide.

References

Auer, P. (2020). Genau! Der auto-reflexive Dialog als Motor der Entwicklung von Diskursmarkern. In *Fachinformationsdienst Linguistik, Verfestigungen in der Interaktion* (pp. 263–294). De Gruyter. https://doi.org/10.1515/9783110637502-011

Bilmes, J. (1988). The concept of preference in conversation analysis. *Language in Society*, 17(2), 161–181. https://doi.org/10.1017/S0047404500012744

Couper-Kuhlen, E. & Selting, M. (2018). *Interactional Linguistics*. Cambridge University Press.

Hopper, R. (1989). Speech in telephone openings: Emergent interaction v. routines. *Western Journal of Communication,* 53(2), 240–252. https://doi.org/10.1080/10570318909374299

Imo, W. (2016). Im Zweifel für den Zweifel: Praktiken des Zweifelns. In A. Deppermann, H. Feilke & A. Linke (Eds.), *Sprachliche und kommunikative Praktiken* (pp. 153–176). De Gruyter. https://doi.org/10.1515/9783110451542-007

Klosa, A., Koplenig, A. & Töpel, A. (2014). Benutzerwünsche und -meinungen zu dem monolingualen deutschen Onlinewörterbuch elexiko. In C. Müller-Spitzer (Ed.), *Using Online Dictionaries* (pp. 281–384). De Gruyter.

Kupietz, M., Lüngen, H. & Witt, A. (2023). DeReKo im Kontext deutschsprachiger Gegenwartskorpora: Perspektiven – Ziele – Visionen. In M. Beißwenger, E. Gredel, L. Lemnitzer & R. Schneider (Eds.), *Korpusgestützte Sprachanalyse. Grundlagen, Anwendungen und Analysen* (pp. 61–77). Narr.

Lerner, G. H. & Kitzinger. C. (2019). Well-prefacing in the organization of self-initiated repair. *Research on Language and Social Interaction*, 52(1), 1–19. https://doi.org/10.1080/08351813.2019.1572376

Müller-Spitzer, C. (2014). *Using Online Dictionaries.* De Gruyter.

Pillet-Shore, D. (2017). Preference organization. In J. Nussbaum (Eds.), *Oxford Research Encyclopedia of Communication*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190228613.013.132

Pomerantz, A. (1984). Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shapes. In J. M. Atkinson & J. Heritage (Eds.), *Structures of social action: Studies in conversation analysis* (pp. 57–101). Cambridge University Press.

Reineke, S., Deppermann, A. & Schmidt, T. (2023). Das Forschungs- und Lehrkorpus für gesprochenes Deutsch (FOLK). Zum Nutzen eines großen annotierten Korpus gesprochener Sprache für interaktionslinguistische Fragestellungen. In A. Deppermann, M. Kupiez & C. Fandrych (Eds.), *Korpora in der germanistischen Sprachwissenschaft. Mündlich, schriftlich, multimedial* (pp. 71–102). Leibniz-Institut für Deutsche Sprache (IDS). https://doi.org/10.1515/9783111085708-005

Rejón-Guardia, F., Molinillo, S. & Anaya-Sánchez, R. (2026). Generative Engine Optimization. In R. Ladhari, (Eds), *Encyclopedia of Artificial Intelligence in Marketing.* Springer. https://doi.org/10.1007/978-3-031-75316-9_68-1

Sacks, H. & Schegloff, E. A. (1973). Opening up closings. *Semiotica*, 8(4), 289–327. https://doi.org/10.1515/semi.1973.8.4.289

Sacks, H., Schegloff, E. A. & Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4), 696–735.

Schegloff, E. A. (1996). Turn organization: One intersection of grammar and interaction. In E. A. Schegloff, E. Ochs & S. Thompson (Eds.), *Interaction and grammar* (pp. 52–133). Cambridge University Press. https://doi.org/10.1017/CBO9780511620874.002

Selting, M., Auer, P., Barth-Weingarten, D., Bergmann, J., Bergmann, P., Birkner, K., Couper-Kuhlen, E., Deppermann, A., Gilles, P., Günthner, S., Hartung, M., Kern, F., Mertzlufft, C., Meyer, C., Morek, M., Oberzaucher, F., Peters, J., Quasthoff, U., Schütte, W., Stukenbrock, A. & Uhmann, S. (2009). Gesprächsanalytisches Transkriptionssystem 2 (GAT 2). *Gesprächsforschung - Online-Zeitschrift Zur Verbalen Interaktion*, 10, 353–402.

Wittner, J. & Ender, A. (2025). VIELLEICHT – DU HAST RECHT. Zu sprachlichem Gebrauch und Akzeptanz von Deklarativsätzen mit V2- vs. V3-Stellung in Erst- und Zweitsprache Deutsch. *Korpora Deutsch als Fremdsprache*, 5(2), 138–163.

Websites

LeGeDe: *Lexikon Gesprochenes Deutsch*. Retrieved May 18, 2026, from https://www.owid.de/legede/

DeReKo: *Deutsches Referenzkorpus*. Retrieved May 18, 2026, from https://cosmas2.ids-mannheim.de/cosmas2-web/

Xponent21. (2025, March 11). *How to optimize content to rank in AI search results*. Xponent21 Insights. Retrieved July 13, 2026, from https://xponent21.com/insights/optimize-content-rank-in-ai-search-results/

Contact information

Sophia Fiedler
Leibniz-Institut für Deutsche Sprache, Mannheim
sophia.fiedler@ids-mannheim.de

Acknowledgements

I thank Petra Storjohann for her careful reading of a first draft of this paper.

Ivana Filipović Petrović & Slobodan Beliga

MODELLING PHRASEOLOGICAL SEMANTIC STRUCTURES IN DMLEX

Aligning a Croatian Lexonomy-Based Idiom Dictionary with a Thematic Index

Abstract Phraseological resources contain complex semantic structures that are often difficult to represent in interoperable lexical models. Although digital phraseological dictionaries increasingly use hyperlinks, thematic indices, and semantic groupings to improve navigation, many of these relations remain implicit and are primarily intended for human interpretation. This study explores the expressive capacity of the DMLex model for representing conceptual and semantic relations in a Croatian phraseological resource originally developed in Lexonomy. The resource consists of the *Online Dictionary of Croatian Idioms* and a manually curated thematic index grouping phraseological units into semantic domains and subdomains. The study adopts a scenario-based modelling approach focused on recurring phraseological structures observed in the resource. Three modelling scenarios are analysed: (1) variant and constructional relation networks, (2) polysemy and sense-level thematic classification, and (3) multiple thematic affiliation. The analysis demonstrates how lexical variation, constructional relations, synonymy, and thematic organisation can be represented through explicitly encoded semantic relations within DMLex. The results show that DMLex can accommodate much of the semantic complexity of phraseological resources while also revealing challenges related to overlapping classifications and conceptual hierarchies. The proposed approach contributes to the interoperable modelling of phraseological data for lexicographic and NLP-oriented applications.

Keywords Phraseological units; semantic networks; relational modelling; digital lexicography; DMLex

1 Introduction

Phraseological units (PUs) have long been recognised as a particularly demanding area for both lexicography and natural language processing (NLP) because they combine formal stability with semantic opacity (Sag et al., 2002; Baldwin & Kim, 2010; Constant et al., 2017). Unlike non-phraseological lexical units, PUs function simultaneously on several interconnected levels. Their formal structure consists of lexical items that usually retain literal meanings outside the phraseological construction, while the phraseological meaning emerges only at the level of the expression as a whole. This semantic layering complicates both interpretation and representation, especially in computational environments.

In addition to semantic non-compositionality, phraseological systems are characterised by dense networks of conceptual and semantic relations. Many PUs express similar meanings despite having entirely different lexical structures. For example, the Croatian PUs *kad na vrbi rodi grožđe* ('when grapes grow on a willow tree') and *na sveto Nigdarjevo* ('on Saint Never's Day') both express the meaning 'never', despite having no lexical overlap. In multilingual contexts, semantically equivalent expressions may also differ completely in formal structure and imagery, as in Croatian *kad na vrbi rodi grožđe*, Italian *alle calende greche*, and English *when pigs fly*, all conveying the meaning 'never'. These characteristics pose considerable challenges for computational phraseology and the development of multilingual phraseological resources, since semantically equivalent units often cannot be identified through formal similarity alone.

Recent research has shown that many of these problems persist in transformer-based and large language models despite major advances in language technology (Miletić et al., 2024; Filipović Petrović & Beliga, 2025). These limitations suggest that advances in language technology alone are not sufficient to resolve the challenges posed by phraseological data. Their representation requires modelling approaches capable of capturing not only formal structure, but also conceptual organisation, semantic relations, and sense-level distinctions.

The present study investigates whether DMLex, a standardised model for the representation of digital dictionary data, can adequately represent the semantic organisation of a phraseological resource whose conceptual structure was originally designed primarily for human navigation. The study is based on the *Online Dictionary of Croatian Idioms*, an open-access phraseological resource developed in Lexonomy and enriched with a manually constructed thematic index.

The paper focuses on several recurring modelling scenarios identified within the resource, including variant and conceptual relation networks, polysemy and sense-level thematic classification, and multiple thematic affiliations. The aim is to explore how semantic structures originally encoded through hyperlinks, thematic lists, and interface-level organisation can be formalised within an interoperable lexical modelling framework.

This paper is organised as follows. Section 2 presents the theoretical and lexicographic background of the study. Section 3 describes the source resource and the methodological approach. Section 4 analyses several modelling scenarios and discusses their representation within DMLex. Section 5 concludes the paper and outlines possible directions for future work.

2 Background

2.1 Phraseological complexity and conceptual organisation

The layered nature of phraseological units has long required complex lexicographic solutions. Phraseological expressions involve not only the formal construction itself, but also figurative meaning, relations with semantically related expressions, and affiliation with broader conceptual and thematic domains. Such a multidimensional organisation has traditionally posed challenges for phraseological lexicography, particularly because semantically related PUs are often formally unrelated and therefore difficult to organise within linear alphabetical structures.

The conceptual organisation of phraseological resources has a long lexicographic tradition. Earlier thesaurus-based approaches attempted to organise lexical material according to ideas and conceptual domains rather than alphabetic order (McArthur, 1986; Geeraerts, 1989). Similar principles were later adopted in phraseological lexicography, where PUs were grouped into semantic domains such as emotions, conflict, success, or failure. In Croatian phraseography, conceptual organisation was explored in previous work on the *Online Dictionary of Croatian Idioms*, where PUs were manually grouped into thematic fields and connected through hyperlinks and semantic cross-references (Filipović Petrović & Parizoska, 2019; Filipović Petrović & Beliga, 2025).

This research demonstrated that thematic organisation frequently reflects interactions between phraseological structure and semantic interpretation. For example, adjectival phraseological constructions often describe states or stable properties, while verbal constructions typically express processes, transitions, or events. Such distinctions correspond to thematic

organisation within the analysed phraseological resource and contribute to the semantic structuring of phraseological networks.

2.2 Digital phraseography and implicit semantic structures

Digital lexicography has significantly transformed the representation and accessibility of lexical resources. One of the major advantages of digital dictionaries is the possibility of moving beyond the limitations of linear alphabetical organisation through hyperlinks, cross-references, and thematic navigation structures. This development has been particularly important for phraseological dictionaries, where semantically related expressions are often formally unrelated and therefore difficult to connect within traditional print-based structures.

Within digital phraseography, hyperlinks and thematic indices made it possible to organise PUs as interconnected semantic networks rather than isolated alphabetically ordered entries. Such solutions substantially improved user navigation and enabled concept-oriented access to phraseological meaning. In the *Online Dictionary of Croatian Idioms*, this organisation gradually developed from hyperlink-based semantic connections between individual entries to a separate thematic index grouping PUs into broader conceptual domains and subdomains.

However, although digital environments have made semantic relations more visible and navigationally accessible, many of them remained implicit from a computational perspective. The relations are visible to users but are not explicitly formalised in the underlying data model. Thematic affiliation, conceptual hierarchy, synonymy, and constructional relations frequently exist only as interface-level associations encoded in customised XML structures developed for a specific lexicographic environment.

While digital dictionaries increasingly contain rich semantic information curated by lexicographers, such information often cannot be reliably interpreted or reused by NLP systems without additional semantic formalisation. Phraseological resources therefore illustrate a broader challenge in digital lexicography: the transition from navigational hypertext structures to interoperable semantic representations.

2.3 Interoperable lexical modelling

Recent developments in lexical data modelling have increasingly focused on interoperability, semantic explicitness, and computational reuse. Frameworks such as the Lexical Markup Framework (LMF) (Francopoulo, 2013) and OntoLex-Lemon (McCrae et al., 2017) provide standardised infrastructures for representing lexical information in linked data and NLP-oriented environments. Within these approaches, lexical units, senses, and semantic relations are represented explicitly through formally defined structures intended for reuse across systems and resources.

In this context, DMLex (Filip et al., 2025), developed by the OASIS LEXIDMA Technical Committee, is a standardised, serialization-independent data model for digital dictionaries. It defines a common structural vocabulary for objects such as entries, forms, senses, examples and relations, while allowing resource-specific categories and relation types to be declared through its Controlled Values and Linking modules. The model was selected for the present study because these mechanisms make it possible to represent relations between lexicographic objects explicitly and to distinguish relations anchored at entry level from those

anchored at sense level. This is particularly important for a phraseological resource in which semantic and constructional relations are currently encoded in a Lexonomy-specific structure and are largely intended for human navigation. Recent studies have primarily explored DMLex as a conversion or transport format for integrating legacy dictionaries into collaborative digital infrastructures (Krek et al., 2025). The present study focuses on a different problem: the expressive capacity of DMLex for modelling semantically complex phraseological resources.

Particularly relevant for the present study are relational approaches to lexicographic modelling that avoid deeply recursive hierarchical structures. Měchura (2023) argues that lexicographic relations are often represented more transparently through explicitly encoded relations between objects than through recursive embedding of subsenses and subentries. This principle is especially important for phraseological resources, where semantic organisation frequently involves overlapping classifications, thematic intersections, sense-level distinctions, and constructional relations that cannot easily be represented through strictly hierarchical structures.

3 Methodology

3.1 Source resource and thematic organisation

The study is based on the *Online Dictionary of Croatian Idioms* (ODCI) (Filipović Petrović & Parizoska, 2023), an open-access, corpus-driven phraseological resource developed in the Lexonomy environment. The dictionary was conceived as a born-digital resource combining manual lexicographic work with corpus-based methods and digital lexicographic tools. Entries were compiled manually by linguistically trained lexicographers using corpus evidence from the Croatian web corpus hrWaC accessed through Sketch Engine, while the dictionary itself was written and published in Lexonomy.

The semantic organisation of the resource evolved gradually during dictionary development. In the initial stage, semantically related PUs were connected through manually curated hyperlinks placed within dictionary entries in the *see also (compare)* section. These links primarily connected synonymous or conceptually related phraseological units and enabled non-linear navigation between expressions that would otherwise remain distant in an alphabetically organised dictionary. As the number and complexity of these semantic relations increased, a separate thematic index was developed, grouping PUs into conceptual domains and subdomains such as LJUTNJA ('ANGER'), RASPOLOŽENJE ('MOOD'), POTEŠKOĆE ('DIFFICULTIES') or LJUDSKO PONAŠANJE I KARAKTER ('HUMAN BEHAVIOUR AND CHARACTER').

The current thematic index contains 64 concepts and includes 430 PUs. It was compiled manually by the lexicographers based on the meanings established during dictionary compilation. PUs were assigned to one or more domains or subdomains according to shared conceptual meaning, while subdomains frequently reflect systematic correspondences between grammatical structure and semantic interpretation. In the domain LJUTNJA ('ANGER'), for example, adjectival constructions such as *crven od bijesa* ('red with anger') are classified under the subdomain BITI LJUTIT ('TO BE ANGRY'), whereas verbal constructions such as *pozelenjeti od bijesa* ('to turn green with rage') belong to the subdomain NALJUTITI SE ('TO BECOME ANGRY'). The organisation therefore reflects not only semantic similarity but also constructional distinctions related to stative and inchoative meanings.

Figures 1–3 illustrate the structure of the resource. Figure 1 shows the dictionary entry as presented in the user interface, including grouped variants, examples, and hyperlinks to



FIGURE 1 Dictionary entry in the *Online Dictionary of Croatian Idioms*

```
<entry>
  <headword>crven od bijesa</headword>
  <headword>zelen od bijesa</headword>
  <idiom>
    <definition>jako ljutit</definition>
    <example>Nakon pola sata uvjeravanja Ferdo je još bio odjeven, a ja već <h>crvena od bijesa</h> jer se
      nismo pomakli sa početne točke.</example>
    <example>Ti obijesni gade! - zaurlah i <h>zelen od bijesa</h> pojurih u njegovom smjeru.</example>
  </idiom>
  <idiom2>
    <idiom2form>pocrvenjeti od bijesa</idiom2form>
    <idiom2form>pozelenjeti od bijesa</idiom2form>
    <definition>jako se naljutiti, ražestiti se</definition>
    <example>Udario sam još jednog muškarca, a on je sav <h>pocrvenio od bijesa</h>.</example>
    <example>Babi sam izbeljila jezik u facu u svoj njegovoj dužini (a kažu da imam dug jezik). Stara <h>je
      pozelenila od bijesa</h>. Dobila sam dobro dvije tri po riti.</example>
  </idiom2>
  <theme>
    <themetext>Usporedi</themetext>
    <themeIdioms>ljut kao ris</themeIdioms>
    <themeIdioms>ljut kao vrag</themeIdioms>
    <themeIdioms>kipjeti od bijesa</themeIdioms>
    <themeIdioms>para ide na uši komu</themeIdioms>
  </theme>
</entry>
```

FIGURE 2 Underlying Lexonomy structure of the dictionary entry



FIGURE 3 Thematic index showing conceptual domains and subdomains in the *Online Dictionary of Croatian Idioms*

semantically related PUs. Figure 2 presents the underlying Lexonomy/XML-like structure of the same entry. Although the resource already contains considerable structural organisation, semantic relations are not formally typed or explicitly encoded in a standardised machine-readable way. Figure 3 shows the separate thematic index, where PUs are grouped into conceptual domains and subdomains through manually curated lists and hyperlinks.

From the perspective of computational modelling, the resource therefore contains a rich but largely implicit semantic organisation. Thematic affiliation, synonymy, and conceptual proximity are primarily represented through interface-level associations intended for human navigation and interpretation. While digital lexicographic platforms successfully overcome many limitations of linear dictionary organisation by enabling hyperlinking and conceptual browsing, the underlying semantic relations often remain insufficiently explicit for computational reuse and interoperability. The present study addresses this problem by aligning the existing phraseological structures with the DMLex model in order to externalise these relations as formally represented semantic structures.

3.2 Identification of recurring lexicographic structures

The study began with an analysis of the source resource in order to identify recurring lexicographic structures relevant for DMLex modelling. The objective was to determine which types of semantic and structural information recur throughout the resource and therefore require explicit representation in an interoperable model.

The analysis revealed several recurrent patterns. These include lexical and constructional variation, polysemy, synonymic relations, thematic classifications, and links between conceptual domains and subdomains. Although realised through individual entries and thematic lists, these patterns repeatedly connect multiple lexicographic objects across the resource and therefore extend beyond the boundaries of single dictionary entries.

The identified structures were subsequently grouped into a set of representative modelling scenarios. The scenarios were designed to capture recurring organisational principles that

can be applied to multiple phraseological units across the resource. The first scenario focuses on lexical variation, constructional relations, and thematic classification. The second examines polysemous phraseological units whose thematic affiliation depends on the interpretation of individual senses. The third addresses multiple thematic affiliations and the network-like organisation of conceptual domains.

The analysis was informed by relational approaches to lexicographic modelling proposed by Měchura (2023), who argues that lexicographic information is often represented more transparently through explicitly encoded relations between objects than through complex nested structures. This perspective proved particularly useful for the present resource, where thematic classifications, synonymic relations, constructional correspondences, and sense-level distinctions frequently connect entries, senses, and conceptual categories in ways that are not adequately captured by entry-internal organisation alone.

3.3 Aligning phraseological structures with DMLex

The structures identified in Section 3.2 are mapped onto DMLex by the modelling principles set out below, applied to specific scenarios in Section 4 and illustrated in Figures 4–8. In the present study, DMLex modelling refers to the identification of relevant lexicographic objects in the source resource and their representation through DMLex entries, senses and explicitly typed relations, supplemented by resource-specific controlled values. DMLex separates a normatively fixed structural vocabulary — *entry, sense, relation, member, partOfSpeech*, and properties such as *headword, type, role*, and *ref* — from the lexicographic content expressed through it (Filip et al., 2025). Property values are determined by the resource itself and declared through the Controlled Values and Linking modules, which allow the introduction of domain-specific relation types without deviating from the standard. We therefore introduce five resource-specific types: *lexicalVariation* (formal variants sharing a constructional pattern), *stativeInchoative* (paired stative and inchoative constructions expressing the same conceptual content through different aspectual profiles), *phraseologicalSynonymy* (semantic equivalence between phraseological units), *belongsToDomain* (thematic affiliation), and *subDomainOf* (hierarchy between conceptual categories), each with the corresponding *memberType* declarations.

A central decision concerns the granularity at which relations are anchored. DMLex allows a member's reference to point either to an entry or to a sense; we adopt throughout the principle that form-related relations target entries while meaning-related relations target senses. Lexical variation and constructional alternations are accordingly modelled at the entry level (Figure 4); thematic affiliation and phraseological synonymy at the sense level (Figures 5–7). This distinction is not explicitly encoded in the source Lexonomy resource and becomes essential for polysemous units (Section 4.2) and for senses with multiple thematic affiliations (Figure 8; Section 4.3).

Thematic categories themselves are represented as entry objects rather than labels, with their hierarchy expressed through subDomainOf (Figures 5). This makes the conceptual structure queryable as a graph and provides potential anchor points for future links to external resources such as OntoLex-Lemon (McCrae et al., 2017), FrameNet, or Wikidata. Because *partOfSpeech* is the normative DMLex property name, it is retained in the model; however, the values used here describe the constructional category of the phraseological unit as a whole (*verbalPhraseme, adjectivalPhraseme*), not the part of speech of an individual lexical head.

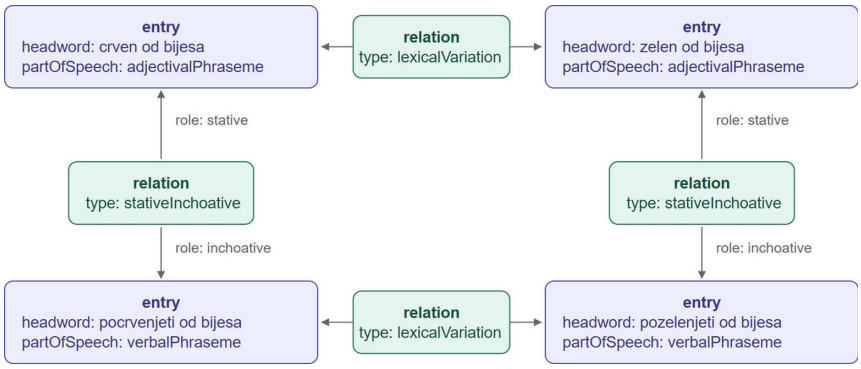


FIGURE 4 Paradigmatic structure: variant and constructional relations among phraseological entries

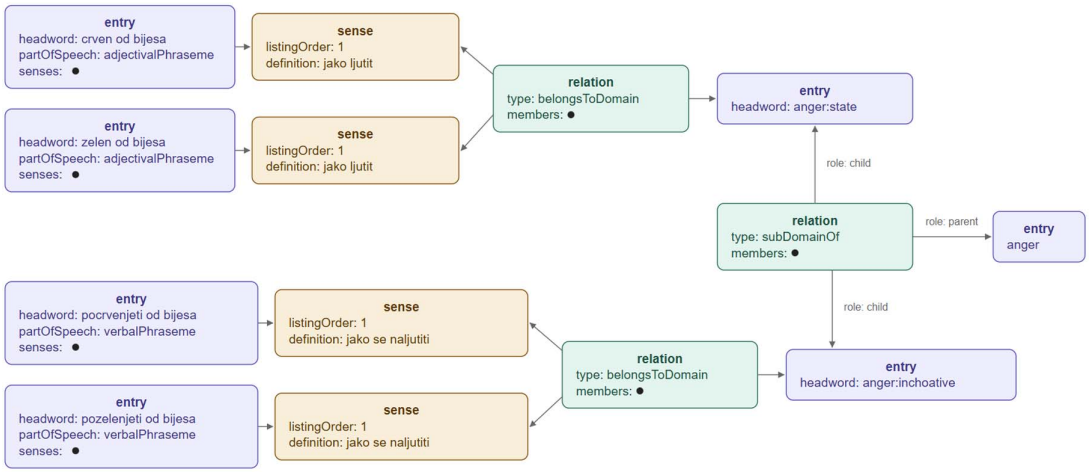


FIGURE 5 Thematic classification: phraseological senses linked to concept entries via *belongsToDomain* and *subDomainOf*

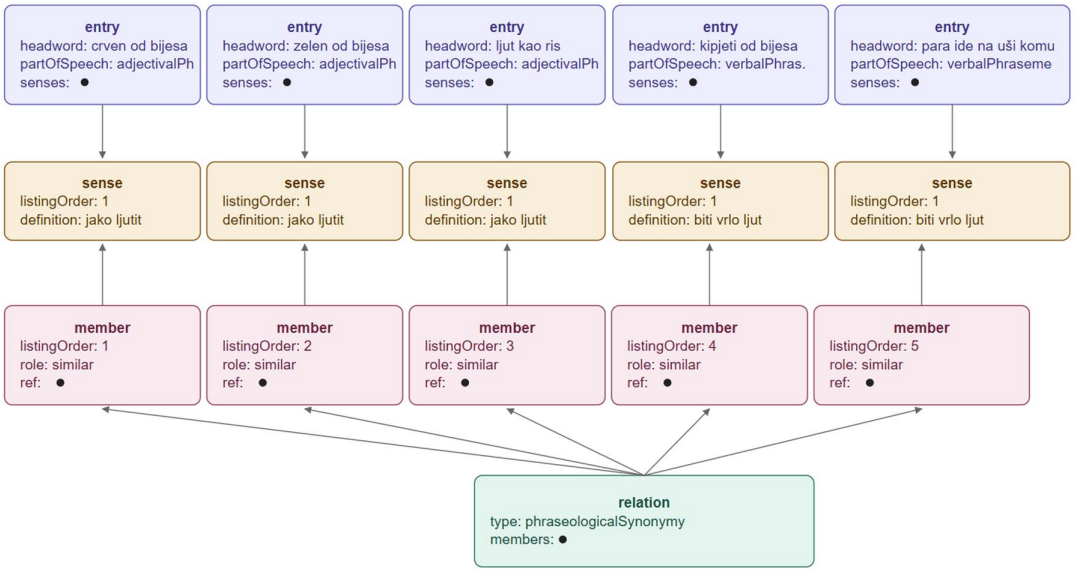


FIGURE 6 Phraseological synonymy at sense level: explicit member objects connect the relation to senses

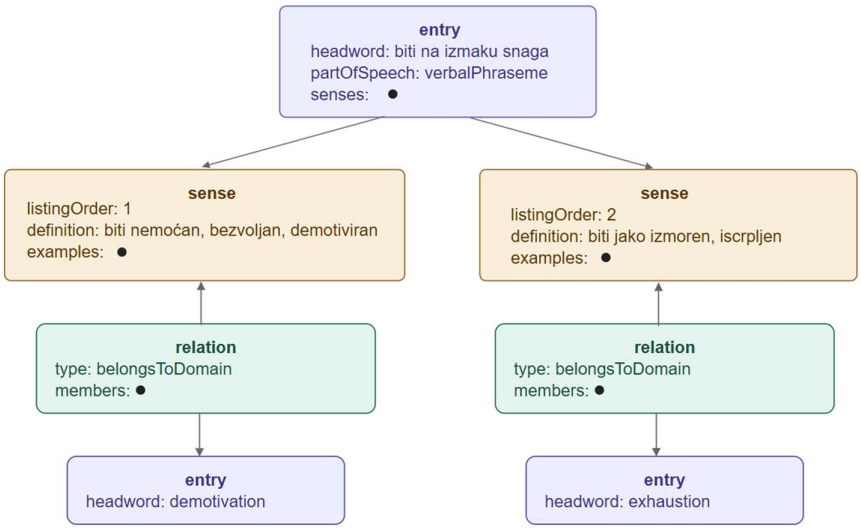


FIGURE 7 Polysemy and sense-level thematic affiliation: one entry with two senses, each affiliated to a different thematic concept

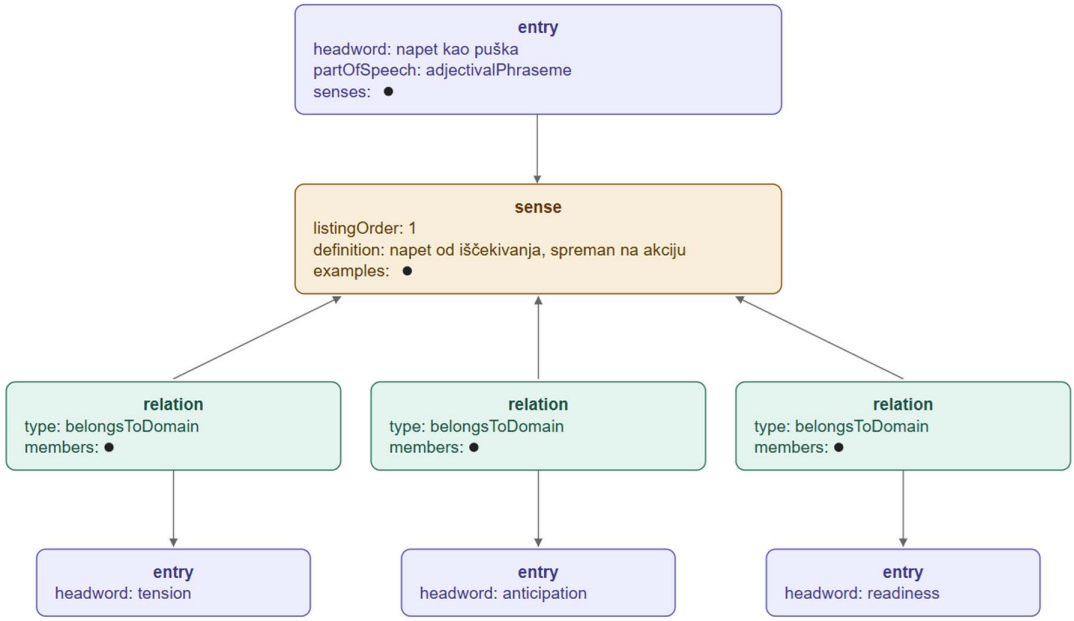


FIGURE 8 Multiple thematic affiliation: one sense referenced by multiple *belongsToDomain* relations, each targeting a different concept

4 Analysis: Modelling scenarios

4.1 Variant and conceptual relation networks

The first modelling scenario is based on a cluster of semantically related PUs associated with the thematic domain LJUTNJA (ANGER), already illustrated in the source resource through both the dictionary entry and the thematic index (Figures 1–3). This example is particularly suitable for demonstrating the interaction between phraseological variation, constructional structure, and conceptual classification. In the original Lexonomy presentation, the PUs *crven od bijesa* ('red with anger') and *zelen od bijesa* ('green with anger') are grouped within the

same entry and defined as 'very angry', while *pocrvenjeti od bijesa* ('to turn red with anger') and *pozelenjeti od bijesa* ('to turn green with anger') are presented as related forms with the meaning 'to become very angry'. The entry also contains links to semantically related PUs such as *ljut kao ris* ('angry as a lynx'), *kipjeti od bijesa* ('to boil with anger') and *para ide na uši komu* ('steam comes out of someone's ears').

The cluster combines several different types of phraseological relations. The relation between *crven od bijesa* and *zelen od bijesa* can be represented as lexical variation, since the two PUs share the same constructional pattern and differ primarily in the lexical component expressing colour. The same applies to *pocrvenjeti od bijesa* and *pozelenjeti od bijesa*. By contrast, the relation between *crven od bijesa* and *pocrvenjeti od bijesa* involves a constructional shift from an adjectival to a verbal construction, together with a semantic shift from a stative meaning to an inchoative one. It is therefore treated as a constructional state/change-of-state relation (Figure 4).

This distinction also motivates the thematic organisation of the cluster (Figure 5). The adjectival constructions *crven od bijesa* and *zelen od bijesa* belong to the subdomain BITI LJUTIT (TO BE ANGRY), whereas the verbal constructions *pocrvenjeti od bijesa* and *pozelenjeti od bijesa* belong to the subdomain NALJUTITI SE (TO BECOME ANGRY). The thematic classification is therefore linked to semantic interpretation: adjectival constructions profile anger as a state, while verbal constructions profile the onset of that state. A further layer of organisation is provided by synonymic relations between phraseological senses (Figure 6). PUs such as *crven od bijesa*, *ljut kao ris*, *kipjeti od bijesa* and *para ide na uši komu* are formally heterogeneous but semantically close.

4.2 Polysemy and differentiated thematic classification

The second modelling scenario focuses on polysemy. The PU *biti na izmaku snaga* ('to be at the end of one's strength') exhibits two related but distinct senses in the dictionary. In the first sense, the PU denotes a psychological or emotional state ('to be helpless, demotivated, emotionally exhausted'), as illustrated in EXAMPLE 1. In the second sense, it refers to physical exhaustion ('to be extremely tired, exhausted'), as illustrated in EXAMPLE 2.

(1) My husband and I are at the end of our strength, which has seriously damaged our relationship; we argue, accuse, and insult each other on a daily basis.

(2) The government requested assistance from the European Union in fighting the fires because the local firefighters were at the end of their strength.

The two senses are associated with different thematic domains. The first belongs to the domain RASPOLOŽENJE (MOOD), whereas the second is classified under UMOR, ISCRPLJENOST (FATIGUE, EXHAUSTION). Although the phraseological form remains unchanged, its thematic affiliation depends on the interpretation of the PU in context.

This example shows that thematic classification is not always determined at the level of the phraseological unit as a whole. In polysemous entries, different senses may participate in different conceptual structures and therefore establish different relations to thematic domains (Figure 7). The model consequently represents thematic affiliation separately for each sense.

4.3 Multiple thematic affiliation and network organisation

The third modelling scenario addresses multiple thematic affiliations within the conceptual organisation of the phraseological resource. Whereas the previous case focused on different

thematic classifications associated with different senses of the same PU, the present scenario concerns phraseological units whose interpretation is associated with more than one conceptual domain.

The PU *napet kao puška* ('tense as a rifle') illustrates this type of organisation. In the dictionary, the PU denotes a state of tension, anticipation, and readiness for action. As illustrated in EXAMPLE 3, the PU characterises a person who is intensely focused on an anticipated event and fully prepared to act. Its semantic interpretation therefore overlaps with multiple conceptual domains, including TENSION, ANTICIPATION, and READINESS (Figure 8). These domains collectively characterise aspects of human behaviour and psychological disposition.

(3) We were running late, of course. Marino, as usual, had arrived at least twenty-five minutes earlier, tense as a rifle and ready for action.

This example shows that thematic organisation in the resource is not limited to a one-to-one correspondence between phraseological units and conceptual categories. A single phraseological sense may participate simultaneously in several thematic domains, reflecting the interconnected nature of conceptual classification in phraseological lexicography. From a modelling perspective, thematic affiliation is represented through explicit semantic relations linking the phraseological sense to multiple conceptual domains. The focus of this scenario is therefore the representation of multiple conceptual affiliations within the semantic network.

5 Discussion

The modelling scenarios analysed in this study suggest that DMLex provides a flexible framework for representing many of the semantic structures encountered in phraseological resources. In particular, the analysed cases indicate that semantic relations originally expressed through thematic organisation, hyperlinks, and other navigational structures can be reformulated as explicit relations within an interoperable data model.

At the same time, the analysis revealed that phraseological resources raise several modelling questions that extend beyond the representation of individual entries and senses. One of these concerns the status of thematic domains and conceptual categories. In the Online Dictionary of Croatian Idioms, thematic domains were originally developed as a user-oriented navigation system designed to facilitate concept-based access to phraseological meaning. When transferred into a formal data model, however, such categories acquire a more explicit semantic status. This raises questions about the nature of the entities being represented and about the boundary between lexicographic organisation and conceptual modelling.

A related issue concerns the representation of overlapping classifications. Thematic organisation in phraseological resources rarely follows a strictly hierarchical structure. Individual phraseological units may participate in several conceptual domains simultaneously, while polysemous units may exhibit different thematic affiliations depending on the interpretation of a particular sense. Such cases suggest that phraseological resources are more naturally represented as networks of relations than as systems in which each unit is assigned to a single conceptual category. The present study therefore supports modelling strategies based on explicitly encoded relations.

The analysed scenarios also demonstrate that many semantic structures remain implicit in traditional digital dictionaries. Hyperlinks, thematic indices, and cross-references often convey rich semantic information to human users, yet these relations are not necessarily

available for computational processing. Formal modelling requires such relations to be identified, interpreted, and encoded explicitly. This process inevitably introduces analytical decisions regarding the distinction between state and change-of-state meanings, the delimitation of thematic domains, and the choice of relation types used to connect phraseological objects.

Because the present study is based on representative modelling scenarios rather than a completed conversion of the entire resource, it is not yet possible to provide a reliable estimate of the total time required. The structural mapping of entries, forms, senses and examples can be systematised once stable identifiers and controlled values have been established, but the identification and typing of semantic relations, the resolution of polysemous senses and the validation of multiple thematic affiliations still require expert intervention. The main challenge for downstream applications is therefore not only formal conversion but also the consistency and granularity of the encoded relations. Semantic search, cross-lingual linking and NLP-oriented use depend on sufficiently complete sense distinctions, stable identifiers and clearly defined relation types. These applications will consequently require task-specific evaluation once the full resource has been converted.

The present study focused primarily on semantic and conceptual organisation. However, phraseological resources frequently contain additional layers of information that are equally important for interpretation and use. In the Online Dictionary of Croatian Idioms, these include stylistic labels such as colloquial and vulgar, usage notes, and other explanatory information intended to guide users in appropriate contextual interpretation. Particularly noteworthy are usage notes describing intentional ad hoc modifications of phraseological units in discourse. Usage notes of this kind remain relatively uncommon in phraseological dictionaries because intentional phraseological modifications are often difficult to identify systematically in corpus data and frequently depend on context-specific communicative intentions. While DMLex provides mechanisms for representing many types of lexicographic information, the representation of discourse-specific modifications and other context-dependent phraseological phenomena lies beyond the scope of the present study. Future work should therefore examine how such pragmatic, stylistic, and discourse-related information can be integrated into phraseological data models while preserving the interpretative distinctions that motivate their inclusion in lexicographic description.

6 Conclusion and future work

This study explored the applicability of DMLex for representing the semantic organisation of a phraseological resource originally developed for human-oriented navigation. Using the *Online Dictionary of Croatian Idioms* as a case study, the analysis focused on several recurring modelling scenarios, including variant and conceptual relation networks, polysemy and thematic classification, and multiple thematic affiliation.

The results indicate that DMLex provides sufficient expressive capacity for representing not only phraseological entries and senses, but also the semantic relations that connect them within broader conceptual structures. Relations originally realised through hyperlinks, thematic indices, and other interface-level mechanisms can be reformulated as explicit and interoperable semantic representations.

More broadly, the study demonstrates that a substantial amount of semantic knowledge already present in digital phraseological resources can be made computationally accessible through formal modelling. The challenge is therefore not only to represent individual

phraseological units, but also to capture the conceptual networks, thematic classifications, and semantic relations that organise them.

Future work will focus on extending the present approach from individual modelling scenarios to the complete phraseological resource. The long-term objective is to represent the entire *Online Dictionary of Croatian Idioms* within a DMLex-compatible framework and to develop a user interface based on the resulting data model. Such a representation would facilitate interoperability with other lexical resources, support multilingual linking, and create new opportunities for computational processing and cross-linguistic phraseological research.

References

Baldwin, T. & Kim, S. N. (2010). Multiword expressions. In N. Indurkhya & F. J. Damerau (Eds.), *Handbook of natural language processing* (2nd ed., pp. 267–292). CRC Press.

Constant, M., Eryiğit, G., Monti, J., van der Plas, L., Ramisch, C., Rosner, M. & Todirașcu, A. (2017). Multiword expression processing: A survey. *Computational Linguistics*, 43(4), 837–892. https://aclanthology.org/J17-4005/

Filip, D., Jakubíček, M., Kovář, V., Krek, S., McCrae, J. & Měchura, M. (Eds.). (2025). *Data Model for Lexicography Version 1.0* (OASIS Standard). OASIS. https://docs.oasis-open.org/lexidma/dmlex/v1.0/dmlex-v1.0.html

Filipović Petrović, I. & Beliga, S. (2025). Can AI understand Croatian idioms? Assessing large language models in lexicographic tasks. *Prispevki za novejšo zgodovino*, 65(3), 218–242. https://doi.org/10.51663/pnz.65.3.09

Filipović Petrović, I. & Parizoska, J. (2019). Konceptualna organizacija frazeoloških rječnika u e-leksikografiji. *Filologija*, 73, 27–45.

Filipović Petrović, I. & Parizoska, J. (2023). *Frazeološki rječnik hrvatskoga jezika*. https://lexonomy.elex.is/#/frazeoloskirjecnikhr

Francopoulo, G. (2013). *LMF lexical markup framework*. Wiley.

Geeraerts, D. (1989). Principles of monolingual lexicography. In F. J. Hausmann (Ed.), *Wörterbücher. Ein internationales Handbuch zur Lexikographie* (Vol. 1, pp. 287–296). Walter de Gruyter.

Krek, S., Ponikvar, P., Repar, A., Kosem, I. & Lindemann, D. (2025). DMLex on Wikibase: Legacy dictionaries as collaboratively editable dataset. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda & A. Salgado (Eds.), *Electronic lexicography in the 21st century (eLex 2025): Intelligent lexicography. Proceedings of the eLex 2025 conference* (pp. 174–188). Lexical Computing CZ.

McArthur, T. (1986). *Worlds of reference: Lexicography, learning, and language from the clay tablet to the computer*. Cambridge University Press.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P. & Cimiano, P. (2017). The OntoLex-Lemon model: Development and applications. In *Proceedings of the eLex 2017 Conference* (pp. 587–597). Lexical Computing.

Měchura, M. (2023). Avoiding recursion in the representation of subsenses and subentries in dictionaries. *International Journal of Lexicography*, 36(3), 260–278. https://doi.org/10.1093/ijl/ecad012

Miletić, F. & Schulte im Walde, S. (2024). Semantics of multiword expressions in transformer-based models: A survey. *Transactions of the Association for Computational Linguistics*, 12, 593–612. https://doi.org/10.1162/tacl_a_00657

Sag, I. A., Baldwin, T., Bond, F., Copestake, A. & Flickinger, D. (2002). Multiword expressions: A pain in the neck for NLP. In *Proceedings of the 3rd International Conference on Intelligent Text Processing and Computational Linguistics* (CICLing 2002) (pp. 1–15).

Contact information

Ivana Filipović Petrović
Linguistic Research Institute, Croatian Academy of Sciences and Arts
ifilipovic@hazu.hr

Slobodan Beliga
Faculty of Informatics and Digital Technologies, University of Rijeka
sbeliga@inf.uniri.hr

Acknowledgements

This research was supported by the project Hybrid AI Approaches to Natural Language Processing and Knowledge Generation – HyAI (uniri-iz-25-215), funded by the European Union – NextGenerationEU.

Marta Garcia-Casado

NEW SENSES, NEW METHODS

Clustering as a Guide for Detecting Semantic Neologisms

Abstract Identifying semantic neology remains a significant challenge for lexicography due to the gradual nature of meaning shifts. This study proposes a methodological framework for the automatic detection of verbal semantic neology in Catalan, leveraging recent advances in Natural Language Processing (NLP). Utilizing a general-language corpus spanning 2000-2021, this research employs a Large Language Model to generate contextualized embeddings for 90 target verbs, capturing nuanced semantic variation. The HDBSCAN clustering algorithm is then applied to group these representations by usage similarity. To validate the results, a lexicographic baseline is established using the *Diccionari essencial de la llengua catalana*; clusters that do not align with registered senses are flagged as candidate neologisms for qualitative analysis. This qualitative analysis focuses on shifts in argument structure and semantic implicatures. The findings emphasize the value of NLP tools for low resource languages like Catalan that have growing but comparatively few digital resources. Ultimately, this study advocates for a hybrid lexicographic model where machine learning serves as an exploratory tool to augment, rather than replace, human expertise. This approach ensures that modern dictionaries can responsibly integrate technological innovation while maintaining academic and ethical standards.

Keywords Lexicography; computational semantics; semantic neology; NLP; HDBSCAN

1 Introduction

Linguistic research has undergone a notable evolution following the growing development of new technologies. With the emergence of computational linguistics and natural language processing (NLP), many tasks have been automated or are assisted, such as machine translation or the exploitation of large corpora. More specifically, in the field of neology, although various technological advances have allowed for faster detection of formal neology, the automatic detection of semantic neology (the emergence of new meanings in existing signifiers) remains a goal yet to be achieved. This difficulty lies in the fact that semantic change involves gradual shifts in use and subtle contextual variations that traditional tools fail to detect accurately.

This situation highlights the need to adopt a new perspective. Current approaches (Torres-Rivera et al., 2020), based on primarily static methodologies and centered on manual analysis, have proven insufficient to provide systematic and fast detection given the massive volume of data that should be analyzed. The present work proposes to overcome these limitations through a hybrid model that combines the rigor of qualitative lexicographic analysis with the benefits of machine learning.

The theoretical starting point of this research is distributional semantics, which assumes that the meaning of a word can be induced from its contexts of occurrence (Harris, 1954; Firth, 1957). Under this premise, a change in the argument structure of a verb or in the semantic implicatures of its arguments is a likely indicator of a change in its meaning. Thus, we recover and expand the proposal of Bastuji (1974) —focused on features like [± concrete] or [± human]— and integrate it into a deeper analytical framework that considers argument structures (Kroeger, 2004) and semantic implicatures (Hanks, 2013) in the transition from non-neological to neological use.

To operationalize this proposal, this study leverages recent advances in natural language processing, specifically the use of contextualized embeddings obtained through BERTa, a language model developed for Catalan (Armengol-Estape et al., 2021). These embeddings capture context-dependent semantic variations, which is ideal for identifying the shifts in meaning that characterize neological verbs.

In this context, the present work aims to evaluate a methodological proposal for the automatic detection of verbal semantic neology in Catalan. The specific objectives are:

1. To generate vector representations of verbal uses to detect distribution patterns through clustering algorithms (HDBSCAN).
2. To propose a methodology for the detection and validation of the results obtained from (1), checking whether these clusters correspond to meanings already registered in the *Diccionari essencial de la llengua catalana* (2024), the main authoritative dictionary for Catalan, or if instead they constitute candidates for semantic neologisms.

To achieve these objectives, a general corpus covering the period 2000-2021 was used, provided by the Institut d'Estudis Catalans and based on the Corpus Textual Informatitzat de la Llengua Catalana (CTILC). From this corpus, two Catalan verbs are analyzed for the purpose of this study. The procedure combines automatic sense induction with a qualitative analysis: those usage groups (clusters) that cannot be assigned to any dictionary definition are subjected to an examination of their argument structure and semantic implicatures. In this way, the proposed methodology does not seek to replace the lexicographer but to augment their analytical capacity, conceiving automatic neologism detection as a first step in the work pipeline that allows for progress towards a lexicographically relevant description of semantic change in the digital era.

2 Semantic neology and its current detection

Every language undergoes incessant changes at the lexical-syntactic level, which has consequences for both the structure of the language and the lexicon that comprises it (Lüdkte, 1998), since "no hay lengua que no se renueve ni período que no experimente cambios" (Fernández-Sevilla, 1982). These changes are closely linked to social evolution; the malleability of language allows it to adapt to new realities, generating unprecedented meanings or nuances that respond to social changes. As a result of this capacity for innovation, language constantly produces new terms (Alvar Ezquerra, 1998).

This creation of new lexical units is the central axis of neology, the linguistic discipline that analyzes novel words that emerge in a language (Cabré, 2002). Its object of study is neologisms, which can originate either through a new signifier —formal neology— or by attributing a different meaning to an already existing word —semantic neology. Although its etymological root simply suggests the idea of a "new word," the definition of a neologism requires the consideration of various factors to achieve a polyhedric characterization.

To delimit the concept and its detection method, this work adopts the lexicographical criterion (Rey, 1976), which identifies a neologism based on its absence within an exclusion corpus —in this case, a reference dictionary. Despite the challenges that this approach entails, it allows for the establishment of a baseline for comparison that guarantees the collection of systematic and objective data. By relying on this criterion, it is possible to minimize subjectivity and individual biases in the detection of new units (Bernal, 2022).

2.1 Delimitation of the concept

While defining neologisms in general is often complicated by inconsistent detection criteria, the scientific community shows solid unanimity when describing semantic neology (Díaz-Hormigo, 2020). Authors such as Torijano (2024) view this theoretical hegemony positively, arguing that the phenomenon is situated on an independent conceptual scheme that is clearly differentiated from formal neology.

The historical starting point for approaching its formal definition is located in the works of Bréal (1897, p. 295), who for the first time conferred a dual nature upon the neologism by stating that "un pays où il serait interdit d'innover, retirerait à son langage toute chance de se développer". This characterization was subsequently supported by theorists such as Matoré (1952) and Gardin et al. (1974), who ratified that neology manifests exclusively in units that already exist formally in the language, but to which a new semantic load is assigned. In the same vein, Rey (1976) reaffirmed that the neological quality can operate exclusively on the semantic level, without requiring transformations in the graphic or phonological structure of the word.

To further delve into the characterization of this phenomenon, it is essential to go back to the formulations of Bastuji (1974) and Guilbert (1975), who directly link and condition the existence of the semantic neology to the concept of polysemy. Bastuji (1974, p. 6) formally defines semantic neology as "un cas particulier de la polysémie, avec un trait diachronique de nouveauté dans l'emploi, donc dans le sens". Following this premise, Guilbert (1975, p. 65) approaches it as the addition of a new meaning to the same already existing phonological segment, remarking that "c'est précisément cette possibilité de la variation dans la combinación des sèmes qui défi nit la néologie sémantique". Thus, polysemy, understood as the coexistence of a new sense with those already consolidated under the same signifier constitutes the exclusive defining axis of semantic neologisms, a position widely endorsed by contemporary literature (Cabré et al., 2002; Renouf, 2014; Adelstein, 2017; Díaz-Hormigo, 2020; Klosa-Kückelhaus et al., 2020). Additionally, this definition can acquire stylistic nuances; Sánchez-Manzanares (2009, p. 128) associates the neology of sense with the classical rhetorical tradition, arguing that "el tropo se define como un cambio de significado" which introduces a literary dimension into the explanation of its formation processes.

However, a comprehensive definition of semantic neology requires transcending the isolated unit and adopting a contextual and dynamic perspective. Guilbert (1973, p. 23) emphasizes that it is methodologically impossible to detect a semantic neologism without observing its environment of appearance, given that mutations of meaning remain hidden and imperceptible in the abstract lexicon until the word is inserted into the sentence and generates a new distribution of syntagmatic components (Harris, 1954). For example, the figurative use of a word implies establishing syntagmatic relations with units belonging to semantic classes distinct from those of its canonical use (as occurs when projecting [human] combinatory features onto terms that originally possessed [non-human] features). The major shortcoming of traditional approaches has been the maintenance of a static perspective focused on describing the consequences or the abstract result of semantic change. A rigorous definition must incorporate a dynamic vision where alterations of the immediate contextual environment are considered the object of study for semantic neology (Guilbert, 1973). Finally, Torijano (2024, p. 80) underlines that a persistent obstacle in the study of this discipline is its scarce presence in grammars and its deficient identification in dictionaries, a gap justified because this lexicogenetic process has traditionally tended to be located in the branch of general and historical semantic change, instead of being fully integrated into studies of neology, which habitually prioritize formal lexical change.

2.2 Current approaches to the detection of semantic neology

Since the beginning of the century, the automation of formal neology detection has been considerably solved —except for syntagmatic compounds and conversions— and several approaches have been developed to automate this task (Cabré et al., 2009; Cartier, 2016; Kerremans et al., 2018; among others). In this context, Castro et al. (2021, p. 384) state that "the comparison of corpus data against a list of words obtained from dictionaries and other sources is a widely used methodology for automatic detection of lexical neology". This inclination towards the lexicographical criterion may be due to the fact that it is the easiest to operationalize computationally, given that, provided a corpus, the words appearing in it are compared with the wordlist of the chosen dictionary, and those that are not registered are extracted. However, although this method provides satisfactory results (Janssen, 2005; Cabré et al., 2009), it cannot be ignored that it presents limitations regarding the detection of some word-formation processes: the method does not allow for the detection of syntagmatic units, subcategorical changes, or semantic neologisms.

In this regard, Renau (2023, p. 59) reaffirms this limitation and highlights the difficulty of resolving it compared to the detection of formal neology, given that "it is much more difficult to detect semantic neology than lexical neology, because in semantic neology the change is not formal, thus, it must be detected in alternative, more complex ways, and the error rate can be high". This complexity, caused by the fact that the changes experienced by semantic neologisms are not merely formal and visible, means they can easily go unnoticed by the neologist (Freixa et al., 2015).

Therefore, to automatically detect semantic neology, it is necessary to develop a new method. Renau (2023, p. 58) emphasizes that "addressing novel sense identification for lexicographic purposes requires covering large segments of vocabulary in a fast and systematic way"; hence, it is essential to establish criteria to systematize the task. The most recent studies (references) conducted on the detection of this neological process propose some methods to achieve this initial goal. In any case, according to Renau (2023, p. 59), all attempts to resolve this issue agree on employing a diachronic approach that compares two or more corpora, using various metrics and strategies to identify semantic neology.

One of the earliest approaches is that of Janssen (2009), based on comparing word frequency between two corpora (a current one and an old one). In this case, the proposed method was that "a word used ten times in texts from earlier dates (reference corpus), but a thousand times in new texts (study corpus), is probably a word that has changed its usage, and thus is probably a semantic neologism" (Janssen, 2009, p. 71). This proposal remained at an abstract level and never materialized into any tool, although similar ones have been developed subsequently.

A common thread among existing attempts at the automatic detection of semantic neology is that they are approached from a static perspective. Torres-Rivera et al. (2020), for instance, go beyond unit frequencies by combining this methodology with tools like word2vec (Mikolov et al., 2013). Although this is an innovative contribution, it may not have been further developed because this tool works with static embeddings (which, unlike contextual embeddings, assign a single fixed representation to a word), reinforcing the current static view of semantic neology. This vision presents clear limitations: it fails to account for the relationship between a semantic neologism and the rest of the words in the sentence within their real contexts of occurrence, as it only considers the changing word, thus attempting to explain how it has changed.

Against this backdrop, the proposal presented here explicitly examines the benefits offered by a dynamic and contextualized representation of meaning in the detection of semantic neology. Given our current stage in the automatic detection of semantic neology, we propose to approach it from the perspective of distributional semantics to orient it toward detection using language models.

3 Emerging computational resources: a paradigm shift

As mentioned above, Bastuji (1974) was one of the first authors to highlight the relevance of contextual modifications in the field of semantic neology. This observation, which was overshadowed by subsequent classifications of said neology, is based on the principles of distributional semantics and posits that meaning arises from the use made of words in real contexts. Thus, a contextual modification will necessarily imply a semantic modification (Harris, 1954; Firth, 1957).

In this way, distributional semantics—which derives from structuralism—proposes that the understanding of meaning must be based on the systematic observation of contexts of use. Thus, and following the path that Bastuji (1974) began to chart, through the analysis of textual corpora, this approach allows for the identification of patterns and changes in semantic implicatures and in the argument structure of words, specifically of verbs, which should indicate possible semantic shifts (Baroni et al., 2010; Erk, 2012; Mikolov et al., 2013) and, therefore, potential neological uses.

Distributional semantics allows us to stop defining meaning exclusively by the intrinsic features of the word or by its membership to rhetorical classifications, so that we can define it by the network of relations it establishes with other units with which it co-occurs in real contexts of use (Boleda et al., 2016). This approach offers a little-explored perspective for the systematic study of semantic neology, as it allows for the objective and systematic detection of new senses, thereby overcoming the limitations of previous traditional classifications based on rhetorical figures.

In recent years, with the emergence of artificial intelligence and new language models, distributional semantics has become the theoretical foundation for the development of new natural language processing tools. Thus, as Martí (2018, p. 18) states, “la tecnología digital ha hecho posible que estas propuestas teóricas aparecidas en los años 50 del pasado siglo dispongan ahora de modelos matemáticos de base estadística para su representación y procesamiento”.

In this approach, the meaning of lexical units is modeled from a massive input of real texts to shape a computational lexicon. In this space, words are represented mathematically as vectors, which are lists of numbers projecting coordinates in a multidimensional space (Boleda, 2020). The fundamental premise of this system is that vector similarity directly correlates with distributional similarity and, therefore, with semantic similarity (Boleda, 2020). To determine how close two words are in terms of meaning, the cosine of the angle formed by their trajectories is commonly used: the closer their vectors are in space, the higher their similarity and conceptual affinity.

The way of generating and structuring these vector representations has undergone a radical methodological evolution in recent decades. In its beginnings, one-hot vectors were employed, a low-computational-cost technique that differentiated terms based exclusively on their surface form, leaving the model completely unable to reflect the actual semantic proximity between words. The true paradigm shift arrived with the introduction of static word embeddings, materialized in algorithms such as word2vec (Mikolov et al., 2013) and GloVe (Pennington

et al., 2014). These models succeeded in learning lower-dimensional representations by capturing semantic relations from co-occurrence in large-scale corpora. However, static embeddings suffered from an insurmountable limitation: they mapped each word to a single, fixed vector. This blocked any attempt to effectively process phenomena such as polysemy, homonymy, or semantic change, given that a word with multiple senses always received the same mathematical representation, regardless of variations in its immediate environment.

To capture the variability inherent to linguistic use, contextualized or dynamic embeddings emerged, which are capable of generating a specific vector for each occurrence of a word based on its syntactic and semantic context. Following the first steps taken by neural networks like ELMo (Peters et al., 2018), the breakthrough of the transformer architecture (Vaswani et al., 2017) and the development of models such as BERT (Devlin et al., 2019) and RoBERTa (Liu et al., 2019) drastically elevated the performance of these models. By being trained in a self-supervised manner with massive corpora, these models assimilate deep contextual dependencies that static technologies completely ignored.

The crucial difference between both methodological approaches becomes evident when observing how verbs behave in the vector space. In a static model, terms like *cuinar* ('to cook') and *coure* ('to bake/boil') will share a very close position because, generally across the training data, both tend to combine with units of the [food] semantic class. However, the static model is blind to the figurative or novel uses within a specific sentence. Conversely, vectors extracted from contextualized embeddings adapt to the dynamic environment of the clause.

Currently, the computational study of distributional semantics relies heavily on transformer-based architectures such as BERT, which bidirectionally analyzes both the left and right contexts of each term (Laicher et al., 2021). By processing text through 12 distinct analytical layers, a highly precise mathematical representation of each token is achieved. To describe and visualize the patterns hidden within these massive representations, contemporary linguistics systematically resorts to advanced dimensionality reduction techniques and clustering algorithms (Geeraerts, 2024; Schlechtweg et al., 2024).

4 Methodology

Once the concept of semantic neology, its classifications, and its current detection models have been analyzed, a new methodological proposal has been designed to efficiently automate its detection. To achieve this, new language models have been leveraged, which, as previously mentioned, allow for the vectorization and characterization of word meanings based on their context of appearance.

Thus, this new methodology starts with a selection of neologisms—specifically, verbs—since they form predicates, making the contextual changes reflected in the argument structure and the semantic implicatures of their arguments more evident. After selecting the neologisms, a corpus was compiled from which in-context uses of these verbs could be extracted, making it possible to analyze the occurrences of each verb using a language model trained exclusively for Catalan: BERTa (Armengol-Estape et al., 2021). This model vectorized each use, and these vectors were then grouped into clusters according to their vectorial similarity and, consequently, their semantic similarity. Once the clusters for each verb were generated, a reference dictionary was consulted, and each definition was associated with its corresponding cluster. For the purpose of this study, the *Diccionari essencial de la llengua catalana* (DEIEC) has been consulted. This volume has been chosen because, unlike the *Diccionari de la llengua catalana* (DIEC2), this dictionary contains definitions where the argument structure of verbs is visible. This isolates

those uses that are not reflected in the dictionary and, therefore, constitute potential semantic neologisms.

4.1 Neologisms selection and corpus characterization

The methodological framework evaluated in this paper is embedded within a broader, comprehensive research project designed to automatically detect verbal semantic neology across a macro-structural selection of 90 target verbs in Catalan. However, due to length limitations, this study focuses on the presentation of the results obtained for two illustrative cases: *amanir* 'to dress' and *quallar* 'to curdle'. These two verbs were strategically selected because they represent two highly distinct typologies of contextual distribution shifts, as explained in section 5.

To conduct this analysis, the primary data source consulted is the Corpus Textual Informatitzat de la Llengua Catalana (CTILC), which has been explicitly granted and provided by the Institut d'Estudis Catalans (IEC) for the specific purposes of this investigation. The CTILC is the institutional reference corpus for contemporary Catalan, designed to capture real linguistic usage. In terms of its composition, the CTILC is a balanced text corpus comprising over 72 million words. It aggregates a vast variety of written documents, from literary works to highly diverse non-literary texts, including press, academic essays, administrative documents and technical texts. For the current study, focusing on contemporary semantic changes, occurrences were extracted from the chronological blocks covering the 2000-2021 period.

4.2 BERTa: a language model for Catalan

To capture the semantic variation of the analyzed verbs, this research incorporates BERTa (roberta-base-ca-v2[1]), a Transformer model based on the RoBERTa architecture and trained for Catalan by the Barcelona Supercomputing Center within the Projecte AINA (Armengol-Estape et al., 2021). Having been pre-trained on a massive corpus of 18 gigabytes of Catalan text, and featuring 12 encoding layers with a dimensionality of 768 variables, this model offers crucial advantages for the detection of semantic neology. Mainly, it overcomes the limitations inherent in traditional static embeddings by generating dynamic, contextualized representations for each occurrence. Furthermore, its self-attention mechanisms bidirectionally analyze the argument structure and verbal complements, mathematically detecting when the verb co-occurs with abstract entities that differ from those found in its non-neological usage.

To carry out this context-adapted vectorization, each sentence in the corpus is subjected to an analysis where the language model evaluates the bidirectional relations between the target verb and the words surrounding it to both its left and right. Once the model processes the contextual relations of the target verb with the other co-occurring units, it assigns to the verb a 768-dimensional vector based on the relationships established with each contextual unit in that specific concordance.

The final step of this pipeline ensures that this vector information can be reused on a massive scale in the subsequent phases of the experiment. To this end, each generated vector is linked to its associated metadata, which includes the complete original sentence and the year in which it was registered.

[1] Accessible at https://huggingface.co/projecte-aina/roberta-base-ca-v2

4.3 Selection and implementation of the clustering algorithm

Once the chosen verbs are transformed into 768-dimensional vector representations, the workflow addresses the challenge of identifying semantic clusters within this high-dimensional space. Direct distance calculation in spaces with so many dimensions tends to scatter the data points and reduce the effectiveness of geometric classification algorithms. To mitigate this limitation, the system first applies a dimensionality reduction process using the t-SNE (t-Distributed Stochastic Neighbor Embedding) technique (Van der Maaten et al., 2008). While non-linear dimensionality reduction techniques like t-SNE can introduce global distance distortions, its application in this study is justified by its ability to preserve local neighborhood structures, which are essential for identifying closely related semantic senses. To mitigate any potential information loss and ensure the semantic validity of the results, the t-SNE projection is strictly used to identify cluster assignments via HDBSCAN. Crucially, the subsequent semantic characterization, specifically the calculation of cluster centroids and the selection of prototypical sentences, is performed by reverting to the original 768-dimensional vector space. This hybrid approach leverages the density-discovery strengths of t-SNE while maintaining the high-dimensional representational richness of the BERTa model for final qualitative validation.

On top of these dimensionality reductions, the pipeline implements the HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) clustering algorithm (Campello et al., 2013). Unlike parametric clustering methods such as k-Means (MacQueen, 1967), which require predefining the exact number of groups and force every element into a category, HDBSCAN autonomously assigns clusters based on data density. This property is essential for the detection of semantic neology, as it allows consolidated verbal usages and new sense extensions to be grouped into independent clusters, while automatically isolating atypical, erroneous, or residual contexts under the label of "noise". All of this is achieved without assigning a predefined number of clusters, thereby allowing the creation of as many clusters as there are real-world usages. To ensure the interpretative coherence of these results, a refinement stage is applied: in cases where clusters exhibit excessive proximity, they are automatically merged to avoid the fragmentation of similar semantic uses, employing a distance threshold of ≥ 2 between their prototypical sentences.

The process concludes with the extraction of prototypical sentences designed to facilitate subsequent qualitative validation by the lexicographer. In order to avoid losing the conceptual richness blurred during dimensionality reduction, this abstraction phase returns to the original high-dimensional semantic space. For each of the clusters identified by HDBSCAN, the algorithm computes its centroid by calculating the vector mean of all the embeddings that comprise that group. Then, the Euclidean distance between all the cluster elements and their corresponding centroid is evaluated, the sentence closest to the centroid as its representative. This sentence is thus chosen as the most prototypical instantiation of the senses grouped by a given cluster. This allows for a subsequent contrastive analysis against the senses registered in the reference dictionary.

5 Results

5.1 Amanir

The verb *amanir* 'to dress' has a total of 92 occurrences in the CTILC corpus between 2000 and 2021. For these 92 occurrences, 3 clusters have been detected, with 5, 23, and 41 occurrences in each, as shown in Figure 1. Of the total recorded contexts, 22 were not grouped into any cluster, as the algorithm read them as noise —that is, as sporadic uses that did not correspond to the general uses detected.

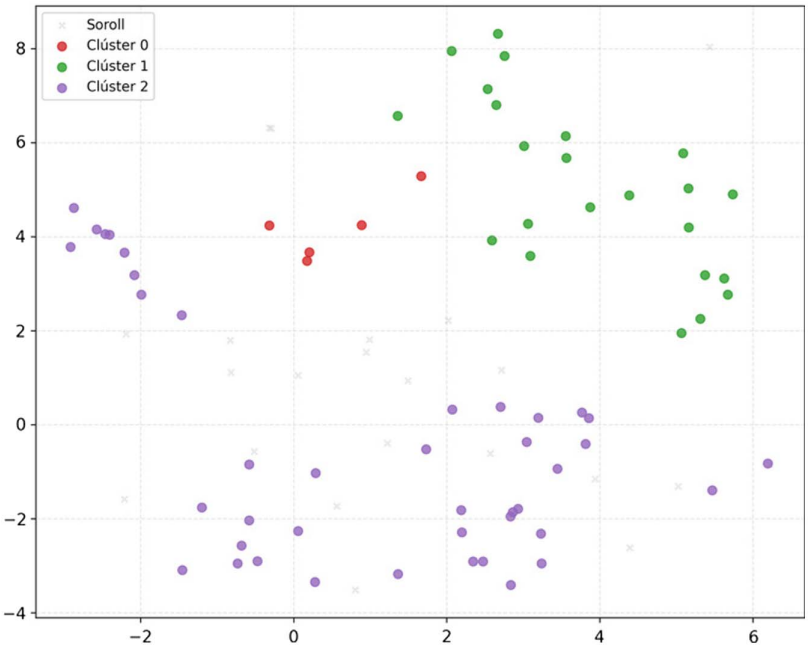


FIGURE 1 Clusters generated for the verb *amanir*

Of the three clusters identified for this verb, the most numerous is cluster 2, with a total of 41 occurrences. This grouping corresponds to the general dictionarized usage, aligning with the definition 'of dressing a food item'. In line with this usage, we find cluster 0, containing 5 occurrences, all corresponding to the same meaning. Nevertheless, a split was made between cluster 0 and cluster 2 because, despite sharing the same lexicographical entry, cluster 0 isolated those contexts that refer to the specific construction *tomàquet d'amanir* 'salad tomato'. Since the contexts are highly similar in these five cases, HDBSCAN assigned them to a standalone group.

Beyond these already dictionarized uses, as shown in Table 1, a third cluster, cluster 1, stands out prominently with 23 occurrences. This cluster features an abstract use of the verb *amanir*, as their contextual environments differ substantially from the exclusively culinary uses found in the other clusters. In the case of the verb *amanir*, as observed in both the prototypical sentences and the selected examples, there is a very clear distributional shift. Although its argument structure remains the same, we observe a shift in one of its arguments, changing from an agentive construction to a causative one. Furthermore, regarding the semantic implicatures of these arguments, significant modifications with major repercussions on the contexts of appearance are also visible: in the dictionarized usage (clusters 0 and 2), the agent is always human and the theme is a food item. In contrast, in the neological uses (cluster 1), both the cause and the theme correspond to abstract entities, thereby generating environments that differ sharply from non-neological uses. These changes are reflected in the distinct vectors that BERTa assigns to the neological instances, leading HDBSCAN to group them into a standalone cluster.

5.2 Quallar

The verb *quallar* ('to curdle') has a total of 140 occurrences in the CTILC corpus between 2000 and 2021. For these 140 occurrences, 2 clusters have been found, with 60 and 47 occurrences in each, as shown in Figure 2. Of the total recorded contexts, 32 were not grouped into any cluster, as the algorithm identified them as noise —that is, as sporadic uses that did not correspond to the general uses detected.

TABLE 1 Summary of the results obtained for the verb *amanir*

Cluster	Prototypical sentence	Usage examples	DEIEC definition
0	"[. . .] clar de les últimes escarxofes de temporada, moll de rosada dels manats de cebes tendres, capficat de bledes i espinacs, translúcid dels tomàquets d'**amanir** —i amb el vermell— vellut empiocat dels primers micacos, rosa delicat de les camises de les cebes dolces, [. . .]"	"Sovint tu em manes que vagi a comprar un enciam i tomates d'**amanir** a casa de la senyora Rosalia. O una dotzena d'ous."	"Condimentar [hortalisses crues, vianda a punt per a ésser menjada] a base d'oli, vinagre, sal o condiments similars."
		"Ingredients: 4 tomàquets d'**amanir**, 1 kg de patates, 1 ceba, 1 pebrot vermell, 400 g de bacallà esqueixat, 100 g d'olives arbequines, oli d'oliva i sal."	
1	"I en neixen d'altres que també mengen, beuen i dormen i, per no embrutir-se d'avorriment, **amaneixen** la seva vida amb la maledicència, el vodka, les cartes, les trifulgues."	"La fortuna del vencedor no concedeix cap glòria —va dir el cavaller des de terra—, per més que l'**amaneixi** amb paraules vanes! Jo soc el cavaller de Muntalt i ningú no em pot retreure no haver estat pietós ni haver tingut misericòrdia amb molta gent."	None
		"Els telenotícies no s'estan al marge d'aquest clima, tot i que donen una mica més de relleu a la informació sobre la cosa pública, que **amaneixen** amb unes gotes d'informació internacional."	
2	"Si és verge i ecològic, el podeu fer servir per **amanir** en cru qualsevol plat. S'extreu de la polpa fresca i madura del coco."	"Se'n menjava molta, la vianda l'**amania** tota amb suc de cansalada, i per cuinar es feia servir molt greix."	"Condimentar [hortalisses crues, vianda a punt per a ésser menjada] a base d'oli, vinagre, sal o condiments similars."
		"A continuació, escorreu-les i **amaniu**-les amb oli, pebre, alls picats, pinyons i panses, i deixeu-ho macerar durant 2 o 3 hores."	

In the case of the verb *quallar*, the proposed methodology proves to be particularly fruitful. The algorithm identified only two clusters: one corresponding to the non-neological usage—directly associated with a consolidated lexicographical sense—and another comprising the neological uses, which lack a record in the reference dictionary. The dictionarized usage was concentrated in cluster 2, with a total of 47 occurrences. On the other hand, cluster 0, which groups 60 occurrences, corresponds to the non-dictionarized and, consequently, neological variant.

Unlike the trends observed with the verb *amanir*, the contextual variations in *quallar* manifest prominently in the argument structure and the number of arguments linked to the verbal base, as shown in Table 2. In the dictionarized occurrences, the verb behaves as a mono-argumental

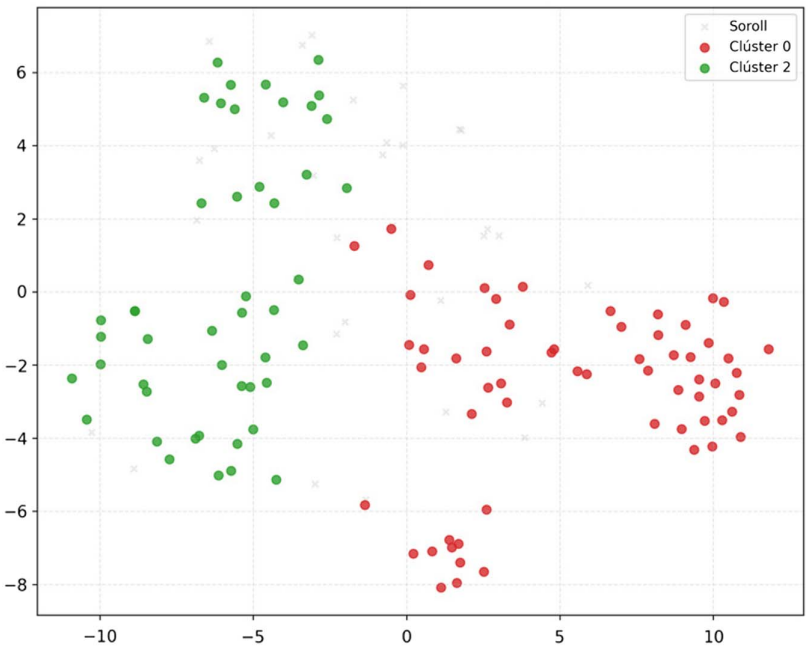


FIGURE 2 Clusters generated for the verb *quallar*

TABLE 2 Summary of the results obtained for the verb *quallar*

Cluster	Prototypical sentence	Usage examples	DEIEC definition
0	"A finals dels 60, el moviment s'havia estès arreu del món, però a Espanya no arribà a **quallar** del tot, la dictadura ens aïllava. Va ser sobretot a Eivissa on se n'establiren algunes comunes."	"Les causes d'aquest fenomen van ser variades i, entre elles, cal esmentar la medicina higienista de l'època, que començava a **quallar** entre les classes altes més o menys cultes i els metges, i que aviat va arribar a tothom."	None
		"En resum, amb excepcions la natura de les quals es perd en el temps, a les grans planes no **ha quallat** un poder polític prou autònom i poderós."	
2	"Finalment, es tria la figuera amb variacions en la preparació, atès que en alguna ocasió **ha quallat** la llet. Per a reduir el temps d'espera i aconseguir resultats més precisos i regulars, s'opta per mantenir la llet amb el quall a una temperatura ambient."	"Aboquem al damunt els ous batuts i deixem que **qualli** . Donem la volta a la truita i la deixem coure un parell de minuts més per l'altre costat."	"[Una substància, un procés] fer passar d'estat líquid a semisòlid [un líquid albuminós, una matèria líquida, un col·loide]"
		"Només haureu de **quallar** la llet, i ho aconseguireu posant-hi una mica de suc de llimona i un parell de minuts al microones."	

unit that only requires a theme, whose semantic implicatures are strictly restricted to the food domain. In contrast, the neological uses of this verb exhibit a bi-argumental structure consisting of a cause and a theme, where both arguments correspond to abstract entities. In this scenario, the transition from a highly specific contextual environment restricted to the culinary domain toward abstract conceptual frameworks led BERTa to assign sharply differentiated vectors for each type of usage, facilitating their subsequent automatic and precise clustering via the HDBSCAN algorithm.

6 Discussion and conclusions

The results from *amanir* and *quallar* showcase the efficiency of applying a density-based clustering algorithm (HDBSCAN) to contextualized BERTa embeddings to isolate semantic neologisms without requiring predefined annotation. As theorized by Bastuji (1974) and Guilbert (1973), meaning shifts reorganize syntactic-semantic combinatorics. For *amanir*, the framework successfully separated literal culinary uses (Clusters 0 and 2) from abstract extensions (Cluster 1). This transition from a human agent and edible theme to abstract causative arguments altered BERTa's bidirectional attention weights, translating linguistic variation into a distinct vector distribution.

Similarly, for *quallar*, the algorithm produced a clean binary split: cluster 2 consolidated the dictionarized, mono-argumental culinary usage ("quallar la llet"), while cluster 0 gathered 60 occurrences of the neological, abstract variant ("quallar un poder polític"). This successfully captured a significant shift in the argument structure and semantic implicatures of the constituents. While traditional approaches relying on frequency (Janssen, 2009) or static embeddings (Torres-Rivera et al., 2020) fail to identify concurrent semantic layers within the same temporal corpus, this dynamic pipeline maps polysemic nuances effectively. Thus, this methodology became an automatic filtering tool that accelerates systematic vocabulary screening (Renau, 2023).

This study has successfully developed and evaluated a hybrid methodological proposal for the automatic detection of verbal semantic neology in Catalan. By combining the contextual modeling of BERTa combined with the density-based classification of HDBSCAN, the framework showcases that distributional semantics can be operationalized efficiently to identify meaning shifts. Crucially, the methodology fulfills its primary objective: it does not attempt to replace the lexicographer with fully automated decisions, but rather aims to augment human expertise. The abstraction of a "prototypical sentence" for each cluster via centroid Euclidean proximity effectively condenses massive textual corpora into a few highly representative, contrastive examples, allowing lexicographers to validate candidate neologisms rapidly against reference baselines like the DEIEC.

Future lines of research should expand this pipeline to other lexical categories, such as nouns and adjectives, where semantic extensions may follow different distributional patterns than those observed in verbal predicates. Additionally, implementing this framework within a strict diachronic timeline could allow researchers to track the exact lifecycle of a semantic change: observing the precise chronological threshold at which an innovative usage transitions from being treated as isolated algorithmic "noise" into forming a dense, stable neological cluster. Finally, this approach establishes a scalable and responsible path forward for lexicography in the digital era, ensuring that resources for languages like Catalan can be enriched through the balanced integration of machine learning and linguistic expertise.

References

Adelstein, A. (2017). Neología semántica y voces especializadas en la conformación del léxico común del español de la Argentina. In O. Conde (Ed.), *Los argots hispánicos. Convergencias y divergencias en las hablas populares hispanoamericanas* (pp. 103–122). Universidad Nacional de Lanús.

Alvar Ezquerra, M. (1998). Palabras nuevas en los periódicos de hoy. In Caja de Burgos (Ed.), *La lengua española a finales del milenio* (pp. 13–44). Caja de Burgos.

Armengol-Estape, J., Pio Carrino, C., Rodriguez-Penagos, C., Gibert Bonet, O., Asrmentano-Oller, C., Gonzalez-Agirre, A., Melero, M. & Villegas, M. (2021). Are multilingual models the best choice for moderately under-resourced languages? A comprehensive assessment for Catalan. In C. Zong, F. Xia, W. Li & R. Navigli (Eds.), *Findings of the Association for Computational Linguistics: ACL 2021* (pp. 4933–4946). Association for Computational Linguistics.

Baroni, M. & Lenci, A. (2010). Distributional memory: a general framework for corpus-based semantics. *Computational Linguistics*, 36(4), 673–721.

Bastuji, J. (1974). Aspects de la néologie sémantique. *Languages*, 36, 6–19.

Bernal, E. (2022). Tendencias neológicas del español peninsular (2015-2019). In E. Bernal, J. Freixa, & S. Torner (Eds.), *La neología del español: del uso al diccionario* (pp. 73–103). Iberoamericana; Vervuert.

Boleda, G. (2020). Distributional Semantics and Linguistic Theory. *Annual Review of Linguistics*, 6, 213–234.

Bréal, M. (1897). *Essai de sémantique*. Hachette.

Cabré, M. T. (2002). La neologia avui: el naixement d'una disciplina. In M. T. Cabré, J. Freixa & E. Solé (Eds.), *Lèxic i neologia* (pp. 29–42). Institut Universitari de Lingüística Aplicada - Universitat Pompeu Fabra.

Cabré, M. T. (2009). La classificació dels neologismes: una tasca complexa. In M. T. Cabré & R. Estopà (Eds.), *Les paraules noves. Criteris per detectar i mesurar els neologismes* (pp. 11–37). Eumo.

Campello, R., Moulavi, D., & Sander, J. (2013). Density-Based Clustering Based on Hierarchical Density Estimates. In J. Pei, V. Tseng, L. Cao, H. Motoda, & G. Xu (Eds.), *Lecture Notes in Computer Science* (pp. 160–172). Springer.

Cartier, E. (2016). Neoveille, système de repérage et de suivi des néologismes en sept langues. *Neologica*, 10, 101–131.

Castro, A., Nazar, R. & Renau, I. (2021). New verbs and dictionaries: method for the automatic detection of neology in Spanish verbs. *International Journal of Lexicography*, 34(3), 382–399.

Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran & T. Solorio (Eds.), *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1(3) (pp. 4171–4186).

Díaz-Hormigo, M. T. (2020). Precisiones para una caracterización de la neología semántica. *ELUA: Estudios de Lingüística*, 34, 73–94.

Erk, K. (2012). Vector space models of word meaning and phrase meaning: a survey. *Language and Linguistics Compass*, 6, 635–653.

Fernández-Sevilla, J. (1982). *Neología y neologismo en español contemporáneo*. Editorial Don Quijote.

Firth, J. R. (1957). *Studies in Linguistic Analysis*. Blackwell.

Freixa, J., Bernal, E., & Cabré, M. T. (Eds.). (2015). *La neologia lèxica catalana*. Institut d'Estudis Catalans.

Gardin, B., Lefèvre, G., Marcellesi, C. & Mortureux, M.-F. (1974). À propos du sentiment néologique. *Langages*, 36, 45–52.

Geeraerts, D., Speelman, D., Heylen, K., Montes, M., De Pascale, S., Franco, K. & Lang, M. (2024). *Lexical Variation and Change. A Distributional Semantic Approach*. Oxford University Press.

Guilbert, L. (1973). Théorie du néologisme. *Cahiers de l'Association internationale des études françaises*, 25, 9–29.

Guilbert, L. (1975). *La créativité lexicale*. Larousse.

Hanks, P. (2013). *Lexical analysis: Norms and exploitations*. The MIT Press.

Harris, Z. (1954). Distributional structure. *Word*, 10, 146–162.

Janssen, M. (2009). Detección de neologismos: una perspectiva computacional. *Debate Terminológico*, 5(5), 68–75.

Kerremans, D., & Prokic, J. (2018). Mining the Web for new words: Semi-automatic neologism identification with the NeoCrawler. *Anglia: Zeitschrift für englische Philologie*, 136(2), 239–268.

Klosa-Kückelhaus, A. & Kernerman, I. (2020). Global viewpoints on lexicography and neologisms: An introduction. *Dictionaries*, 41(1), 1–9.

Laicher, S., Kurtyigit, S., Schlechtweg, D., Kuhn, J. & Schulte im Walde, S. (2021). Explaining and Improving BERT Performance on Lexical Semantic Change Detection. In I.-T. Sorodoc, M. Sushil, E. Takmaz & E. Agirre (Eds.), *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop* (pp. 192–202).

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L. & Stoyanov, V. (2019). RoBERTa: A robustly optimized BERT pretraining approach. *arXiv preprint arXiv:1907.11692*.

Lüdtke, H. (1998). *El cambio lingüístico*. Universitat Autònoma de Barcelona.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In L. Le Cam & J. Neyman (Eds.), *Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability*, 5(1) (pp. 281–297).

Martí, M. A. (2018). Modelos de semántica distribucional. In M. Díaz Ferro, G. Vaamonde, A. Varela Suárez, M. C. Cabeza Pereiro & J. M. García-Miguel Gallego (Eds.), *Actas do XIII Congreso Internacional de Lingüística Xeral* (pp. 16–22). Universidade de Vigo.

Matoré, G. (1952). Le néologisme: naissance et diffusion. *Le français moderne*, 20(2), 87–92.

Mikolov, T., Chen, K., Corrado, G. & Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3787*.

Pennington, J., Socher, R. & Manning, C. (2014). Glove: Global vectors for word representation. In A. Moschitti, B. Pang & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1532–1543).

Peters, M., Neumann, M., Iyyer, M., Gardner, M., Clark, K., Lee, K. & Zettlemoyer, L. (2018). Deep contextualized word representations. In M. Walker, H. Ji & A. Stent (Eds.), *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (pp. 2227–2237).

Kroeger, P. (2004). *Analyzing syntax: A lexical-functional approach*. Cambridge University Press.

Renau, I. (2023). A corpus-based study of semantic neology of the Covid-19 pandemic. *Quaderns de Filologia: Estudis Lingüístics*, 28, 55–76.

Renouf, A. (2014). Neology: from word to register. In L. Vandelanotte, K. Davidse & C. Gentens (Eds.), *Recent Advances in Corpus Linguistics* (pp. 173–205). Brill.

Rey, A. (1976). Le néologisme: un pseudo-concept?. *Cahiers de Lexicologie*, 28, 3–17.

Sánchez Manzanares, M. C. (2009). Procedimientos trópicos en la neología semántica: sistematicidad y creatividad. *Revista de Investigación Lingüística*, 12, 123–146.

Schlechtweg, D., Zamora-Reina, F., Bravo-Marquez, F. & Arefyev, N. (2024). Language Resources and Evaluation. *Language Resources and Evaluation*, 1, 1–25.

Torijano, J. A. (2024). Neología semántica: aspectos favorecedores y obstaculizadores para el cambio lingüístico. *Annali - Sezione romanza*, 66(1), 79–96.

Torres-Rivera, A. & Torres-Moreno, J. M. (2020). Detecting new word meanings: A comparison of word embedding models in Spanish. In P. Bellot & L. Soulier (Eds.), *Conférence en Recherche d'Informations et Applications* (pp. 1–16). ARIA.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A., Kaiser, L. & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems* (pp. 5998–6008).

Contact information

Marta Garcia-Casado
Universitat Pompeu Fabra
marta.garciac@upf.edu

Acknowledgements

The author would like to thank the Institut d'Estudis Catalans for kindly providing the corpus used in this study.

Alexander Geyken & Gregor Middell

TREND WORD DETECTION FOR DWDS, A LEXICAL INFORMATION SYSTEM FOR CONTEMPORARY GERMAN

Abstract This paper addresses the challenge of prioritizing dictionary revisions in corpus-based lexicography, specifically for the Digital Dictionary of German (DWDS), which relies heavily on legacy content. Given the massive data volume in modern corpora, manual selection of relevant words for updating is no longer feasible. We investigate two methods for identifying "trend words"—terms showing an overproportionally high present-day frequency relative to past intervals. The first method applies established linear regression to large monitor corpora, highlighting limitations due to reliance on the chosen time period. To mitigate this, the second method employs a dispersion-oriented keyness metric on a specialized "discourse corpus" of current online text. This metric scores lemmas based on their spread across unique sources, effectively prioritizing words relevant to contemporary discourse. Evaluation of the ranked candidate list confirms that the high-score group contains a significantly higher density of relevant trend words compared to lower-ranked groups. This system allows lexicographers to systematically prioritize legacy entries for revision, such as those that are outdated or missing a current sense, thereby optimizing the lexicographical workflow and ensuring the dictionary reflects current language use.

Keywords Trend word detection; keyness metrics; legacy dictionaries; entry revision; lexicographic process

1 Introduction

Trend words, as employed in this work, are words or phrases exhibiting an overproportionally high frequency of occurrence in the present relative to preceding time intervals. While trend words are primarily established linguistic units used over an extended period, they can also, albeit much less frequently, be neologisms. The automated identification of trending words from large corpora for documentary dictionaries of contemporary language can serve as a strategic mechanism for prioritizing editorial updates. Given the massive data volumes in modern corpora, corpus-based lexicography faces the difficulty of selecting from an overwhelming number of candidates those that possess particular relevance. An automated detection system could significantly optimize this workflow: it allows the editorial team to focus specifically on those lemmas whose current usage frequency suggests a heightened public need for information.

Traditionally, the update cycles within lexicographical projects have frequently been governed by opportunistic triggers: revisions are initiated either in response to specific user feedback regarding perceived shortcomings or when editorial staff discovers outdated or wrong dictionary articles. A more methodical, yet remarkably labor-intensive alternative involves systematic reading programs, wherein researchers evaluate dictionary articles sequentially to pinpoint candidates for revision. Such exhaustive manual reviews are not only time-consuming but also inherently susceptible to oversight, as mere inspection cannot substitute for rigorous corpus-based analysis. In recent years, automated methodologies have emerged to streamline this workflow, such as semantic shift detection techniques designed to reveal unrecorded lexical senses (Schlechtweg et al., 2025). While these modern approaches effectively suggest lexical units for editorial consideration, they do not provide a mechanism for the systematic prioritization of entries based on current public information needs or elevated trend metrics.

In this work we apply our trend identification method to the Digital Dictionary of German (Digitales Wörterbuch der deutschen Sprache, DWDS, www.dwds.de), a comprehensive lexical information system for contemporary German. DWDS contains several large diachronic corpora that are useful for the work at hand. DWDS also has a comprehensive monolingual dictionary that relíes on legacy print dictionaries as a "backbone" as well as newer digitally born entries compiled by the editorial staff (Geyken et al., 2025). Hence DWDS is a typical representative for those dictionary projects that build to a considerable extent on legacy content that continuously needs to be revised.

The remainder of this paper is organized as follows. Section 2 describes related work pertaining to the automatic identification of trend words from large corpora. In section 3 we describe three different corpora as well as the dictionary resources to which the trend identification method is applied (section 4). In section 5, we present and evaluate the results on the basis of the corpora and the trend identification method. In section 6, we discuss the results and present some ideas for future work.

2 Related work

Apart from their utility in structuring lexicographical editorial processes, established platforms like Sketch Engine (Kilgarriff et al., 2014) already provide for the retrieval and statistical evaluation of word frequencies from large linguistic corpora. When coupled with precise temporal metadata for individual documents, frequency distributions over time can be derived and abstracted via linear regression models (Herman & Kovář, 2013). The estimated slopes of linear functions fitted to such time series data then indicate for each word, whether a word's occurrence is generally increasing over the measured time interval (and the word thus a trend candidate), or whether its usage remains stable or is in decline. Additionally, correlation coefficients – Kendall's τ coefficients in case of the cited approach – serve as a measure of similarity for each frequency time series and the associated linear estimate. In this way, estimated trends can be classified in terms of the estimate's accuracy: To what extent is an estimated, linear progression actually representative of its underlying frequency distribution over time, or do outliers and/or seasonal effects result in points of inflection that would be counterindicative of such an assumed, linear trend.

Traditionally, frequencies are measured on the basis of tokens. Newer approaches, on the other hand, use contextual word embeddings (Horn, 2021) for trend analysis not only of word tokens but also for nuanced tracking of different meanings of a lexeme, thus facilitating the study of semantic change of words. Evaluations of these methods immediately bring the dependencies of the expected results on the available data basis into focus (Herman, 2025). Outlier in the time series data, caused by insufficient balance in the underlying text corpora, may be alleviated by robust statistical methods. In case of unclean data, however, such as opportunistically crawled web sources used in previous studies, the analysis led to bad results with high rates of linguistically irrelevant and misidentified trends.

3 Resources

In this section we present three different corpora and discuss their quality as a prerequisite for reliable results. Additionally, we briefly describe the DWDS dictionary used for the evaluation of the trend word detection method.

For trend word detection in the context of the DWDS, we adopt established statistical methods while improving on the quality of the underlying dataset. DWDS draws on a very large corpus

basis, currently comprising over 75 billion tokens. Out of this text collection, two substantial subcollections are employed as monitor corpora and are consequently suitable for serving as the basis for trend detection: a web corpus and a large newspaper corpus. The web corpus ("webmonitor") is the result of focused crawling (Barbaresi, 2019) with a carefully selected seed set of more than 1,000 popular German-language web pages and their corresponding RSS feeds. It covers a large variety of topics. Particular emphasis is put on specialized websites with broader readerships, including finance, IT, lifestyle, pharmaceutical journals, etc. The corpus was created in 2021, is updated daily by harvesting new documents from RSS feeds in the seed set via Trafilatura (Barbaresi, 2021), and by now comprises more than 6 billion tokens. With its daily updates, "webmonitor" is generally suited for observing short-term trends. The newspaper corpus ("newspapers") on the other hand is a collection of major regional and nationwide daily and weekly newspapers licensed by publishing houses. It covers the period from 1946 until today, totalling more than 26 billion tokens. It allows tracking long-term trends spanning multiple years and enables the identification of periodically fluctuating word frequencies that look like short-term trends as components of a seasonal frequency distribution.

Both monitor corpora are already utilized in the ongoing work on the DWDS dictionary. For this study, we augment them with two further resources. Having a bias towards journalistic writing in both monitor corpora, firstly we draw comparisons with query frequencies of chosen words (aka search terms) on Google Search. Such frequency data has been collected by the search engine provider dating back until at least 2004. It is made available via its product "Google Trends" that is accessed for our purposes via an API. Secondly, we substantially extended the RSS feed set underlying the web monitor corpus and conducted a one-off crawl harvesting a snapshot of all web documents discovered via this larger seed set ("discourse corpus"). This snapshot is taken as a proxy for the online discourse in April 2026 and allows to measure the dispersion of words in this discourse as another indicator for their relative relevance. To make the coverage of the "discourse corpus" as broad as possible, we reused a dataset of German online news articles from last year (Kriesch & Losacker, 2025). The dataset of about 50 million web resources was reduced to the web domains containing those resources; the homepages of these domains in turn were used to discover RSS feeds which have not been part of the "webmonitor" seed list yet. 755 additional RSS feeds were thus added to the seed list, bringing it to 1,835 feeds in total. Of those, 1,467 feeds showed activity when harvested at the end of April 2026, delivering 42,911 web resources, of which 26,872 resources were dated as published in the same month and kept for further analysis. Boilerplate content was subsequently removed via a custom implementation of jusText (Pomikálek, 2011), the remaining text was linguistically annotated (Köhler et al., 2025) and indexed by lemma (Nolda, 2023), with recognized named entities removed. The resulting index of frequencies per lemma, web resource, feed source and day of publication contained about 7.9 million entries.

The DWDS Dictionary is a comprehensive monolingual dictionary for contemporary German with currently 270,000 entries. Based on very large text corpora, it compiles information on form (spelling, grammar), meaning (definitions, semantic relations), and usage. The lexicographical work is oriented toward both authentic language use (descriptive) and language norms, the latter primarily in the area of spelling. The DWDS Dictionary draws on the lexical knowledge from the three major print dictionaries of the 20th/21st century German language: the Wörterbuch der deutschen Gegenwartssprache (Dictionary of Contemporary German, WDG), compiled between 1952–1977), which is supplemented by sections from the Großes Wörterbuch der deutschen Sprache (Comprehensive Dictionary of the German Language) (Duden-GWDS, 1999) and the WAHRIG Deutsches Wörterbuch (WAHRIG, 2011/2018). The DWDS dictionary is updated in two ways by the DWDS team: revision of dictionary articles of the legacy dictionaries and the compilation of new dictionary entries for words that were not described in the legacy dictionaries. These words mainly consist of neologisms. In total,

over 40,000 headwords have been updated or completely revised since the beginning of the lexicographical work at DWDS. The current version of the DWDS Dictionary is updated weekly.

4 Method

Utilizing the aforementioned resources, we conducted our study in two phases. In the first phase, we applied established statistical methods for trend calculation on selected trend word candidates chosen for their journalistic or thematic relevance (Word of the Year, key terms from sociopolitical discourse, etc.). In the second phase these methods were complemented by incorporating dispersion-oriented keyness metrics applied to the discourse corpus. This hybrid approach was designed to mitigate the inherent methodological challenges and pitfalls typically associated with longitudinal frequency analysis.

Looking at a first trend word candidate, the merits of the established method based on linear regression become as apparent as some of its challenges for the task at hand. Figure 1 illustrates frequency distributions over time in our two monitor corpora and in Google Trends for the word *Heizungsgesetz* ('Heating Act', a heavily debated law in Germany since 2023).

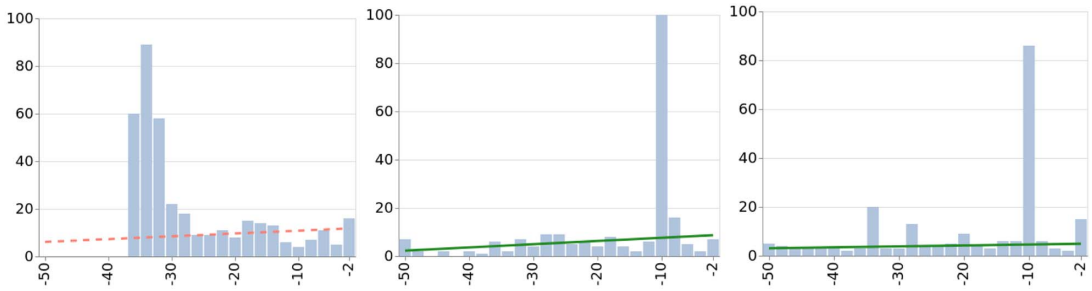


FIGURE 1 Frequency distribution (scaled linearly in [0,100]) over time for *Heizungsgesetz* in (from left to right) "newspapers" (number of months), "webmonitor" and "Google Trends" (both number of weeks)

Referring to German legislation that requires newly installed heating systems to generate at least 65% of their heat from renewable energy, the *Heizungsgesetz* has been a major subject of public debate. Particularly in the first quarter of 2026, when the newly elected German government introduced a revised version of the act, rolling back on some of its highly contested requirements, the term figured prominently in various publication formats. The two histograms on the right in figure 1 conveniently support this observation. Looking at weekly word frequencies in the "webmonitor" corpus as well as in Google Trends in a 50 weeks interval from April 2026 (week 0) to April 2025 (week 50), we observe increasing occurrences of *Heizungsgesetz* over the last 50 weeks, peaking in week 10. The slope of the estimated linear function (superimposed on the histogram) is slightly positive in both cases (0.13 for "webmonitor", 0.04 for "Google Trends"). The same applies to the correlation coefficient (Kendall's τ: 0.352 for "webmonitor", 0.2118 for "Google Trends"). From a purely numerical point of view, *Heizungsgesetz* thus qualifies as a clear trend candidate when looking at its usage over roughly the last year. Moreover, the correlation coefficient passes an arbitrary lower threshold (established as $|\tau| \geq 0.2$ in our cases and depicted by a solid green line), so we can assume that the estimated positive linear progression represents the underlying frequency data with sufficient accuracy. In contrast, when we look at frequencies of the term in monthly intervals as observed in the newspaper corpus, going back again 50 steps, but this time for a longer total period of about 4 years, the numeric indicators suggest a different conclusion. The slope of the linear progression fitted by the Theil-Sen estimator is still positive (0.12), due to the fact that the

term was completely missing from the collected newspapers in the months before initial introduction of the legislation, 3 years ago. But the slightly lower correlation coefficient (0.1606, below the arbitrary 0.2 lower threshold and therefore depicted as a dashed red line) hints to something that is even easier to grasp by the human eye: the term was more prevalent at the introduction of the act than during its revision, hardly qualifying it as a clear trend candidate, at least when looking at its usage beyond the last 50 weeks. More generally, low slope estimates in combination with low correlation coefficients can hint to recurring trends, as can be illustrated by *Spargel* 'asparagus', a seasonal vegetable.

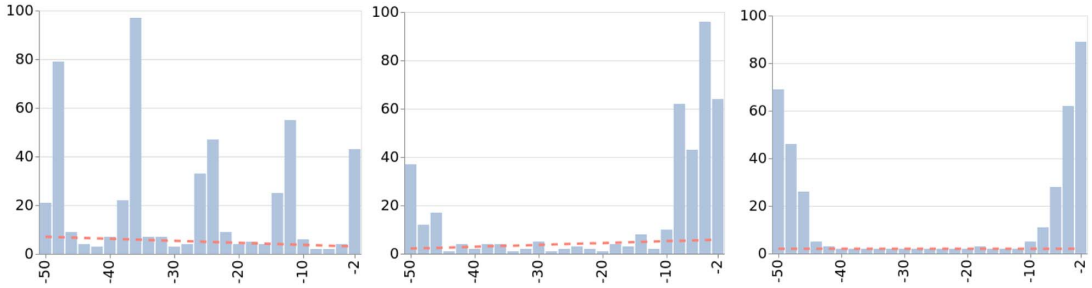


FIGURE 2 Frequency distribution (scaled linearly in [0,100]) over time for *Spargel* in (from left to right) "newspapers" (number of months), "webmonitor" and "Google Trends" (both number of weeks)

The retrospective plots for *Spargel* in figure 2 depict the same time periods and data sources as for *Heizungsgesetz*. Plotted at the end of April, hence, at the beginning of the yearly asparagus season, the week-based plots stretching over roughly a year show expected increases of word frequency at the start and end of their timelines, while the year-based plot exhibits recurring peaks in 12-month intervals. The slopes of all three estimated linear progressions revolve around zero (-0.08, 0.08, 0) while the absolute value of their correlation coefficients is also low (-0.1924, 0.1582, 0.1291). So if the investigated time period covers the interval between recurring trends and we assume a suitable threshold for acceptable correlation coefficients, such effects can be reliably detected. Whether this is instructive for prioritizing lexicographic work, is a different question, as one could argue that independent of the chosen time period under investigation and the general ability to account for recurring patterns, a comprehensive reference dictionary of contemporary German shall provide an up-to-date description of *Spargel* every year, and most certainly in spring.

A third and last example for the application of traditional trend analysis shows that varying time periods under investigation can even yield contradictory results. With the 2026 Iran war and the blockading of the Strait of Hormuz, the accompanying rise of oil prices and resulting fuel shortages are a pervasive topic worldwide. Mentions of the jet fuel *Kerosin* 'kerosine' should be on the rise accordingly.

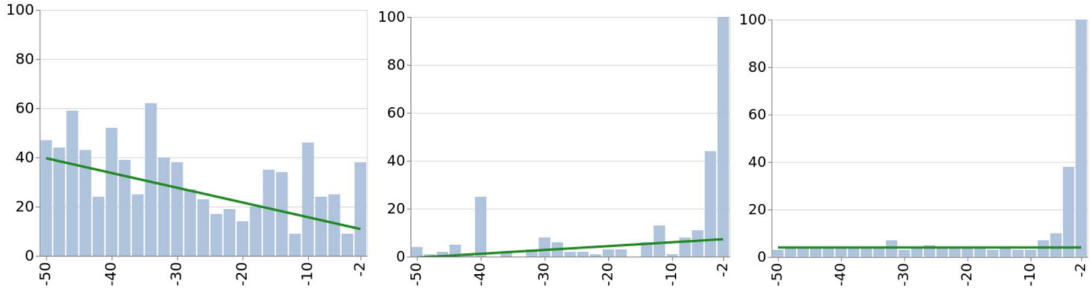


FIGURE 3 Frequency distribution (scaled linearly in [0,100]) over time for *Kerosin* in (from left to right) "news-papers" (number of months), "webmonitor" and "Google Trends" (both number of weeks)

The plots for *Kerosin* in figure 3 do not support this unanimously. Only the "webmonitor" plot in the middle reflects the expected increase in occurrences over the last weeks. An estimated slope of 0.16 and a correlation coefficient of 0.352 would also suffice to identify this word as a trend candidate when taking the perspective of a one-year development. Trying to verify that via the "Google Trends" plot on the right, which depicts the same time period, there is also a visible increase of occurrences over the last weeks; in contrast to the monitor corpus though and because of a constant base level of search queries for *Kerosin*, the estimated slope over all measured weekly intervals amounts to 0 (τ: 0.2934) and thus does not indicate a trend at all. The long-term view on the left suggests a third conclusion. In newspapers over the last 4 years, occurrences of *Kerosin* are actually decreasing in number, with an estimated slope of -0.6 and comparatively solid correlation coefficient of -0.4566.

To summarize, estimating linear trends in a frequency distribution over time allows us to discern trend candidates on a case-by-case basis. The same holds for recurring patterns. Unfortunately, the results depend on the chosen time period. If the comparison between the monitor corpora (news-oriented text collections) versus Google Trends (general interest) for *Kerosin* is a generalizable indicator, they will furthermore depend on the dataset under investigation. For discovering trend candidates automatically, these dependencies pose a significant challenge, in terms of picking suitable corpora, in terms of investigating suitable time periods, as well as in terms of handling ambiguous or even contradictory results.

We mitigated this problem in the second phase of our study by shifting our attention from the discernability of trend candidates to their generality as a keyword in a target corpus which we deem representative of the current German online discourse. Adopting a dispersion-oriented metric of keyness (Sönning, 2024), we design trend word detection as a two-stage process where first of all we filter the discourse corpus for keywords with high dispersion, and only then conduct a deeper analysis to discern true trend candidates among the highly dispersed keywords.

In our "discourse corpus" of April 2026, we ignored the available daily resolution and aggregated frequencies for the whole month per lemma, source (RSS feed) and document (URL). The lemmas were filtered by the following three inequality criteria:

1. A lemma has to be present in at least 3 sources. This lower bound on the dispersion eliminates lemma which are deemed specific to a particular source or ensemble of sources, even if highly frequent in documents of those sources.

2. The number of documents in which a lemma occurs has to be (strictly) greater than the number of sources. We eliminate lemmas this way, which on average are mentioned only once per source.

3. The total frequency of a lemma has to be at least 3 times greater than the number of documents it appears in. This lower bound ensures the topicality of lemmas with regard to the analyzed documents.

Each lemma satisfying the criteria is attributed a numerical score: $s = \frac{freq(feeds)}{freq(\mathrm{doc}) + \sqrt{freq(tokens)}}$.

Lexical units in the "discourse corpus" are assigned a higher score if their dispersion across diverse feed sources is high, whereas the metric simultaneously penalizes for high document frequency and raw token counts.

5 Results

In Section 4, three examples (*Heizungsgesetz* 'heating law', *Spargel* 'asparagus' and *Kerosin* 'kerosine') were used to show the challenges in applying trend recognition methods based on linear regression: trend candidates are recognized, but only if the appropriate time interval has been determined beforehand. In lexicographical practice, however, a procedure is needed that determines trend words at a certain point in time, independent of the setting of the time interval.

Therefore, we employed a dispersion-oriented metric and introduced a score together with some filter criteria, as described in section 4, phase 2. On this basis, the best-rated lemmas were extracted for the "discourse corpus" and ordered according to the score value. Named Entities were filtered out beforehand, as we are only interested in lexical units. The score itself rates function words, but also high-frequency everyday words, lower, because they occur in almost all documents and therefore the quotient of the number of feeds and the number of documents is small. In contrast, a high score is given to words that are typical for a discourse but are not present everywhere at the same time. The result is presented based on a list of 500 words ordered by score. The top 10 words are displayed in Table 1.

TABLE 1 Top 10 lemmas

rank	lemma	class	discourse	#src	#doc	#tok	score	entry status
1	Kerosin	ts	Iran	60	62	322	75.05 %	legacy
2	radioaktiv	ts	memorial	62	67	260	74.59 %	legacy
3	kurdisch	ts	Iran	27	28	103	70.78 %	revised
4	Welthandel	ts	Iran	35	39	123	69.87 %	legacy
5	Social-Media-Verbot	ts	internet	36	39	176	68.88 %	missing
6	Hospiz	tl	politics	20	21	65	68.82 %	revised
7	Revolutionsgarde	ts	Iran	44	51	171	68.7 %	legacy
8	Aktenzeichen	n	TV	42	49	148	68.67 %	n.a.
9	Nest	tc	nature	54	65	196	68.35 %	revised
10	Minijob	tl	politics	25	28	85	67.17 %	revised

The data in the "class" and "discourse" columns are the result of manual analysis. In the "class" column, we distinguish three trend types and a residual class:

- ts = short-term trend (e.g. *Kerosin*)
- tl = longer-term trending word (e.g. *Minijob*)
- tc = cyclical trend (e.g. *Nest*)
- n.a = no trend (e.g. *Aktenzeichen*)

In the "discourse" column, we assign the lemmas to a trend event. In Table 1, four lemmas are assigned to the Iran conflict, either because they are frequently mentioned protagonists in the conflict, e.g. *Revolutionsgarden* 'revolutionary guards' or *kurdisch* 'kurdish [troops]' or because the consequences of the conflict on *Welthandel* 'world trade' or a shortage of *Kerosin* are feared. The word *radioaktiv* 'radioactive' refers to a commemoration day: on April 28, 2026, the Chernobyl reactor accident has marked its 40th anniversary. Longer trends correspond to

the politics discourse type in Table 1: the two lemmas *Hospiz* 'hospice' and *Minijob* 'marginal employment' have been discussed for a longer time with varying intensity. The word *Nest* 'nest', on the other hand, is a typical cyclical trend term related to nature in the spring season. Finally, Table 1 contains a false lemmatization: the word *Aktenzeichen* 'file card' stands for the name of a popular TV show (*Aktenzeichen XY*). The columns #src, #doc, #tok and score contain the values from the "discourse corpus" (cf. section 3) and the score was calculated according to the formula in section 4. The "entry status" column records whether the article originates unchanged from one of the printed dictionaries ("legacy"), whether it has already been revised ("revised"), or whether the entry is missing ("missing"). In the overall list of 500 lemmas, we annotated two subsets: the set of all lemmas with a rank of 1-50 ("R1-50"), and the list of all words with a rank of 201-250 ("R201-250"), cf. Table 2.

TABLE 2 Trend word classification

class	R1-50	R201-250
ts	17	7
tl	18	17
tc	6	2
n	9	24

To evaluate annotation reliability, two experienced DWDS lexicographers independently annotated subsets "R1-50" and "R201-250". The four-category system (ts, tl, tc, nt) yielded substantial agreement (85.0% raw agreement; Cohen's Kappa = 0.78). A confusion matrix analysis indicated that under-specified guidelines caused discrepancies primarily between short-term (ts) and long-term (tl) trend words. Consequently, conflating these two categories improves raw agreement to 91.0% and increases Kappa to 0.84.

The results show above all that the number of lemmas corresponding to short-term trends (ts) is significantly higher in "R1-50" than in "R201-250". In contrast, there are hardly any differences for longer-term trends (tl). For cyclical trend words, the number in "R1-50" is slightly higher than that in "R201-250". In contrast, the number of words that are not trend-relevant and therefore not to be prioritized for review is significantly lower in the "R1-50" list than in "R201-250". Overall, this shows that the score fulfills its function as a measure of the quality of a trend word: the density of trend words in the list with a higher score ("R1-50") is significantly higher than in the list with a lower score ("R201-250"). To evaluate the list's usefulness for revision prioritization, we analyzed the ts and tl classes from Table 2; the resulting data are presented in Table 3.

TABLE 3 Prioritization of entry revision

class	entry state	R1-50	R201-250
ts	legacy	10	3
ts	missing	2	4
ts	revised	5	0
tl	legacy	0	1
tl	missing	8	9
tl	revised	10	7

Obviously short-term trend words that are missing in the dictionary are natural candidates for a dictionary update. Entries that have only recently been revised, on the other hand, are not a

priority for revision. Particularly interesting for prioritization are short-term trend words (ts) that originate from the legacy print dictionaries and have not yet been revised (legacy), as these entries in the DWDS dictionary are between 25 and 75 years old (cf. section 3) and therefore represent special candidates for revision, either because these words have undergone a change in meaning or at least require an update of the definition. A good example for the first case is the word *Kerosin*. The word is defined in the legacy dictionary (WDG, 1969) as 'durch Destillation von Erdöl gewonnener Treibstoff für Dieselmotoren' (en: 'fuel obtained by the distillation of crude oil for diesel engines'), which obviously is outdated since it omits the most common field of use, namely as jet engine fuel. An example for the second case is the compound noun *Welthandel*. Within the sociopolitical discourse surrounding the Iran conflict, the term emerges as a significant trend candidate, particularly as the potential blockade of the Strait of Hormuz is perceived as a critical vulnerability for global commercial stability. A query of *Welthandel* in the current DWDS system retrieves a legacy entry defined as 'der internationale Handel im Weltmaßstab' (international trade on a global scale; WDG, 1977). While factually accurate, this description lacks the specificity required for a contemporary lexical information system; it necessitates a modern revision to reflect a complex economic reality characterized by the cross-border exchange of goods, services, and digital assets.

6 Discussion and further work

The empirical findings presented in the preceding section indicate that utilizing a dispersion-oriented metric within the discourse corpus yields effective candidates for trend identification. A comparative analysis involving the top 500 lexical units, harvested from the "discourse corpus" and sorted according to the established score function, demonstrates that the "R1-50" subset contains a significantly higher concentration of short-term trend words than the lower-ranked "R201-250" group. Furthermore, the overall density of relevant trend candidates exhibits a marked increase in the high-score list: while only nine entries in the top 50 were classified as non-trending, this number rises to twenty-four for the "R201-250" interval. The operational integration of this system into the daily lexicographical workflow is remarkably efficient. By generating candidate lists on a monthly or weekly basis and enriching them with metadata—identifying missing entries, recently updated articles, or legacy content—we provide a robust foundation for editorial review. Lexicographers can then manually evaluate these lists, systematically prioritizing for revision those legacy dictionary entries whose contemporary relevance suggests a clear need for a revision of semantic structure or for update(s) or definition(s).

While these initial results are promising, a deeper comprehension of the discourse corpus's thematic stratification is necessary. To address this, we intend to implement topic modeling techniques in future research iterations. Subsequent investigations must also establish a methodological distinction between lexical units signaling genuine linguistic change and ephemeral terms surfacing solely in response to isolated media events. Platforms like "Google Trends" vividly illustrate how specific words achieve temporary popularity through high query volumes linked to singular occurrences. Within our detection framework, such short-lived phenomena are naturally penalized; the scoring mechanism's design effectively lowers the rank of terms with limited distribution across unique RSS feed sources.

Beyond its utility as a strategic prioritization tool, we envisage featuring the most significant weekly trend words on the DWDS website, with direct links to their corresponding dictionary articles and corpus citations. By making the contemporary "trendiness" of specific lexical units visible, we aim to render linguistic dynamics transparent, allowing users to perceive the shifting status of words as a direct reflection of contemporary history.

References

Barbaresi, A. (2019). The vast and the focused: On the need for domain-focused web corpora.In P. Bański, A. Barbaresi, H. Biber, E. Breiteneder, S. Clematide, M. Kupietz, H. Lüngen & C. Iliadi (Eds.), *Proceedings of CMLC-7* (pp. 29–32). Leibniz-Institut für Deutsche Sprache. https://doi.org/10.14618/ids-pub-9025

Barbaresi, A. (2021). Trafilatura: A web scraping library and command-line tool for text discovery and extraction. In H. Ji, J. C. Park & R. Xia (Eds.), *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: System Demonstrations* (pp. 122–131). Association for Computational Linguistics.

DWDS corpora: webmonitor: https://www.dwds.de/d/korpora/webmonitor; newspapers: https://www.dwds.de/d/korpora/zeitungenxl

Geyken, A., Lemnitzer, L., Solf, M., & Wiegand, F. (2025). Umbau auf hoher See – Bericht aus dem Inneren des DWDS. In W. Blanck, R. H. Gouws, & A. Lobenstein-Reichmann (Eds.), *Lexikographisch-grammatische Perspektiven – Tradition, Veränderung und Vielfalt in Lexikographie und Wörterbuchforschung, Stefan J. Schierholz zum 70. Geburtstag* (pp. 187–201). De Gruyter. https://doi.org/10.1515/9783111676531

Herman, O. (2025). Automatic detection of word sense shift from corpus data. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga Š. Arhar Holdt, T. Munda & A. Salgado (Eds.), *Electronic Lexicography in the 21st Century (eLex 2025): Intelligent Lexicography. Proceedings of the eLex 2025 conference* (pp. 237–251). Lexical Computing CZ s.r.o.

Herman, O., & Kovář, V. (2013). Methods for Detection of Word Usage over Time. In A. Horak & P. Rychly (Eds.), *Seventh Workshop on Recent Advances in Slavonic Natural Language Processing. (RASLAN 2013)* (pp. 79-86). Masaryk University Faculty of Informatics.

Köhler, L., Knaebel, R., & Middell, G. (2025). *German spaCy models trained on German UD-HDT and a collection of German NER datasets* [Computer software]. Zenodo. https://doi.org/10.5281/zenodo.17566592

Kriesch, L., & Losacker, S. (2025). A geolocated dataset of German news articles. *Scientific Data*, *12*(1), 1128. https://doi.org/10.1038/s41597-025-05422-w

Nolda, A. (2023). *DWDSmor: A toolbox for morphological analysis and generation in German, based on the DWDS lexicon and an SMOR-style grammar* [presentation at eLex 2023]. https://elex.link/elex2023/wp-content/uploads/10_00-DWDSmor-A-toolbox-for-morphological-analysis-and-generation-in-Germanbased-on-the-DWDS-lexicon-and-an-SMOR-style-grammar.pdf

Pomikálek, J. (2011). *Removing boilerplate and duplicate content from web corpora* [Doctoral dissertation, Masaryk University]. https://theses.cz/id/nqo9nn/?lang=en

Schlechtweg, D., Sköldberg, E., Virk, M. S., White, J., & Hengchen, S. (2025). Automatic non-recorded sense detection for Swedish through word sense induction with fine-tuned word-in-context models. In I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga Š. Arhar Holdt, T. Munda & A. Salgado (Eds.), *Electronic Lexicography in the 21st Century (eLex 2025): Intelligent Lexicography. Proceedings of the eLex 2025 conference* (pp.160-174). Lexical Computing CZ s.r.o.

Sönning, L. (2024). Evaluation of keyness metrics: Performance and reliability. *Corpus Linguistics and Linguistic Theory*, *20*(2), 263–288. https://doi.org/10.1515/cllt-2022-0116

Contact information

Alexander Geyken
Berlin-Brandenburg Academy of Sciences and the Humanities
geyken@bbaw.de

Gregor Middell
Berlin-Brandenburg Academy of Sciences and the Humanities
middell@bbaw.de

Ondřej Herman

SENSE CLASSIFICATION FOR WORD SKETCHES

Abstract The Word Sketch is a core tool for modern corpus-based lexicography, providing a concise summary of a word's grammatical and collocational behavior. A longstanding limitation of standard Word Sketches is the lack of polysemy awareness: when a lexicographer examines a highly polysemous word such as *bank*, the resulting collocate lists mix financial terms like *account* and *deposit* with geographical terms like *river* and *sand*. In this paper, we present and evaluate a fully automated method for enriching Word Sketches with word sense information. Using the Adaptive Skip-gram model, we perform unsupervised word sense induction and project the induced senses onto the type-level collocate lists of a Word Sketch, so that collocates with sufficiently confident predictions are associated with the sense they most strongly represent. On a manually annotated test set of 23 polysemous English words, our method achieves a Precision of 0.64 and Recall of 0.77 (F1 $\approx$ 0.70). The method requires no manual training data and is already deployed in Sketch Engine for eight languages. We describe the user interface, which allows lexicographers to filter Word Sketches by sense, and discuss implications for dictionary-writing workflows.

Keywords Word sense induction; word sketch; polysemy; corpus lexicography; sketch engine

1 Introduction

The Word Sketch (Kilgarriff et al., 2014) is a widely used tool in corpus-based lexicography. For any given headword, it summarizes the word's combinatorial behavior, organizing collocates into grammatical relations such as modifiers, objects, and subjects. This allows lexicographers to quickly understand the usage patterns of a word and select appropriate collocates, examples, and usage notes for a dictionary entry.

However, an important shortcoming of the standard Word Sketch is its lack of treatment of polysemy. The collocate lists are computed over lemmas, so grammatical relations conflate the different senses of the headword. In the modifier relation of the noun *bank*, for instance, *commercial, central*, and *investment* sit alongside *steep, grassy*, and *opposite*, with no indication that they describe two unrelated meanings. The lexicographer is left to mentally disambiguate the senses, scanning long collocate lists to identify which entries belong together. For words with many senses, this is a time-consuming and error-prone process.

The value of addressing this limitation became apparent during the development of a semi-automatic Tagalog–English–Korean dictionary (Baisa et al., 2019), in which an earlier, experimental word sense induction method was used to assist annotators in grouping collocates by sense. The experience suggested that even imperfect sense clustering can assist lexicographic work, motivating us to develop a more reliable, scalable method.

In this paper, we present a method for enriching Word Sketches with automatically induced sense information. The approach uses the Adaptive Skip-gram model (Bartunov et al., 2016), an unsupervised word sense induction method that requires no manually labeled training data and is therefore applicable to any language with a sufficiently large corpus. We project the model's token-level sense predictions onto the type-level collocate lists of a Word Sketch and color-code each collocate by its dominant sense. We evaluate the result on a manually annotated test set of 23 polysemous English words, where it outperforms an earlier clustering-based approach (Herman et al., 2019).

2 Background

2.1 Word Sketches

A Word Sketch (Kilgarriff et al., 2014) is an automatically generated summary of a word's grammatical and collocational behavior in a given corpus. It is computed by the Sketch Engine corpus management system (Jakubíček et al., 2013) using a set of pattern-matching rules called a sketch grammar, which identify collocations and classify them into grammatical relations. The output is a set of triples of the form (*headword, relation, collocate*), each accompanied by an association score that reflects the statistical salience of the collocation.

For example, the Word Sketch for the noun *bank* may contain triples such as (*bank*, modifier, *central*), (*bank*, object_of, *rob*), or (*bank*, modifier, *river*). The collocates within each grammatical relation are ranked by their association score, so that the most characteristic collocates appear at the top.

Word Sketches are well established in practical lexicography: they have been used in the creation of numerous dictionaries and form a standard component of the dictionary-writing workflow in Sketch Engine.

2.2 Word Sense Induction

Word Sense Induction (WSI) is the task of automatically discovering the different senses of a word from unlabeled text. It differs from Word Sense Disambiguation (WSD), which assigns occurrences of a word to senses drawn from a fixed inventory such as WordNet. WSI is, in principle, more flexible: it does not require a predefined sense inventory and can discover senses that are not listed in any existing dictionary.

WSI methods build on the distributional hypothesis (Harris, 1954): words that occur in similar contexts tend to have similar meanings. The corollary is that different meanings of the same word tend to occur in different contexts, which makes it possible to identify word senses by clustering contextual patterns.

In this paper, we use *sense* to mean a cluster of usages occurring in related contexts. The sense induced by the WSI algorithm does not map directly to an entry in a fixed dictionary inventory. This approach treats word senses as abstractions over specific usage patterns (Kilgarriff, 1997).

A relevant observation for our work is the "one sense per collocation" hypothesis (Yarowsky, 1993), which states that a given collocation (e.g., *bank account*) very often corresponds to a single sense of the ambiguous word. If a collocate strongly prefers one sense of the headword, we can use the sense distribution observed in concordance lines to assign the collocate to that sense.

The WSI task has been the subject of several shared tasks in the SemEval evaluation campaigns, including SemEval-2007 Task 02 (Agirre & Soroa, 2007), SemEval-2010 Task 14 (Manandhar et al., 2010), and SemEval-2013 Task 13 (Jurgens & Klapaftis, 2013). These campaigns established standard evaluation methodologies and demonstrated the difficulty of the task, with the best systems achieving moderate agreement with gold-standard sense inventories.

An earlier approach to WSI for Word Sketches (Herman et al., 2019) used spectral clustering and HDBSCAN applied to averaged fastText word embeddings. While that method produced

usable sense clusters for many words, results were sensitive to parameter settings and often left a large proportion of collocates unclassified. Because of these limitations, the present work instead adopts the Adaptive Skip-gram model described in this paper.

3 Methodology

3.1 The Adaptive Skip-gram Model

The Adaptive Skip-gram model (Bartunov et al., 2016) is an extension of the Skip-gram word embedding model. While standard Skip-gram assigns a single vector to each word in the vocabulary, Adaptive Skip-gram is capable of learning multiple vectors for a single word, where each vector approximates a recurring contextual usage pattern.

The key advantage of Adaptive Skip-gram over other multi-sense embedding methods is that the number of senses per word is not fixed in advance. Instead, the model uses a non-parametric Bayesian framework to automatically determine how many senses each word needs based on the training data. A granularity parameter, denoted α (alpha), controls the tendency of the model to split words into multiple senses. A low α produces fewer, coarser senses; a high α encourages finer distinctions.

To handle the large corpora used in Sketch Engine (often exceeding one billion tokens), we re-implemented the Adaptive Skip-gram training algorithm in the Rust programming language, which allows us to train models on corpora of this scale within practical time limits. A fuller technical account of the implementation and its use for sense-shift detection is given by Herman (2026).

3.2 Projecting Senses onto Word Sketches

The method we use for projecting token-level WSI predictions onto the type-level Word Sketch collocate lists follows these steps:

1. **Model Training.** An Adaptive Skip-gram model is trained on the target corpus. The model learns multiple usage prototypes for polysemous words and a single vector for monosemous words. A single model is trained per language and applied to all corpora in that language, whether preloaded or user-uploaded.
2. **Instance Retrieval.** For a specific Word Sketch triple (*headword, relation, collocate*), we retrieve all positions in which this particular collocation occurs.
3. **Sense Prediction.** For each retrieved concordance line, the Adaptive Skip-gram model predicts the probability distribution over the senses of the headword given the surrounding context, so each concordance line generates a vector of sense probabilities.
4. **Aggregation and Thresholding.** The sense probability distributions are averaged across all concordance lines for a given collocate. This provides a single sense distribution for the collocation. Based on the "one sense per collocation" hypothesis, we expect most stable collocations to exhibit a strong preference for a single sense. If the averaged probability for a specific sense exceeds a predefined threshold, the collocate is assigned to that sense. If no sense crosses the threshold, or if the collocation is used in divergent contexts, the collocate remains unassigned ("No sense" is shown in the user interface).

To illustrate, consider the collocate *commercial* in the modifier relation for the noun *bank*. The model retrieves all corpus positions where *commercial* acts as a modifier for *bank*, predicts the distribution of senses of *bank* for each context, and averages them. If the result assigns 85% probability to a "financial institution" sense and 15% to a "river bank" sense, the financial sense exceeds the chosen threshold, so the collocate *commercial* is assigned to the financial sense of *bank*.

3.3 Sense Labeling

Because the method is unsupervised, the induced senses do not come with human-readable labels such as "financial institution" or "river bank." Instead, we represent each sense by its nearest semantic neighbors in the Adaptive Skip-gram vector space. For the noun *bank*, Sense 1 might be represented by the words *institution, fund, lending, lender*, while Sense 2 is represented by *edge, along, river, side*. These labels are generated automatically and serve as hints that help the user identify what each sense corresponds to. While not always perfectly transparent, they are often sufficient for a lexicographer to understand the grouping.

4 User Interface and Visualization

We designed a visualization within Sketch Engine that makes the sense information accessible. Figure 1 shows the Word Sketch for the noun *bank* with the sense classification enabled.

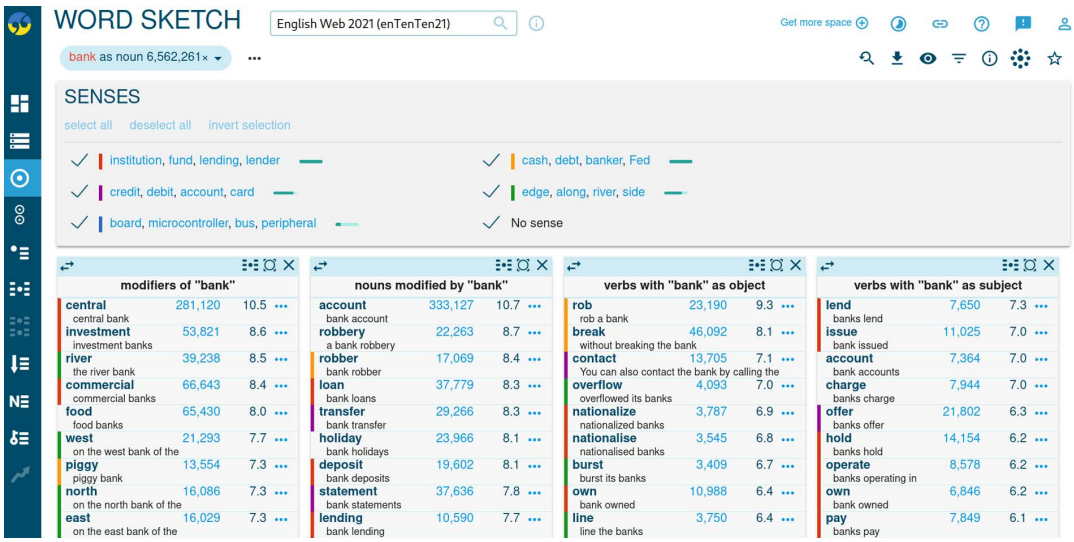


FIGURE 1 Word Sketch for the noun *bank* with sense classification; colored bars indicate the assigned sense for each collocate, and the legend at the top shows automatically generated sense labels

At the top of the Word Sketch view, a "Senses" legend is displayed. Each induced sense is represented by the nearest-neighbor labels described above and is assigned a distinct color. In the case of *bank*, the system identifies five senses; alongside the financial and geographic senses, these include a payment card sense (represented by *credit, debit, account, card*).

In the Word Sketch columns below, each collocate is marked with a colored vertical bar corresponding to its assigned sense, so that the financial and geographic collocates of *bank* can be told apart at a glance. The user can also filter the Word Sketch by sense, hiding all collocates that do not belong to a selected sense.

5 Evaluation

5.1 Test Set

To evaluate the quality of the sense classification, we created a manually annotated test set of 23 known-polysemous English words, including nouns such as *bank, band, chip*, and *crane*, as well as two verbs, *fire* and *lie*. For each word, a single annotator inspected the top 150 collocates across all grammatical relations and grouped them by sense based on their corpus contexts, without consulting a fixed dictionary sense inventory. Only collocates that could be assigned to a specific sense with high confidence were annotated; ambiguous cases were excluded from the evaluation. The number of gold senses per word ranges from 2 (for *bark, beam, club, file, lie*) to 7 (for *base*), with a mean of 3.5 senses.

5.2 Metrics

We report the following evaluation metrics, defined in terms of the relationship between the gold-standard senses and the clusters produced by the system. **Gold Senses** is the number of senses we annotated for a word. **Found Senses** is the number of those senses that at least one cluster can be associated with, even if that cluster also covers other senses. **Pure Senses** are the Found Senses whose associated clusters contain no entries mapping to a different annotated sense. From the cluster side, **Mixed Clusters** are clusters that map to more than one annotated sense, while **Unique Clusters** map to at most one annotated sense. Together they account for every cluster the system produces.

Precision is the number of Pure Senses divided by the total number of clusters (Mixed Clusters + Unique Clusters). High precision means that the clusters the system produces tend to be sense-coherent.

Recall is the number of Found Senses divided by the number of Gold Senses. High recall means that the system identifies most of the senses present in the test set.

We also report **Pure Recall**, the number of Pure Senses divided by the number of Gold Senses. We summarize the results using the harmonic mean (F1) of Precision and Recall.

5.3 Results

We evaluated the system using the best-performing configuration from a parameter sweep: a model with an embedding dimension of 256, a context window of 14 tokens, and a sense assignment threshold of 0.75. The per-word results are shown in Table 1.

The system achieves a mean Precision of 0.64 and a mean Recall of 0.77, yielding an F1 score of approximately 0.70. Three words — *bark, file*, and *fire* — achieve perfect scores (P = 1.0, R = 1.0), meaning that the system recovers all gold senses with perfectly pure clusters. Several other words also achieve high scores, including *tank* (P = 1.0, R = 0.80), *bat* (P = 1.0, R = 0.67), and *bow* (P = 1.0, R = 0.60).

The most challenging words are *base* (P = 0.17, with 7 gold senses and many mixed clusters), *jumper* (P = 0.00, all collocates assigned to a single mixed cluster), and *change* (P = 1.0 but R = 0.33, only one of three senses is found).

TABLE 1 Evaluation results for 23 polysemous English words. P = Precision, R = Recall, PR = Pure Recall, the Mean row averages across the evaluated words

Word	Gold Senses	Found Senses	Pure Senses	P	R	PR
band-n	6	5	5	0.62	0.83	0.83
bank-n	4	2	2	1.00	0.50	0.50
bark-n	2	2	2	1.00	1.00	1.00
base-n	7	6	1	0.17	0.86	0.14
bat-n	3	2	2	1.00	0.67	0.67
beam-n	2	2	2	0.67	1.00	1.00
board-n	5	3	3	0.50	0.60	0.60
bow-n	5	3	3	1.00	0.60	0.60
change-n	3	1	1	1.00	0.33	0.33
chip-n	4	2	2	0.40	0.50	0.50
club-n	2	1	1	0.33	0.50	0.50
crane-n	3	3	1	0.33	1.00	0.33
deck-n	3	2	2	1.00	0.67	0.67
file-n	2	2	2	1.00	1.00	1.00
fire-v	3	3	3	1.00	1.00	1.00
iron-n	3	3	3	0.43	1.00	1.00
jumper-n	3	2	0	0.00	0.67	0.00
lie-v	2	2	2	0.40	1.00	1.00
party-n	4	3	3	0.60	0.75	0.75
pole-n	3	2	2	0.67	0.67	0.67
spring-n	4	3	3	0.43	0.75	0.75
tank-n	5	4	4	1.00	0.80	0.80
tear-n	3	3	1	0.20	1.00	0.33
Mean	**3.52**	**2.65**	**2.17**	**0.64**	**0.77**	**0.65**

5.4 Discussion

An F1 score of approximately 0.70 may seem modest by the standards of many NLP tasks, but it is useful as a pre-filtering step in a manual workflow. Our approach is to give the lexicographer an initial view to refine, instead of an undifferentiated collocate list. Perfect scores are unlikely, because the system groups usage patterns drawn from a limited sample of a language, without any task-specific tuning. The practical question is therefore whether the grouping is coherent enough to reduce manual effort while still allowing the editor to inspect the underlying corpus evidence.

Compared to an earlier approach based on HDBSCAN clustering of averaged fastText embeddings (Herman et al., 2019), the method based on Adaptive Skip-gram performs substantially better. The earlier method achieved a mean Precision of 0.43 and Recall of 0.63 (F1 $\approx$ 0.51) on the same test set. The new method improves Precision by 0.21 and Recall by 0.14, raising F1 by 0.19. Moreover, we observed that the Adaptive Skip-gram method is more stable over different training runs and produces fewer unclassified collocates.

Note that WSI is a difficult task even for human annotators. In the ShadowSense evaluation (Herman & Jakubíček, 2024), the average inter-annotator Rand Index obtained for English was 0.82, which is high, but far from perfect.

6 Conclusion and Future Work

We have presented a method for enriching Word Sketches with automatically induced word sense information. The method is based on the Adaptive Skip-gram model, which learns multiple usage prototypes for polysemous words without requiring any manually labeled training data. By projecting sense predictions from concordance lines onto the type-level collocate lists of a Word Sketch, we enable a visualization in which collocates are color-coded by their dominant sense. On a test set of 23 polysemous English words, the method achieves an F1 score of approximately 0.70, representing a substantial improvement over the earlier clustering-based approach.

The feature is already deployed in Sketch Engine for all corpora in eight supported languages: English, Spanish, Italian, French, Ukrainian, Czech, Slovak, and Estonian. It groups collocates based on distributional semantics.

We are currently working on expanding the language coverage to all languages supported by Sketch Engine. We also plan to improve automatic sense labeling to produce more transparent labels, to explore more sophisticated models for sense induction, and to extend the sense awareness to other features within Sketch Engine. The first target is the Trends functionality, which tracks changes in word frequency over time. Combining the per-sense information from Adaptive Skip-gram with diachronic corpus data makes it possible to detect newly appearing word senses. The combination has already been investigated (Herman & Rychlý, 2026). The work that remains is primarily its engineering integration into Sketch Engine's production infrastructure.

References

Agirre, E. & Soroa, A. (2007). SemEval-2007 Task 02: Evaluating word sense induction and discrimination systems. In E. Agirre, L. Márquez i Villodre & R. Wicentowski (Eds.), *Proceedings of the 4th International Workshop on Semantic Evaluations* (pp. 7–12). Association for Computational Linguistics.

Baisa, V., Blahuš, M., Cukr, M., Herman, O., Jakubíček, M., Kovář, V., Medveď, M., Měchura, M., Rychlý, P. & Suchomel, V. (2019). Automating dictionary production: A Tagalog–English–Korean dictionary from scratch. In I. Kosem, T. Zingano Kuhn, M. Correia, J.P. Ferreria, M. Jansen, I. Pereira, J. Kallas, M. Jakubíček, S. Krek & C. Tiberius (Eds.), *Electronic Lexicography in the 21st Century: Proceedings of the eLex 2019 Conference* (pp. 805–818). Lexical Computing CZ, s.r.o.

Bartunov, S., Kondrashkin, D., Osokin, A. & Vetrov, D. (2016). Breaking sticks and ambiguities with adaptive skip-gram. In A. Gretton & C. C. Robert (Eds.), *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics* (pp. 130–138). PMLR.

Harris, Z. S. (1954). Distributional structure. *Word*, 10(2–3), 146–162.

Herman, O. (2026). *Automatic Detection of Word Sense Shift* [Doctoral dissertation, Masaryk University].

Herman, O. & Jakubíček, M. (2024). ShadowSense: A multi-annotated dataset for evaluating word sense induction. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 14763–14769). ELRA.

Herman, O., Jakubíček, M., Rychlý, P. & Kovář, V. (2019). Word sense induction using Word Sketches. In C. Martín-Vide, M. Purver & S. Pollak (Eds.), *Statistical Language and Speech Processing: 7th International Conference, SLSP 2019, Proceedings* (pp. 83–91). Springer.

Herman, O. & Rychlý, P. (2026). Detecting subtle sense shift with polysemy-aware trends. In V. Demberg, K. Inui & L. Marquez (Eds.), *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 2: Short Papers)* (pp. 60–65). Association for Computational Linguistics.

Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P. & Suchomel, V. (2013). The TenTen corpus family. In *7th International Corpus Linguistics Conference* (pp. 125–127).

Jurgens, D. & Klapaftis, I. (2013). SemEval-2013 Task 13: Word sense induction for graded and non-graded senses. In S. Manandhar & D. Yuret (Eds.), *Second Joint Conference on Lexical and Computational Semantics, Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)* (pp. 290–299). Association for Computational Linguistics.

Kilgarriff, A. (1997). I don't believe in word senses. *Computers and the Humanities*, 31(2), 91–113.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, 1(1), 7–36.

Manandhar, S., Klapaftis, I. P., Dligach, D. & Pradhan, S. S. (2010). SemEval-2010 Task 14: Word sense induction & disambiguation. In K. Erk & C. Strapparava (Eds.), *Proceedings of the 5th International Workshop on Semantic Evaluation* (pp. 63–68). Association for Computational Linguistics.

Yarowsky, D. (1993). One sense per collocation. In M. Kaufmann (Ed.), *Human Language Technology: Proceedings of a Workshop Held at Plainsboro* (pp. 266–271).

Contact information

Ondřej Herman
Lexical Computing, Brno, Czech Republic
ondrej.herman@sketchengine.eu

Dominika Kováříková & Laura A. Janda

EMOTIONS IN CONTEXT

A Semantic Field Viewed through Grammar and Collocations

Abstract We examine Czech emotion nouns as a structured semantic field, adopting a field-level perspective that complements entry-level lexicographic description. The analysis is based on 65 nouns denoting emotions and related affective states, selected with reference to psychological classifications and corpus frequency. It combines grammatical profiling, correspondence analysis, and a brief collocational overview in order to show that these lexemes share certain patterns, some of them characteristic of the field as a whole and others pointing to its internal structuring. The results suggest that grammatical behaviour does not simply reflect semantic similarity in the usual psychological sense, but reveals different ways in which emotions are represented in language, for example as experienced states, emotional contents, agent-like forces, or states one is driven into. Collocational evidence points to broader tendencies across the field, including loss of agency, emotional intensity, bodily manifestation, regulation, and combination with other emotions. We argue that this kind of field-based description can complement traditional lexicographic practice and may also be attractive to dictionary users interested in contrasts between related words.

Keywords Semantic field; emotion nouns; grammatical profiles; collocations

1 Introduction

Emotion terms are represented in dictionaries through sense distinctions, definitions, grammatical information, collocations, examples of use, and semantic relations to other lexical items. In digital lexicographic resources, these relations can be made explicit through links between related entries, and recent lexicographic discussion has explored how such connections can be represented more systematically within interconnected data structures (e.g., Gracia et al., 2017; Měchura, 2016). A further perspective can be gained by systematically examining grammatical and collocational behaviour across a larger set of semantically related lexemes. The semantic field of emotions provides a useful case for such an approach, since it contains many closely related and partly overlapping nouns whose similarities and differences can be examined across the field as a whole.

This article approaches Czech emotion nouns from this field-level perspective. It also builds on approaches that treat lexical meaning as relational and emphasize the importance of lexical fields and combinatorial behaviour in lexicographic description (Mel'čuk, 1988). We examine how selected lexemes behave in actual usage: which grammatical cases they favour, which constructions they typically occur in, which verbs and adjectives they combine with, and how they co-occur with other emotion terms. By combining grammatical profiling with collocational analysis, we aim to describe the affective lexicon as an internally structured domain and to show how evidence from across a semantic field can add a comparative layer to entry-level lexicographic description.

Grammatical behaviour is an important part of this picture. Grammatical profiles do not encode meaning directly, but they reveal typical syntactic environments and characteristic ways in which emotions are represented in language. When combined with collocational analysis, they make it possible to distinguish lexemes that may appear close in meaning yet differ in their preferred constructions and patterns of use, as well as lexemes that are semantically more distant but show similar grammatical and collocational behaviour.

Correspondence analysis provides an overall view of the grammatical organization of the field. The patterns revealed from this perspective differ from those derived from psychological studies of emotions, which organize the domain according to other dimensions. A focused case study of the anger-related nouns shows in more detail how semantically related lexemes may form a relatively coherent group in corpus data while still differing in their grammatical preferences and typical collocational patterns.

2 Data and methods

2.1 Selection of nouns

The dataset was selected with reference to psychological classifications of emotions and affective states (Aydede, 2018; Brackett, 2019; Russell, 1980), but adapted for lexical and corpus-based analysis in Czech. The final sample contains 65 Czech nouns denoting emotions and related affective states. It is not intended as a complete inventory of emotion terms, but as a working set for comparative analysis.

Only nouns with at least 1,000 attestations in the combined SYN2015 and SYN2020 corpora were included. These corpora represent contemporary written Czech. The threshold was chosen in order to exclude marginal items and to ensure that grammatical profiles were based on enough data to reveal stable recurrent patterns.

Some potential candidates were excluded because they have prominent non-emotional senses that would substantially interfere with the analysis. For example, *pokoj* can mean both 'calm' and 'room'. The full list of analysed nouns and their English glosses, along with the code for the statistical analysis, is available at https://doi.org/10.18710/4TKI0T (Janda & Kováříková, 2026; R Core Team, 2022). For practical purposes, each Czech noun is assigned a single English gloss, used throughout the article.

2.2 Corpus data and grammatical profiles

In Czech, nouns are always obligatorily marked for grammatical case: Nominative, Genitive, Dative, Accusative, Vocative, Locative, or Instrumental. From the perspective of corpus data, each noun has a unique frequency distribution of attestations across the seven grammatical cases, termed a "grammatical profile" (Janda & Lyashevskaya, 2011). For Czech data, the profiles were obtained using the GramatiKat resource (Kováříková & Kovářík, 2023), which facilitates the extraction of grammatical profiles based on the Czech SYN2015 and SYN2020 reference corpora totaling 200M words.

A grammatical profile for each of the 65 emotion nouns was thus obtained, yielding a relative frequency distribution for each noun across the six syntactic grammatical cases (i.e., excluding the Vocative case, which is primarily used to address human beings and therefore of little relevance for most emotion terms). Table 1 displays a sample of this data, showing the ten nouns of highest corpus frequency. For example, 28.816% of attestations of *láska* 'love' appear in the Nominative case, 24.007% in the Genitive case, etc.

TABLE 1 Grammatical profiles of 10 high-frequency emotion nouns across syntactic cases: NOM = Nominative, GEN = Genitive, DAT = Dative, ACC = Accusative, LOC = Locative, INS = Instrumental

Noun and gloss	NOM	GEN	DAT	ACC	LOC	INS
láska 'love'	0.28816	0.24007	0.03197	0.1895	0.0858	0.10933
strach 'fear'	0.23046	0.152	0.01888	0.47333	0.02904	0.07356
radost 'joy'	0.15287	0.09226	0.02745	0.52079	0.01128	0.15243
štěstí 'happiness'	0.18726	0.22093	0.0319	0.46898	0.02672	0.06412
naděje 'hope'	0.20915	0.15385	0.0102	0.25556	0.08866	0.10269
obava 'worry'	0.2597	0.17267	0.04795	0.3721	0.04954	0.09791
touha 'desire'	0.33635	0.13684	0.03307	0.2267	0.04325	0.13431
vina 'guilt'	0.19149	0.26156	0.0049	0.2577	0.0999	0.15461
hrůza 'terror'	0.24837	0.10994	0.02259	0.21099	0.0486	0.24127
překvapení 'surprise'	0.26458	0.09161	0.1462	0.17871	0.00919	0.23392

2.3 Correspondence analysis

The matrix of relative frequency distributions serves as the input data for a correspondence analysis in which the mathematical distances between the rows and the columns are calculated in a multidimensional space, yielding a plot of the two dimensions that account for the most variance in the data. On the first round it was found that four nouns (see Table 2) have extremely skewed grammatical profiles, collapsing the remaining 61 nouns into a single clump. While these nouns are retained in the overall analysis, they have been excluded in Figure 1 to improve the visualization.

TABLE 2 Four nouns with extreme values (shaded): *rozpaky* 'embarrassment' rarely appears in Nominative, *tíseň* 'distress' is an outlier for Locative, and *údiv* 'astonishment' and *úžas* 'amazement' are outliers for both Dative and Instrumental

Noun and gloss	NOM	GEN	DAT	ACC	LOC	INS
rozpaky 'embarrassment'	0.09004	0.32894	0.00927	0.20669	0.27452	0.09053
tíseň 'distress'	0.18087	0.13427	0.0233	0.09687	0.52177	0.02636
údiv 'astonishment'	0.12298	0.12938	0.15869	0.17284	0.07143	0.33726
úžas 'amazement'	0.09907	0.09728	0.14092	0.16023	0.14843	0.353

2.4 Collocational analysis

Collocational analysis was based on the same corpora, SYN2015 and SYN2020. It was used to compare the typical lexical environments of individual emotion nouns and to identify recurrent patterns across the field. We focused mainly on four types of collocational partners: verbs, adjectives, nouns and other emotions. In the discussion, collocational evidence is used mainly as a complement to grammatical profiling and as a source of selected illustrative contrasts.

3 Grammatical profiles of Czech emotion nouns

This section presents the main grammatical patterns found in the analysed set of Czech emotion nouns. It begins with a correspondence analysis based on grammatical profiles, which helps visualize similarities and differences across the field. The aim is to show that grammatical behaviour provides useful evidence for the description of emotion nouns. The section then turns to the main case-based construals that emerge from the data and illustrates them with selected corpus examples.

3.1 The overall correspondence map

As a first step, we used correspondence analysis to provide an overall view of the grammatical profiles of Czech emotion nouns. This visualization makes it easier to see which lexemes are relatively close to one another in their grammatical behaviour and which stand out more clearly.

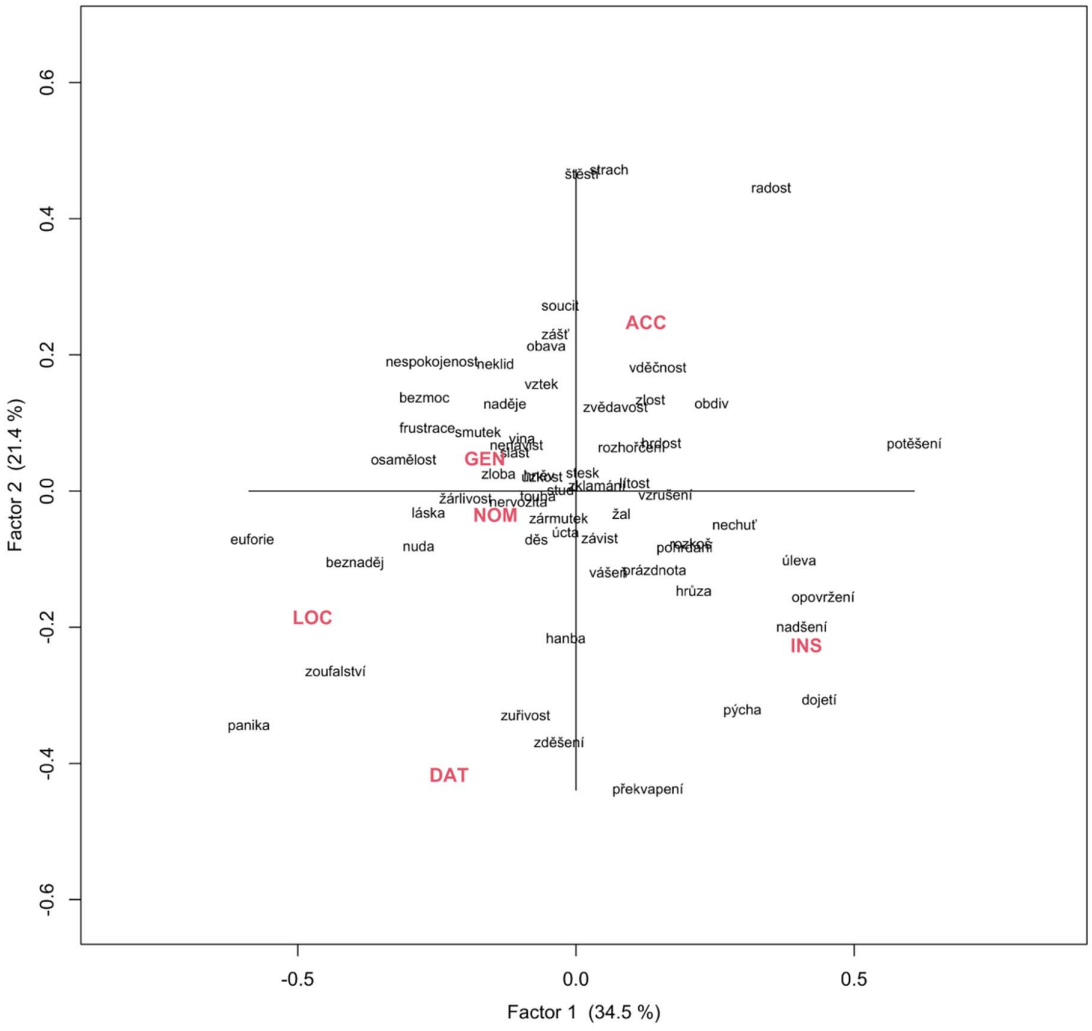


FIGURE 1 Correspondence analysis of grammatical profiles of Czech emotion nouns, accounting for 34.5% + 21.4% = 55.9% of variance (authors' data)

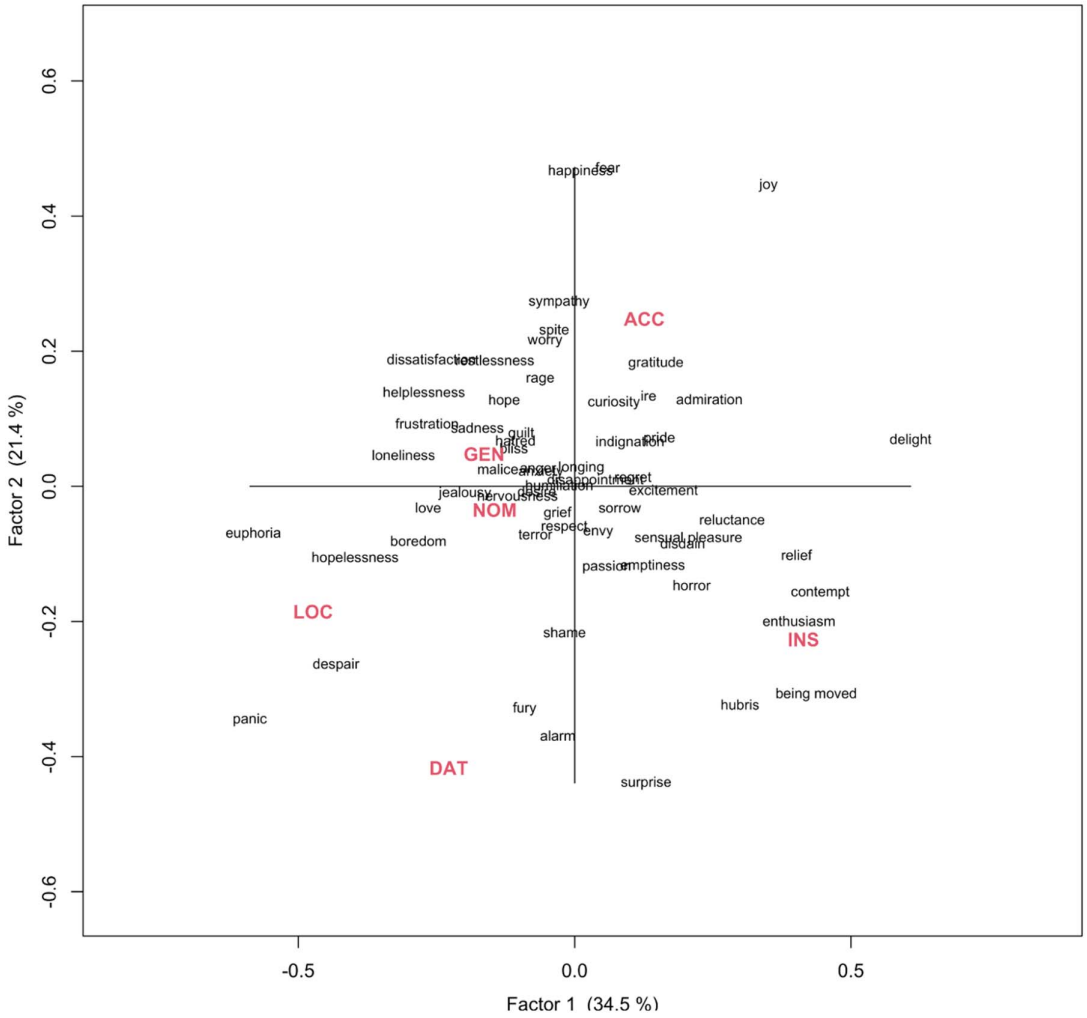


FIGURE 2 Correspondence analysis of grammatical profiles of Czech emotion nouns with English glosses for easier orientation (authors' data)

Figure 1 suggests that grammatical behaviour reflects not synonymy per se, but rather similarity in the ways emotions are grammatically construed. Words with similar meanings may be distributed across different parts of the plot, whereas words that show similar grammatical behaviour tend to be located closer to one another. Figure 2 presents the same map with English glosses for easier orientation.

For example, there are eight closely related positive emotion nouns distributed across different parts of the plot. Along the extreme edges of the top right quadrant, we find *štěstí* 'happiness', *radost* 'joy', and *potěšení* 'delight'. In the bottom right quadrant are *rozkoš* 'sensual pleasure', *úleva* 'relief', and *nadšení* 'enthusiasm'. *Slast* 'bliss' is near the centre in the top left, and *euforie* 'euphoria' is at the edge of the bottom left quadrant. One of the closest neighbours of *slast* 'bliss' is the antonymous *smutek* 'sadness', and *štěstí* 'happiness' is located very close to *strach* 'fear'.

The distribution of emotion nouns in Figure 1 does not correspond to the circumplex distribution of emotions that has been repeatedly validated in psychological studies (see Aydede, 2018; Vittersø, 2025). Psychological models of core affect typically describe affective experience along two bipolar dimensions: valence, which measures how positive or negative a state is (roughly in terms of pleasure vs. misery); and activation or arousal, which measures how much energy is

associated with an emotion (for example *thrilled* vs. *calm*). The psychological model therefore locates emotions according to their similarity in valence and arousal.

In contrast to this experiential perspective, our research investigates how emotions are represented in language, specifically in Czech. The distribution of words according to their grammatical behaviour can contribute to the differentiation of semantically related lexemes and reveal broader patterns in how emotions are linguistically construed. To interpret these patterns, we now turn to the semantic associations of the grammatical cases in Czech.

3.2 Case meanings as recurrent construals

Table 3 presents an overview of the cases relevant for the interpretation of Figure 1, along with both a schematic meaning for each case and a specification of the meanings most relevant for Czech emotion nouns.

TABLE 3 Czech grammatical cases, with their schematic meanings and the specific meanings most encountered in conjunction with emotion nouns

Case	Schematic meaning (Janda & Clancy, 2006)	Subset of meanings most encountered in Czech emotion nouns
Accusative	Object	state experienced or controlled
Genitive	Point of reference	substance, source/cause
Nominative	Subject	agent-like force
Instrumental	Accessory	accompanying state or force
Dative	Entity that projects a force	state one is driven into
Locative	Place	state one is in, situational frame

Cognitive Linguistics (Langacker, 2008) postulates that lexicon and grammar form a continuum along which all items have both form and meaning. From this perspective it is thus reasonable to expect that grammatical categories express meaning, albeit abstractly and always in conjunction with their lexical context and the grammatical constructions they appear in. In Czech the meanings of cases are often further specified by prepositions and other collocates. The semantics of case in Czech, as well as in the related language Russian, have been explored in detail (Janda, 1993, 1999, 2000; Janda & Clancy, 2002, 2006); for the purposes of this study, we refer only to the subset of case meanings most relevant for Czech emotion nouns.

Viewed in terms of these case associations, Figure 1 can be interpreted as a conceptual landscape of different ways in which emotions are linguistically construed. Lexemes in the upper part of the map, particularly towards the Accusative, are more strongly associated with construals of emotions as experienced or controlled states. The lower left is oriented towards Locative and Dative construals, in which emotions are represented as states one is in or is driven into, while the lower right is more strongly associated with Instrumental construals of emotions as accompanying states or forces. In the following section, we use corpus examples to illustrate these meanings in the order presented in Table 3.

3.3 Illustrative constructions by case

3.3.1 Accusative: state experienced or controlled

A prevalent grammatical construction in Czech uses an emotion term as the direct object of the verb *mít* 'have', identifying it as a possession, as in (1).

(1) *Myslela jsem, že budeš* ***mít radost***. 'I thought you would **be happy** [lit. **have joy-ACC**].'

Other verbs treat an emotion in the Accusative as a creature or force that can be aroused (2), suppressed (3), or hidden (4).

(2) *Jeho komentář* ***vyvolal údiv***. 'His comment **provoked astonishment**.'

(3) *Lloyd* ***potlačil hněv*** 'Lloyd suppressed [his] anger'

(4) *Isabella se ani nesnažila* ***skrýt zklamání***. 'Isabella didn't even try to **hide [her] disappointment**.'

3.3.2 Genitive: substance, source/cause

A common collocate of Czech emotion nouns is the adjective *plný* 'full', which governs the Genitive and indicates that the emotion is a substance that fills up a person or some part of their body, as in (5). The amount of emotion-as-a-substance can be measured, usually in large units, like 'wave' in (6). Alternatively, one can be devoid of the substance of an emotion as in (7), where the preposition *bez* 'without' governs the Genitive.

(5) *[ona] na něj hleděla očima* ***plnýma naděje a úžasu***. '[she] looked at him with eyes **full of hope and amazement**.'

(6) *Zaplavila ho* ***vlna touhy***. 'A **wave of desire** washed over him.'

(7) *Rodnou zemi opouštěli* ***bez lítosti***. 'They abandoned their homeland **without regret**.'

In concrete domains the prepositions *z* 'from' and *do* 'to' describe motion away and toward, while in abstract domains they signal sources or causes and destinations, as in (8) and (9).

(8) ***Z vděčnosti*** *ho Viktorie políbila*. '**In gratitude** [lit. **from gratitude-GEN**] Viktorie kissed him.'

(9) ***Přivádělo by ji do rozpaků****, že vám tohle všechno vykládám*. 'It would **embarrass her** [lit. **bring her to embarrassment-GEN**] that I am telling you all this.'

3.3.3 Nominative: agent-like force

As the subject of a clause, an emotion behaves as an agent that can take control of people or grow, as in (10).

(10) *moje* ***nervozita narůstala***. 'my **nervousness was growing**.'

3.3.4 Instrumental: accompanying state or force

In conjunction with the preposition *s* 'with', an emotion in the Instrumental is portrayed as a state that accompanies another action, as in (11). An emotion in the Instrumental can also appear as the item that facilitates an (often uncontrolled) action, as in (12).

(11) *Kassandra jí **s úlevou** pohlédla do tváře*. 'Kassandra looked her in the face **with relief**.'

(12) *Samozřejmě, že by se Charlie **dmul pýchou** až k prasknutí*. 'Of course, Charlie would be **puffed up** to the bursting point **with pride** [lit. **by means of pride**].'

3.3.5 Dative: state one is driven into

An emotion in the Dative wields a power of its own, subjugating people, as in (13) or drawing them into its sphere of influence, as in (14).

(13) *Další naděje zhasla a já znovu **propadla zoufalství***. 'Further hope was extinguished and once again I **succumbed to despair**.'

(14) ***K zděšení** mých druhů jsem se začal smát*. '**To the dismay** of my companions, I started to laugh.'

3.3.6 Locative: state one is in

The Locative case portrays an emotion as a container that holds a person, as in (15).

(15) *Odvrátila obličej a **v zoufalství** zavřela oči*. 'She turned her face away and closed her eyes **in despair**.'

In sum, emotion nouns can be distinguished according to the statistical distribution of the grammatical cases that they appear in. In turn, these case distributions reflect grammatical constructions that cast emotions in various metaphorical roles as objects, substances, locations, and creatures over which human beings have greater or lesser control.

4 Collocational structuring of the field

The grammatical profiles discussed in Section 3 already suggest that Czech emotion nouns form a field with an internal organization. Collocational evidence offers a further perspective by showing the lexical environments in which these nouns typically occur. Collocations reveal semantic tendencies that dovetail with those that emerge from patterns of case usage. This is especially clear in the case of verbs, and to a lesser extent adjectives, whose valency helps shape the grammatical behaviour of emotion nouns by favouring particular constructions and case patterns. We focus on several broader patterns typical of the field and its internal organization. Some of these patterns are more widespread, while others are more strongly associated with smaller groups of lexemes.

4.1 Verbal collocations

Verbs show several common ways in which emotions are described in Czech. Some verbs present them as inner states that people feel or go through, such as *cítit* 'feel', *pociťovat* 'feel/experience', and *prožívat* 'experience/live through'. Especially *cítit* and *pociťovat* are found with a large majority of the emotion nouns in the dataset. Other verbs show that emotions can be caused, for example *vyvolat* 'evoke' and *vzbudit* 'arouse'. Another frequent pattern appears with verbs such as *zaplavit* 'flood', *zmocnit se* 'take hold of', *přepadnout* 'attack', and *zachvátit* 'seize'. In these cases, the emotion takes control and the experiencer loses control. The data also contain many verbs of control, especially *potlačit* 'suppress', *ovládnout* 'control', *skrývat* 'conceal',

and *skrýt* 'hide'. In some parts of the field, especially around anger-related nouns, verbs of release are also common, such as *vybít si* 'discharge' and *vylít si* 'pour out'.

4.2 Adjectival collocations

Adjectives add several further types of information. Especially common is *plný* 'full of (emotion)', which combines with a wide range of emotion nouns and belongs to the most broadly distributed adjectival collocates in the field. Another group expresses intensity, with adjectives such as *nesmírný* 'immense', *hluboký* 'deep', and *naprostý* 'absolute'. This shows that emotions can appear in stronger or weaker form. Another group concerns authenticity: *upřímný* 'sincere', *nelíčený* 'unfeigned', *nepředstíraný* 'unpretended', and *předstíraný* 'pretended' describe emotions as genuine or not genuine. Some adjectives are more typical of socially less acceptable emotions. These include words linked to hiding and control, such as *neskrývaný* 'openly shown', *skrývaný* 'hidden', and *potlačovaný* 'suppressed', as well as evaluative adjectives such as *oprávněný* 'justified' and *spravedlivý* 'just/righteous'.

4.3 Noun collocates

Noun collocates often present emotions as sudden and strong events, through nouns such as *vlna* 'wave', *nával* 'rush, surge', *výbuch* 'outburst, explosion', and *záchvat* 'attack, fit'. This is especially clear with anger-like emotions, which often appear as eruptive states. The frequent noun *směsice* 'mixture' suggests that emotions can be mixed states including both positive and negative affect rather than clearly separate units. Some nouns connect emotion with physical expression and bodily experience, e.g. *slzy* 'tears', *bolest* 'pain', *tvář* 'face', *srdce* 'heart', and *krev* 'blood'. A similar tendency can also be seen in some verbs such as *třást se* 'tremble' and *zrudnout* 'turn red', and in adjectives such as *zrudlý* 'red, flushed' and *ochromený* 'paralysed'.

4.4 Emotion–emotion collocations

Links between emotion nouns are very common in the data. In the top-100 collocate lists, every noun co-occurs with at least a few other emotion nouns, and many of them do so very often. The highest numbers are found for *bezmoc* 'helplessness', *hněv* 'anger', and *smutek* 'sadness', each with 24 other emotion nouns in their top-100 profiles. Similar emotions often occur together, especially in the domains of anger, fear, and despair. Positive emotions also co-occur with other positive emotions, though usually less densely. Combinations of positive and negative emotions are also common, for example *radost* 'joy' with *smutek* 'sadness' or *naděje* 'hope' with *zoufalství* 'despair'. This shows that Czech emotion nouns are often presented in relation to other emotions, not on their own.

These collocational patterns further support the view that the Czech emotion lexicon has an internal organization. They point both to broader tendencies shared across the field and to smaller clusters that are less visible at the level of individual entries. Collocational evidence thus complements the grammatical profiles discussed above.

5 Case study: Anger-related nouns

Within the dataset of 65 emotion nouns, we focus on six semantically related anger nouns: *hněv* 'anger', *vztek* 'rage', *zlost* 'ire', *zloba* 'malice', *zuřivost* 'fury', and *rozhořčení* 'indignation'. These lexemes are closely related in meaning, and their grammatical profiles place them relatively

close to one another, while their collocational behaviour reveals an important shared core. At the same time, the group is not homogeneous: the individual lexemes differ in their case preferences and in the types of constructions in which they typically occur. The anger-related group therefore provides a useful case for examining how corpus evidence can differentiate semantically closely related emotion words.

From the perspective of case distribution, the anger-related group is relatively central rather than extreme. None of the six lexemes is strongly dominated by a single case, and as a group they show a fairly balanced profile. At the same time, clear internal differences can still be observed. *Hněv* 'anger' and *zloba* 'malice' show a stronger preference for Nominative and Genitive, while *vztek* 'rage', *zlost* 'ire', and to a lesser extent *rozhořčení* 'indignation' are more strongly associated with the Accusative. This suggests that the first pair is more often understood as agent-like forces or emotional contents, whereas the second group more often appears as experienced and controlled states.

Some lexemes also stand out in more specific ways. The relatively strong Genitive profile of *zloba* 'malice' is supported by expressions such as *plný zloby* 'full of malice', *tolik zloby* 'so much malice', or *bez zloby* 'without malice', which present malice as emotional content that can be present or absent. The clearest outlier is *zuřivost* 'fury', which has a noticeably stronger Dative profile than the other lexemes, due to constructions such as *dohnat k zuřivosti* 'drive into fury' and *přivádět k zuřivosti* 'bring into fury', where the emotion is presented not as something one has or controls, but as a state one is driven into.

Case distribution helps identify the internal structure of the anger-related group, while collocational data make both its shared tendencies and its internal differentiation more visible. Verbs such as *vzplanout* 'flare up', *vzkypět* 'boil up', *popadnout* 'seize', and *zachvátit* 'grab', as well as nouns such as *výbuch* 'explosion', *záchvat* 'fit', and *nával* 'surge', present anger as a dynamic and episodic phenomenon that flares up or erupts. Other collocates point to bodily manifestation: anger makes people tremble, redden, distort their faces, or show other visible signs of inner tension. At the same time, these emotions are often presented as states that need to be controlled, as shown by verbs such as *potlačit* 'suppress', *ovládnout* 'take control', *skrývat* 'hide', and *utišit* 'mute'. In this group, anger is typically presented as something that arises suddenly, escalates, becomes visible on the body, and often needs to be controlled.

Within this shared core, the individual lexemes remain distinct. *Vztek* 'rage' and *zlost* 'ire' are more often linked to Accusative constructions, especially with verbs such as *mít* 'have', *vylít si* 'pour out one's', and *vybít si* 'work out one's'. This makes them more typical of anger that is strongly felt and then vented. *Zuřivost* 'fury', by contrast, is more often found in the Dative expressions mentioned above, which present it as an extreme state that a person is pushed into. *Rozhořčení* 'indignation' is more strongly associated with moral evaluation and overt expression, as in *spravedlivé rozhořčení* 'righteous indignation', *oprávněné rozhořčení* 'justified indignation', or *vyjádřit rozhořčení* 'express indignation'. *Hněv* 'anger' and *zloba* 'malice' differ in another way. Both show a stronger preference for the Nominative and Genitive, but hněv 'anger' is more often linked to dynamic processes such as *vzplanout* 'flare up', *vzkypět* 'boil up', and *vyprchat* 'fade away', while *zloba* 'malice' more often appears as emotional content or substance-like material (*plný zloby* 'full of malice'). The anger-related group thus shows how semantically closely related emotion words can be differentiated by combinations of grammatical preferences and typical constructions.

In a digital lexicographic resource, an entry for *hněv* could therefore be linked to a field-level comparison of semantically closely related lexemes such as *hněv*, *vztek*, *zlost*, *zloba*, *zuřivost*, and *rozhořčení*. Their grammatical profiles, typical constructions, and collocational patterns reveal

systematic differences in how these lexemes are used and how the emotions they denote are linguistically construed. The comparison could also show lexeme-specific characteristics, such as the stronger association of *rozhořčení* with evaluative and expressive contexts. This information could be presented in a comparative panel linked to the individual entries, allowing users to view the lexemes side by side.

6 Conclusion and lexicographic implications

We show that Czech emotion nouns form a field structured not only by semantic similarity. Their internal organization becomes clearer when grammatical behaviour and collocational patterns are taken into account. Some lexemes pattern as experienced and controlled states, others as emotional contents, agent-like forces, or states one is driven into. The collocational evidence points in a similar direction, revealing broader tendencies such as loss of agency, regulation, bodily manifestation, emotional intensity, and combination with other emotions. The anger-related group shows particularly clearly that semantically closely related lexemes may form a relatively coherent group while remaining internally differentiated. The grammatical organization revealed here also provides a linguistically grounded perspective on the emotion domain that complements psychological models based on experiential dimensions such as valence and arousal.

These findings have several lexicographic implications. Individual dictionary entries do not always make it easy to see how closely related lexemes differ in use, how emotion words are associated with one another, or how lexemes that are less closely related in meaning may nevertheless show similar patterns of use. A field-based description can therefore complement traditional lexicographic practice by situating individual lexemes within a structured network of contrasts and associations. This may also make dictionary resources more useful to readers who are looking for clearer comparisons between related words.

An important contribution of this approach is the attention paid to grammatical behaviour. Grammatical profiles help capture what is typical of a given lexeme across its recurrent constructions. In this sense, grammatical behaviour is not external to lexical meaning, but part of the lexical profile of the word and part of the way its meaning is realised in use. Collocational evidence is equally important, since it shows typical lexical partners and typical scenarios of use, but grammatical profiling may be especially valuable here because it is still rarely included in lexicographic description. This makes the approach particularly relevant for digital corpus-based dictionaries, where such information can be presented more flexibly and systematically than in traditional printed entries. It may be especially useful in learner lexicography, where users often need help not only with meaning, but also with typical constructions and patterns of use.

If such a semantic field is to be described more fully, it would be useful to include not only nouns, but also other lexical classes that participate in the expression of emotional meaning. In the affective domain, nouns can be linked to related adjectives, verbs, and adverbs, for example *zlost* 'ire', *rozzlobený* 'angry', *zlobit se* 'be angry', and *rozzlobeně* 'angrily', or *láska* 'love', *milující* 'loving', *zamilovaný* 'in love', *milovat* 'love [verb]', and *zamilovaně* 'lovingly'. A description of this kind could be useful both for native speakers, who may benefit from a clearer view of lexical relations within the field, and for learners of Czech as a second language, who often need help with meaning differences as well as with typical constructions and usage patterns. A similar approach could also be applied to other frequent semantic fields, such as colours, body parts, time, evaluation, size, quantity, modality, or verbs of motion, speaking and thinking.

References

Aydede, M. (2018). *Sensory versus core affect*. Presentation at the Moods Symposium at the Canadian Philosophical Association, Montreal, June 2018. https://philarchive.org/rec/AYDSVC

Brackett, M. (2019). *Permission to Feel*. Celadon Books.

Gracia, J., Kernerman, I. & Bosque-Gil, J. (2017). Toward linked data-native dictionaries. In I. Kosem et al. (Eds.), *Electronic Lexicography in the 21st Century: Proceedings of eLex 2017* (pp. 550–559). Lexical Computing CZ.

Janda, L. A. (1993). *A Geography of Case Semantics: The Czech Dative and the Russian Instrumental*. De Gruyter.

Janda, L. A. (1999). Peircean semiotics and cognitive linguistics: A case study of the Russian genitive. In M. Shapiro (Ed.), *The Peirce Seminar Papers* (pp. 441–466). Berghahn Books.

Janda, L. A. (2000). A cognitive model of the Russian accusative case. In R. K. Potapova, V. D. Solov'ev & V. N. Poljakov (Eds.), *Trudy meždunarodnoj konferencii Kognitivnoe modelirovanie* (Vol. 4, pp. 20–43). MISIS.

Janda, L. A. & Clancy, S. J. (2002). *The Case Book for Russian*. Slavica Publishers.

Janda, L. A. & Clancy, S. J. (2006). *The Case Book for Czech*. Slavica Publishers.

Janda, L. A. & Kováříková, D. (2026). *Replication Data for: Emotions in Context: A Semantic Field Viewed through Grammar and Collocation* [Data set]. DataverseNO. https://doi.org/10.18710/4TKI0T

Janda, L. A. & Lyashevskaya, O. (2011). Grammatical profiles and the interaction of the lexicon with aspect, tense, and mood in Russian. *Cognitive Linguistics*, 22(4), 719–763.

Kováříková, D. & Kovářík, O. (2023). *GramatiKat, version 2. Nástroj pro výzkum gramatických kategorií a gramatických profilů*. FF UK. http://www.korpus.cz/gramatikat

Langacker, R. W. (2008). *Cognitive Grammar: A Basic Introduction*. Oxford University Press.

Měchura, M. (2016). Data structures in lexicography: From trees to graphs. In A. Horák, P. Rychlý & A. Rambousek (Eds.), *The 10th Workshop on Recent Advances in Slavonic Natural Languages Processing (RASLAN 2016)* (pp. 97–104). Tribun EU.

Mel'čuk, I. A. (1988). Semantic description of lexical units in an explanatory combinatorial dictionary: Basic principles and heuristic criteria. *International Journal of Lexicography*, 1(3), 165–188.

R Core Team. (2022). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161–1178.

Vittersø, J. (2025). *Humanistic Well-being*. Springer. https://doi.org/10.1007/978-3-031-69292-5

Contact information

Dominika Kováříková
Faculty of Science, Humanities and Education, Technical University of Liberec
dominika.kovarikova@tul.cz

Laura A. Janda
UiT The Arctic University of Norway
laura.janda@uit.no

Gunta Nešpore-Bērzkalne, Mikus Grasmanis &
Madara Stāde

POSSIBILITIES OF REPRESENTATION OF MULTI-WORD EXPRESSIONS IN THE LATVIAN ELECTRONIC DICTIONARY *TĒZAURS*

Abstract This paper aims to examine the existing representation of multi-word expression (MWE) entries in the electronic Latvian dictionary *Tēzaurs* and characterise their interaction with other, single-word entries, as well as highlight future improvement for better integration of MWEs in the dictionary data. The study arises from the need to examine MWE entries in more detail, as they are a substantial part of a language, but are often given less attention than single-word entries in a dictionary. The paper gives an overview of MWE representation within the *Tēzaurs* structure, e.g., links with component-word entries, corpus examples, sources, as well as semantic links with single-word senses or other MWEs. Linking to other resources, such as the Princeton English WordNet is also mentioned. This overview shows that MWEs in *Tēzaurs* function as independent units that can freely interact with other entries and be embedded in other entries. MWEs also work well in the resource's system of links between entries. The next step is expanding the representation of MWEs in this regard.

Keywords MWE entries; electronic dictionary *Tēzaurs*; intra-dictionary links

1 Introduction

Tēzaurs[1] is the largest Latvian electronic monolingual dictionary with more than 410,000 entries, which emerged as a compilation from nearly 300 pre-existing dictionaries and other sources but has since developed into a multifunctional lexical resource consisting of granular interlinked entities enriched with various kinds of metadata. The *Tēzaurs* system is built as a hybrid database with relational and graph database features, containing entities of relatively small granularity, such as lexemes, senses, examples, etc., which can all have a rich and extendible set of attributes describing their syntactical, semantical, morphological, derivational, stylistic, normative and other aspects, and which can then be linked in various interesting ways (Klints et al., 2024). This structure was mostly influenced by the need to represent the data of a small initial set of included dictionaries; however, practice has shown that this structure also functions well for other included monolingual dictionaries. Besides entries for single words, *Tēzaurs* also contains ~76,000 entries of multi-word expressions (MWEs; dictionary entries that contain more than one orthographic word (Bauer, 2021)) (Grasmanis et al., 2023). Thus, it is currently the most comprehensive resource of MWEs in Latvian lexicography.

Until recently, MWE entries have been a relatively little-managed category of the *Tēzaurs* data despite their number and variety; other types of entries, such as nouns, verbs, adjectives and adverbs have garnered more revisionary attention. However, in recent years, there have been more concentrated efforts to systematise MWEs, as the scope of their functionality increases within the overall structure of *Tēzaurs* and the language resources it supports. This paper aims to examine *Tēzaurs* specifically from the perspective of MWEs in order to see to what extent the functionality of the resource can support the representation of MWEs and what can still be

[1] Publicly available at www.tezaurs.lv or as data from https://repository.clarin.lv/repository/xmlui/handle/20.500.12574/160.

improved in this regard. Additionally, the paper gives an overview of the relation data (various types of links) the MWE entries have amassed so far.

The thematic structure of the following portion of the article is as follows. The second chapter outlines the number and diversity of MWE entries in *Tēzaurs*, as well as the nuances of their representation in the dictionary. The third chapter describes all possible links and relations within *Tēzaurs* that are compatible with MWE entries, such as component links, Latvian WordNet sense relations, interlingual links, and others. The fourth chapter gives an overview of MWE searchability in *Tēzaurs*. This is followed by conclusions, suggestions for future work, and a list of references.

2 MWE Entries in Tēzaurs: an overview

MWEs in *Tēzaurs* are full-fledged entries alongside single-word entries (like in the resources described, e.g., in (Al-Haj et al., 2013; Ermakova et al., 2022). They have nearly all the same properties and elements as single-word entries (see entry view example in Figure 2): MWE lexical variants; attributes of style and usage context; sense definitions; an option for adding usage examples from corpora; an option for links with other entries or resources. Additionally, MWE entries contain links to their components, as well as MWE category tags according to their function and level of figurativeness (for a more detailed description of MWE classification in *Tēzaurs*, see Rituma et al., 2024).

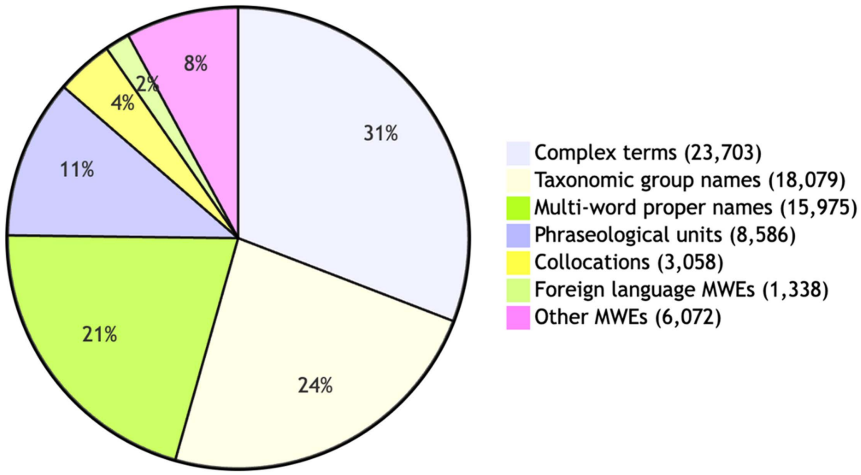


FIGURE 1 MWEs categories in *Tēzaurs* and the distribution of 76,811 MWEs within them

MWEs are sorted into the following categories (see Figure 1):

- **complex terms**, e.g. *centrtieces spēks* 'centripetal force'; *darāmā kārta* 'active voice';
- **taxonomic group names**, such as species, families or classes, both international, e.g. *Canis lupus, Anemone coronaria*, and Latvian, e.g. *parastā bērzubeka* 'birch bolete', *purvu sisenis* 'large marsh grasshopper';
- **multi-word proper names**, e.g. *Aleksandrs Lielais* 'Alexander the Great', *Baltijas jūra* 'Baltic Sea';
- **phraseological units**, i.e., idiomatic MWEs, e.g. *strēbt karstu putru* (lit. 'to gulp hot gruel') 'to act rashly, hastily', *atstiept kājas* (lit. 'to stretch one's legs') 'to die';

- **collocations**, e.g. *kartupeļu biezputra* 'mashed potatoes', *apģērba gabals* 'piece of clothing'. We have adopted a rather narrow understanding of collocations, which are statistically significant co-occurrences of words outside of all previously mentioned groups. In other words, collocations are fixed word combinations with no semantic reinterpretation (Veisbergs, 2019);
- **foreign language MWEs**, e.g., *finita la commedia, de facto*;
- **other MWEs**. A small part of the MWEs has not been categorized yet. This is either because they cannot be assigned to any of the existing MWE categories (such as abbreviations consisting of multiple tokens, e.g., *t. sk.* 'incl.', *a god.* 'esteemed'), or because they have been recently added to *Tēzaurs*, but have not been assigned yet a full set of tags the category detection depends on.

Despite having nearly equivalent properties to single-word entries, MWE entries are deliberately separated from them; they have their own lexical neighbourhood and appear in a different section of search results.

Our aim is to maintain the existing information, add more data and expand the functionality of MWE entries.

3 Possible links and relations

Although single-word and MWE entries share multiple similarities, they also exhibit several differences. The sense definitions and labels of style and usage context are nearly identical, but, unlike single-word entries, MWE entries display no grammatical information (e.g., declension tables and morpho-syntactic information). Furthermore, MWE entries, especially phraseological units, tend to have significantly more MWE lexical variants. At present, these variants are displayed to users as a list; however, for entries with many lexical variants, more user-friendly presentation methods should be considered.

A significant part of the *Tēzaurs* entries (both single-word and MWE) are connected by several types of semantic links within *Tēzaurs* itself, as well as to the Princeton English Wordnet (PWN; Fellbaum, 1998). The created links are navigable in both directions. As the survey by (Mititelu et al., 2025) shows, less than half of the MWE resources examined by the survey have links to corpora or other lexical resources, even though such links increase their applicability for NLP tasks. For the *Tēzaurs* developers, all types of links are viewed as valuable and potentially expandable. This chapter examines the existing links in *Tēzaurs* that also involve MWEs and describes how they function for MWEs specifically.

Figure 2 shows the conceptual data model of *Tēzaurs*. Boxes represent entities in the *Tēzaurs* data model. Consistent lines with bullets at the end represent one-to-many relations between entities. Other lines represent many-to-many relations and "partOf" relations (component links) that are outside the main hierarchies. There are three lexeme types: "word", "multi-word expression" or "word part". Entries inherit their type from the primary lexeme.

The classical dictionary entry representation starts with an entry root, which links to contained lexemes and senses. Senses can have examples, optionally with a corpus reference, as well as subsenses. Lexemes can have wordforms, which provide properties (grammatical, pronuncial, etc.) to specific irregular forms of a lexeme's lemma. Entries can be in a (lexical) "partOf" relation with other entries (typically between a word and an MWE) either directly or via sense. In the latter case, the word's occurrence in the MWE carries that sense. Another

hierarchy is concentrated around synsets, where senses are parts of a synset, and synsets can have both in-dictionary relations (synonymy, antonymy, hyponymy, meronymy) and external relations to PWN (full match (synonymy), or interlingual hypernymy/hyponymy).

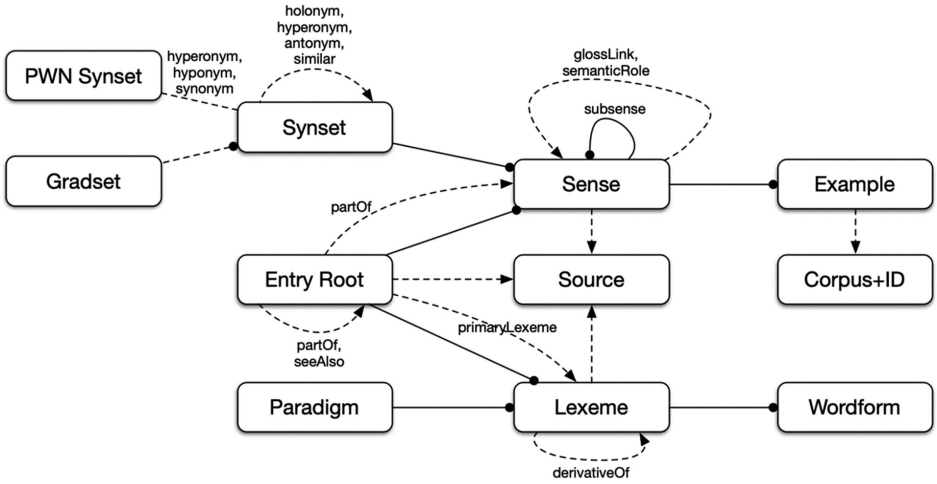


FIGURE 2 Conceptual data model of the dictionary *Tēzaurs*

All entities and links can carry rich metadata in the form of attributes (key-value pairs) representing different grammatical aspects, usage, etc. Currently, there are more than 80 single- and multi-value attributes with more than 700 possible values.

Most of the linking features in *Tēzaurs* initially have been developed and actual links have been created primarily for single-word entries. Therefore, for MWE entries, the number of links currently is comparatively small (except links with component entries) as there have not been concentrated efforts to create links specifically for MWE entries. Nevertheless, the existing linking functionality is fully applicable to MWE entries, although it would benefit from some enhancements.

Table 1 demonstrates how many MWEs of each type have at least one link of a certain kind, as well as the total number of such links in each MWE category. The total number of links for all MWE types is seen at the bottom of the table. Column 1 contains MWE category names. Column 2 indicates the total number of each MWE category in *Tēzaurs*. Columns 3–6 indicate how many entries of the corresponding MWE category have at least one component link, corpus example, Latvian WordNet synset or link to Princeton English WordNet; the numbers in parentheses show the total number of each link type in the respective MWE category. The following subchapters provide a more detailed description of all mentioned link types.

3.1. Links with MWE component entries

Links with component entries are used exclusively for MWE entries. This link enables MWE entries to be linked with content words (or in some cases – other MWEs) they consist of (relation "partOf" in Figure 2; section "Linked Entries" in Figure 3). This is how MWEs are displayed in single-word entries: the entry of the respective word shows all MWEs that the word is linked with.

TABLE 1 The number of MWE entries (by category) with various links and relations

MWE Category	No. of MWEs	Have comp. links (no. of links)	Have corpus ex. (no. of ex.)	Incl. in LWN synsets (no. of synsets)	Have PWN links (no. of links)
1	2	3	4	5	6
Complex terms	23,703	18,433 (24,702)	397 (1,609)	595 (597)	294 (322)
Taxonomic group names	18,079	17,942 (18,256)	14 (29)	47 (47)	402 (429)
Multi-word proper names	15,975	4,619 (4,664)	1 (5)	33 (36)	117 (128)
Phraseological units	8,586	6,407 (12,158)	473 (2,014)	660 (714)	95 (127)
Collocations	3,058	2,944 (4,144)	239 (1,157)	232 (238)	30 (40)
Foreign language MWEs	1,338	530 (537)	0 (0)	6 (6)	63 (70)
Other MWEs	6,072	3,547 (5,526)	227 (1,275)	291 (310)	45 (52)
Total	**76,811**	**54,422 (69,987)**	**1,351 (6,089)**	**1,864 (1,948)**	**1,046 (1,168)**

Component links are not assigned in the same way for all types of MWEs.

For phraseological units, terms (including taxonomic group names) and collocations, our aim is to have links with all their content word components. Currently, almost 80 % of the entries in these three categories have at least one (often more than one) link with component entries (see Table 1).

Multi-word proper names and taxonomic group names are usually linked to one component entry only. Therefore, the numbers of component links for multi-word proper names and taxonomic group names are close to the number of entries of these categories involved in links with component entries. Multi-word proper names mostly are linked to the single-word proper names they contain (if there is a corresponding entry in *Tēzaurs*). For example, the placename *Juglas ezers* 'Jugla Lake' is linked only to the entry of its component *Jugla* (a placename), as it would not be practical to link all lake names to the entry *ezers* 'lake' where they would form a long list of all lakes mentioned in *Tēzaurs*. If a multi-word proper name does not contain any single-word proper name (e.g., *Aklā upīte* (lit. 'The Blind River'), it has no component links. Similarly, multi-word personal names (e.g., *Jānis Kristītājs* 'John the Baptist') often have no component links, since there are no entries corresponding to these components in *Tēzaurs*. Thus, a comparatively small part (less than a third) of MWEs in this category have component links. For example, the entries *Baltpurna delfīns* 'White-beaked dolphin' and *Meža tulpe* 'Wild tulip' have a link only to the component entries *delfīns* 'dolphin' and *tulpe* 'tulip', respectively. This allows users to see, e.g., all tulip varieties whose names contain the word *tulpe* 'tulip' when viewing the entry *tulpe* 'tulip'.

Foreign language MWEs (e.g., *carpe diem* 'seize the day') are linked to their components only if *Tēzaurs* have entries for the corresponding foreign words – usually in the case of some Latin terms or other, smaller groups of foreign words.

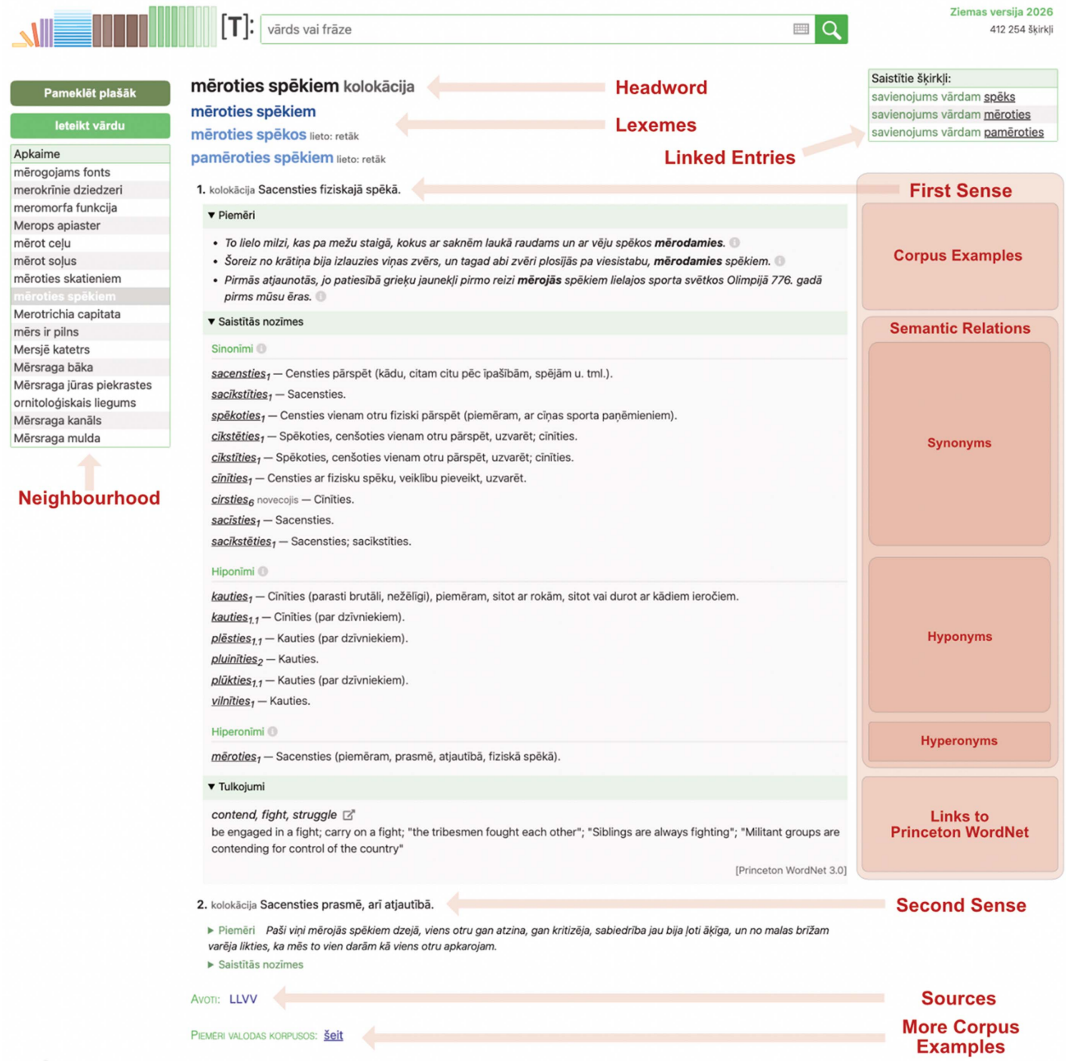


FIGURE 3 Entry for the MWE *mēroties spēkiem* (lit. 'to measure strengths) 'to compete against each other' in the electronic dictionary *Tēzaurs* (https://tezaurs.lv/mwe:304807)

Although component links typically connect MWE entries with single-word entries, it may exist between two MWEs as well, where one is a part of the other (e. g., *līdz ausīm* (lit. 'up to the ears') 'to a large extent' un *iemīlēties līdz ausīm* (lit. 'to fall in love up to the ears') 'to fall head over heels in love').

In total, there are currently 69,987 links between MWEs and their components that involve 54,422 MWE entries, with ongoing work to expand this number.

3.2. Inclusion into Latvian WordNet synsets

The development of the Latvian WordNet (LWN) began in 2019 and is currently ongoing (Paikens et al., 2023). In a wordnet, word senses are grouped into synonym sets (synsets), which in turn are linked to each other based on various other semantic relations, e.g., hypernymy, meronymy, antonymy, etc. Since MWEs constitute an important part of a lexicon, their inclusion in the Latvian WordNet is a significant improvement of the resource. LWN is based on the sense inventory of the entries found in *Tēzaurs*, which is also the largest Latvian resource of

MWEs. The centralised infrastructure makes the inclusion of MWEs in a wordnet easier, which is not always the case for wordnets. For instance, the developers of the Polish WordNet (*plWordNet*) extracted MWE candidates from large corpora and included them in the *plWordNet* after (Piasecki et al. 2015), whereas for Greek, a separate MWE database *IDION* was created, where synonymous MWEs grouped, potentially facilitating future integration into a wordnet (Markantonatou et al. 2019).

MWEs can participate in synset formation just as any single-word entry. Currently in LWN, there are 1,948 synsets involving 1864 MWE entries. However, not all MWE categories participate in synsets with the same frequency (although the overall number of synsets containing MWEs in *Tēzaurs* is still very small, these tendencies can already be observed – see Table 1, Column 5). Overall, synonymy is mostly exhibited by phraseological units and collocations. In taxonomic group names or personal names, synonyms occur only in separate cases; similarly, complex terms do not have many synonyms either. This is because ideally in these groups, one notion corresponds to only one word or MWE, but in practice it is not always so – there are some parallel term cases where one term is native while the other is a loanword (e.g., *kustību nervi* (lit. 'motion nerves') and *motoriskie nervi* (lit. 'motoric nerves') 'motor nerves') and cases of several parallel names for places or nature objects (e.g., a lake that has three names: *Vilku ezers*, *Laimiņu ezers* and *Vilkezers*). Foreign language MWEs are included in synsets only in exceptionally rare cases, as they generally do not have synonyms.

If the meaning of an MWE is synonymous with certain single-word senses, they may form a combined synset; this occurs in most cases, as illustrated by the MWE *mēroties spēkiem* 'to compete against each other', which is placed into a synset with 9 single-word senses with the meaning 'to compete', e.g., *sacensties, sacīkstīties, spēkoties, cīkstēties*, (see Figure 3, section "Synonyms"). The inclusion of idiomatic MWEs into synsets demonstrates the occasional interchangeability of certain types of MWEs and single-word entries and ensures the participation of MWEs in other semantic relations between synsets. Currently there are 5355 semantic links between 875 synsets containing at least one MWE. However, there are also cases, when certain types of MWEs do not match any of the single-word senses since their meaning is too specific or complex for single-word entries to cover. Such MWE senses can be combined into separate, MWE-only synsets, which helps grouping MWEs more efficiently (e.g., a synset containing three MWEs *balts kā krīts* (lit. 'white as chalk'), *balts kā kaļķis* (lit. 'white as limestone'), *balts kā drēbe* (lit. 'white as a cloth') 'very pale').

3.3. Links to Princeton English WordNet

MWEs can also be included in the interlingual links to PWN; there are currently 1,168 such links and 1,046 MWEs involved (see Table 1, Column 6 and Figure 3, section "Links to Princeton WordNet"). Most of these MWEs are a part of larger synsets, however, cases of MWEs linked to PWN as one-element synsets, (i.e., individually) also exist (e.g, *sveiks un vesels* (lit. 'safe and healthy') and *safe_and_sound* in PWN) but are rare as there have not been systematic efforts to translate Latvian MWEs. Although primarily beneficial in the bilingual context, interlingual links have potential for a wider coverage of languages through the Open Multilingual Wordnet.

3.4. Links to corpus examples

The Tezaurs.lv platform offers an option of assisted adding of MWE corpus examples from the Latvian National Corpora Collection (Saulīte et al., 2022). Any provided single word can be searched for in any corpus, and all matching sentences are presented for example selection,

including sentences where the word is in a different form. However, the current example selection system is not fully fitted for MWEs (the system only finds the exact form indicated by the user; discontinuous MWE occurrences and occurrences in inflection variations are not found), it can still be used in MWE entries (see Figure 3, section "Corpus Examples"). Also, example rendering in an entry currently does not support highlighting of all the MWE component-words if they are separated or are in different inflections, opposite to word entries where headwords in examples are highlighted. So far, 6,089 corpus examples have been added to 1,351 MWE entries.

3.5. Links to sources

Since *Tēzaurs* was initially compiled from numerous dictionaries, the name of the dictionary where the entry came from constitutes a part of the entry's meta information. As *Tēzaurs* developed, entries have been merged and refined, so now entries can indicate several sources, and sources can be attributed to lexemes and senses as well. All these features are applicable to MWE entries in the same way as they apply to single-word entries. However, even though links to sources are identical for all entry types, for many phraseological units the situation is more complex, as different sources may list different lexical variants of the MWE that had to be integrated into a single entry not losing the information about the source the lexical variant comes from.

4 MWE search options

By MWE search we understand both looking for the MWE itself and looking for the appearance of the MWE in full sentences – in usage examples and in sense glosses.

In the search function, MWEs have different requirements than single-word entries. There are several different situations where words or MWEs can be searched for: (1) public searches for a dictionary entry (word or MWE), (2) editors' searches in the Tezaurs.lv editing platform for usage examples of MWE occurrences in corpora, (3) various situations in which the editor is looking for an entry or lexeme to edit or link with other entities.

In case of a public search (1) for a single-word entry, *Tēzaurs* supports search by exact match, by match to an inflectional form or by match to a misspelled word (with Levenshtein distance 1), as well as soundex and variations in case. However, these options either do not translate directly to MWEs, or are harder to implement, e.g., for inflections, the capability to inflect the whole MWE would be necessary, and a more sophisticated algorithm would be needed for the search to also identify sparse occurrences and permutations of MWE words in a sentence. Currently the public search option for MWEs in the dictionary requires exact inclusion of all provided words, in the same order, but not necessarily continuous.

The public search (1) is performed differently for words, MWEs, sense glosses and examples, and the results are shown in different sections for matches in each category. In addition, the lexical neighbourhood shows neighbours of the respective type, i.e., other MWEs when looking at an MWE (left column in Figure 3). However, both stemming supported search and syntax tree assisted search are planned as options for future development.

Corpora example search (2) for MWEs does support inflections but requires all words to appear in the usage example in the given order and to be continuous. In this regard, some improvements are possible to at least permit sparse appearances, but a high-quality search should involve the syntactic structure of the example sentence to minimize false positives. This differs

from the public search in *Tēzaurs* (1) as the MWEs themselves are shorter and usually do not have a complex syntactic structure, so treating them as a list of words is sufficient when they are the search target.

In various situations (3) where the editor is looking up an entry or lexeme, the search works on substrings rather than distinct words, which is usually sufficient for such cases, including MWEs. When looking for an MWE occurrence in the sense glosses or in already attached examples, the same considerations regarding inflections and syntactic structure apply.

Another search-related feature is frequency information. For single-word entries, *Tēzaurs* provides information about the frequency per million tokens. Extending this feature to MWEs would require a fast and reliable method for identifying MWE occurrences in corpora data.

5 Conclusions

The paper has shown the potential for MWEs to function well in *Tēzaurs* as independent units that can freely interact with other entries and can be embedded in other (usually single-word) entries. Additionally, the current data shows that MWEs can fully participate in the resource's system of relations (links) between entries. However, further enhancements would be necessary to make representation of MWEs more comprehensive. The next step would be to systematically create different kinds of links specifically for MWE entries, to ensure that they are not underrepresented in this respect compared to single-word entries.

As shown previously, the MWE search experience differs significantly between different use cases. Even if technically not required, it would be beneficial to unify the search functionality between the mentioned scenarios.

Future work would also include improving the function of adding corpus examples for MWEs, adding to the system the capabilities to inflect MWEs, developing morphological and syntactic information of MWEs to improve MWE search and identification in text. The first attempts of syntactical MWE annotation and semi-automated component linking are already underway.

References

Al-Haj, H., Alon, I. & Shuly, W. (2013). Lexical representation of multiword expressions in morphologically-complex languages. *International Journal of Lexicography*, 27, 130–170. https://doi.org/10.1093/ijl/ect036

Bauer, L. (2021). *An Introduction to English Lexicology*. Edinburgh University Press.

Ermakova, M., Geyken, A., Lemnitzer, L. & Roll, B. (2022). Integration of Multi-Word Expressions into the Digital Dictionary of German Language (DWDS). Towards a Lexicographic Representation of Phraseological Variation. In A. Klosa-Kückelhaus, S. Engelberg, C. Möhrs & P. Storjohann (Eds.), *Dictionaries and Society. Proceedings of the XX EURALEX International Congress*, (pp. 851–860). IDS-Verlag.

Fellbaum, C. (1998). *WordNet: An Electronic lexical Database*. MIT Press.

Grasmanis, M., Paikens, P., Pretkalniņa, L., Rituma, L., Stankale, L., Znotiņš, A. & Grūzītis, N. (2023). Tēzaurs.lv – the experiene of building a multufuncional lexical resouce. In M. Medveď, T. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček & S. Krek (Eds.), *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference* (pp. 400–418). Lexical Computing CZ s.r.o.

Klints, A., Grasmanis, M., Nespore-Berzkalne, G., Pretkalnina, L., Stade, M., Gruzitis, N., Lokmane, I., Paikens, P., Rituma, L., Spektors, A. (2024). Tēzaurs as a Digital Multifunctional Lexical Resource. *Baltic Journal of Modern Computing*, 12(4), 513–525. https://doi.org/10.22364/bjmc.2024.12.4.13.

Markantonatou, S., Minos, P., Zakis, G., Moutzouri, V., and Chantou, M. (2019). IDION: A database for Modern Greek multiword expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)* (pp. 130–134). Association for Computational Linguistics.

Mititelu, V. B., Giouli, V., Korvel, G., Liebeskind, C., Lobzhanidze, I., Makhachashvili, R., Markantonatou, S., Markovic, A. & Stoyanova, I. (2025). Survey on lexical resources focused on multiword expressions for the purposes of NLP. In K. O. Atul, G. Voula Giouli, V. B. Mititelu, M. Constant, G. Korvel, A. S. Doğruöz & A. Rademaker (Eds.), *Proceedings of the 21st workshop on Multiword Expressions (MWE 2025)* (pp. 41.–57). Association for Computational Linguistics.

Paikens, P., Klints, A., Lokmane, I., Pretkalniņa, L., Rituma, L., Stāde, M. & Strankale, L. (2023). Latvian WordNet. In G. Rigau, F. Bond & A. Rademaker (Eds.), *12th Global Wordnet Conference, Global Wordnet Association* (pp. 187–196). Global Wordnet Association.

Piasecki, M., Wendelberger, M., & Maziarz, M. (2015). Extraction of the Multi-word Lexical Units in the Perspective of the Wordnet Expansion. In R. Mitkov, G. Angelova & K. Bontcheva (Eds), *Proceedings of the International Conference Recent Advances in Natural Language Processing* (pp. 512–520). INCOMA Ltd. Shoumen.

Rituma, L., Nešpore-Bērzkalne, G., Klints, A., Lokmane, I., Stāde, M. & Paikens, P. (2024). Classifying Multi-Word Expressions in the Latvian Monolingual Electronic Dictionary Tēzaurs.lv. In *6th International Conference Computational Linguistics in Bulgaria (CLIB 2024)* (pp. 113–118). Institute for Bulgarian Language Bulgarian Academy of Sciences.

Saulīte, B., Darģis, R., Grūzītis, N., Auziņa, I., Levāne-Petrova, K., Pretkalniņa, L., Rituma, L., Paikens, P., Znotiņš, A., Strankale, L., Pokratniece, K., Poikāns, I., Barzdins, G., Skadiņa, I., Baklāne, A., Saulespurēns, V. & Ziedin, J. (2022). Latvian National Corpora Collection – Korpuss.lv. In N. Calzolari, F. Béchet, P. Blache, K. Choukri, C. Cieri, T. Declerck, S. Goggi, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, J. Odijk & S. Piperidis (Eds.), *13th Language Resources and Evaluation Conference (LREC)* (pp. 5123–5129). European Language Resources Association.

Andrejs Veisbergs. 2019. The fuzzy concept of idiom and what it might mean for bilingual dictionaries. *Baltic Journal of English Language, Literature and Culture*, 9, 111–129. http://doi.org/10.22364/BJELLC.09.2019.08

Contact information

Gunta Nešpore-Bērzkalne
Institute of Mathematics and Computer Science, University of Latvia
gunta.nespore@lumii.lv

Mikus Grasmanis
Institute of Mathematics and Computer Science, University of Latvia
mikus.grasmanis@lumii.lv

Madara Stāde
Institute of Mathematics and Computer Science, University of Latvia
madara.stade@lumii.lv

Acknowledgements

This work was supported by State Research Programme Letonika – Fostering a Latvian and European Society project “Digital Resources and AI Technologies for the Sustainability of the Latvian Language (DigiLATE)” (VPP-IZM-LETONIKA-2025/1-0004).

Emma Josefin Ölander Aadland

A CRITICAL LEXICOGRAPHICAL DISCOURSE ANALYSIS OF NEW ADJECTIVAL ENTRIES IN THE STANDARD DICTIONARIES FOR NORWEGIAN

Abstract The study presented in the article examines the lemma selection of newly added adjectives in the Norwegian standard dictionaries *Bokmålsordboka* (BOB) and *Nynorskordboka* (NOB). Anchored in *Critical Lexicographical Discourse Studies* (CLDS), the study investigates whether the new adjectives and their sentence examples in the dictionary reflect potential gender-related asymmetries. Methodologically, the new adjectives are classified and analyzed using a socio-semantic category system, combined with concordance analysis of sentence examples to examine patterns of gender representation. The findings show largely comparable patterns across the two dictionaries. There is a clear tendency for new adjectives to be associated with masculine-labelled subcategories versus feminine ones. Furthermore, the sentence examples in the dictionary entries for the new adjectives are predominantly gender-neutral, though a weak tendency towards male overrepresentation is observed, particularly in NOB. Interpreted through a CLDS perspective, the study demonstrates how lexicographic practices related to lemma inclusion may subtly reproduce or cautiously negotiate gendered representations, offering insights relevant to ongoing revision work and lexicographic methodology.

Keywords Gender and Language; Gender Stereotypes; Lemma Inclusion; Lemma Selection; Critical Lexicography; Critical Lexicographical Discourse Studies (CLDS); Gender Critical Corpus Analysis; Socio-semantic Categorization; Concordance Analysis

1 Introduction

The standard dictionaries for Norwegian, *Bokmålsordboka* (BOB) and *Nynorskordboka* (NOB) were thoroughly revised in the years 2018-2024.[1] Both definitions, sentence examples, and fixed expressions were reviewed with the purpose of ensuring that they are in line with contemporary language use (Selback et al., 2025). Furthermore, the revision involved inclusion of new lemmas in both dictionaries with the aim of evening differences between the dictionaries both in lemma selection and structure.[2] The study presented in this paper investigates new adjectives added to the two dictionaries during this revision.[3] The research question is as follows: *Do new adjectival entries added during the revision of the Norwegian standard dictionaries reflect potential gender related asymmetries?*

The present study embraces *Critical Lexicographical Discourse Studies* (CLDS) (Chen, 2019) as its theoretical framework, while the methodological approach for the study involves classifying the newly added adjectives according to a socio-semantic category system further developed from the originally by van Leeuwen (1996). This semantic categorization enables an examination of whether the adjectives convey neutral meanings or are imbued with social and cultural

1 This was the first complete revision since they were printed in 1986 (Rauset 2025, p. 3). The dictionaries have been digital since 1994 (available at https://ordbokene.no/eng) and are jointly owned by the Language Council of Norway and the University of Bergen.

2 Before the revision NOB had a total of 90 000- and BOB 65 000-word entries. After the revision both dictionaries have approximately 100 000-word entries (Rauset, 2025, p. 12).

3 The list of newly added words was generated from the dictionary database 3rd of February 2025 (Standardordbøkene, 2025).

values depending on the context. Additionally, concordance analysis of the adjectives used in the sentence examples in the dictionary entries are conducted to examine gender asymmetries in representation. While the editors explicitly prioritized improving gender representation in sentence examples during the revision process (Selback et al., 2025, p. 142) and a recent study investigating sentence examples in BOB shows a decline in stereotypical portrayals despite some remaining patterns (Ölander Aadland, 2025), no comparable gender critical perspective focusing on gender asymmetry has been applied to newly added words. The present study therefore examines whether structural gender-related patterns emerge in the set of newly added adjectives and discusses implications for lexicographic methodology and revision practice. In doing so, it offers practical insights into ongoing lexicographical work, including suggestions for systematic improvements and consideration of how introducing new words into a dictionary may risk reinforcing gender asymmetries.

The structure of the article is as follows: Section 2 starts with the theoretical framework for the study, followed by a literature review. In section 3, the methodological approach and the data are presented, and in section 4, the findings from the analysis are accounted for. Lastly in section 5, these findings are summarized in a concluding discussion.

2 Theoretical Framework and Previous Research

This section outlines the theoretical framework of the present study, followed by a review of research on neologisms and new entries in dictionaries, with particular attention to gender critical perspectives.

2.1 A backdrop for Critical Lexicographical Discourse Studies (CLDS)

Language is embedded within society, and it shapes and is shaped by the social relationships it expresses. Consequently, language stands in a dialectical relationship with broader societal structures (Fairclough, 2001, p. 19; Skrede, 2017, p. 39). *Critical Discourse Analysis* (CDA), which developed from *Critical Linguistics* (Kress & Hodge, 1979), offers a methodological approach for examining how language functions ideologically and thereby contributes to the maintenance of power relations in society (Skrede, 2017, pp. 11, 20). A defining feature of CDA is its linguistic orientation, which foregrounds systematic analysis of language itself. This emphasis distinguishes CDA from other, more abstract strands of discourse analysis that do not engage directly with linguistic structure or form. This linguistic focus in CDA is further consolidated by Fairclough (Fairclough, 2003). He understands ideology as representations of the world that contribute to establishing and maintaining social power relations (Fairclough, 2003, p. 218), and he highlights the close connection between power, ideology, and language. He also points out that within linguistics, the concept of ideology has been nearly absent from discussions of language and power (Fairclough, 2001, p. 2). Fairclough, who also is viewed as a pioneer for *Critical Language Study* (CLS), outlines that CLS conceptualizes language as a form of social practice, and that social practice in this sense can be called discourse (Fairclough, 2001, p. 34). In his interpretation, discourse may be understood as a concept of "social conditions of production, and social conditions of interpretation" (Fairclough, 2001, p. 20).

Against this backdrop, *Critical Lexicographical Discourse Studies* (CLDS) extends to the lexicographic domain by considering dictionaries as recontextualized discourses that actively construct meaning rather than merely record it (Chen, 2019, p. 368). In this perspective, a dictionary may be interpreted as a lexicographical discourse inherently linked to power and knowledge, and as such dictionaries are representers of both language and society as well as carriers of ideologies (Chen, 2019, p. 371). CLDS originally builds upon critical lexicography

(Kachru & Kahane, 1995), which understands dictionaries as ideologically, socially, and politically embedded rather than neutral artefacts. As Chen (2019) argues, existing research on lexicographical ideology is largely characterized by an internal focus on text. While ideological meanings are identified at the level of textual representation, less attention is paid to the discursive practices of production and use, or to the wider social and institutional conditions shaping these practices. Consequently, the dialectical relationship between lexicographical texts, discursive processes, and social context remains under-theorized. By examining what is included or excluded in a lexicographic discourse, we may deduce and examine any implicit assumptions and ideologies in the dictionary (Chen, 2019, p. 370). This can be related to the distinction between power in discourse and power behind discourse, where the latter is relevant in lexicography, for example in the sense that a dictionary may function as a tool of language standardization (Chen, 2019, p. 369).

The choice of CLDS reflects the aim of the present study, namely, to investigate not only how gendered representations appear in dictionary entries, but also how these representations are shaped by lexicographic practices, editorial decisions, and broader social contexts. While contemporary lexicography is strongly informed by corpus evidence and descriptive principles, lexicographic description necessarily involves editorial interpretation and selection. CLDS is therefore adopted not as an alternative to descriptive lexicography, but as a framework for examining how such interpretative decisions shape the representation of language in dictionaries. This perspective does not assume that such representations are inherently problematic. Rather, it provides analytical tools for examining how they emerge, are maintained, or change across dictionary revisions. In the present study I examine the adjectival lemma inclusion and how the adjectives are presented with sentence examples in the dictionary entries. By applying CLDS as a theoretical framework, I may investigate the ways in which ideological meanings may be embedded in the dictionary content and contextualize these in relation to the production process and the revision practices through which lexicographic choices are negotiated and legitimized. On this basis, the present study approaches newly added adjectives as potential sites of ideological negotiation, where gendered representations may be reproduced or contested. Emphasizing the socially embedded nature of lexicographic practice, this perspective prepares the ground for a discussion of the dictionary's dual role as both a linguistic tool and an authoritative artefact.

2.2 *The Dictionary as Both a Tool and an Authoritative Product*

Dictionaries have a dual position as both practical tools for language users and authoritative cultural artefacts, as people tend to view dictionaries as an authoritative cultural product (Nossem, 2018, p. 174). Understanding the dictionary's dual role is relevant when applying CLDS as a theoretical framework, because it connects everyday language use to the reproduction of broader ideological patterns. Furthermore, this normative and ideological status becomes evident in how dictionaries are commonly perceived as authoritative sources that define which words are acceptable in the language and prescribe their appropriate use (Atkins, 2008, p. 2). In other words, this states that dictionaries are "powerful tools for transmitting power and ideology" (Fuertes-Olivera & Tarp, 2022, p. 110). The dictionary can also be interpreted as a tool for users to solve problems of linguistic nature (ten Hacken, 2020, p. 146). This understates that the work of lexicographers is powerful both linguistically and culturally, and Nossem (2018, p. 184) even claims that a dictionary is "the product of a series of biased decisions by the lexicographers". Conversely, lexicographers often conceive their role primarily as descriptive rather than normative (ten Hacken, 2020, p. 137). This goes in line with the fact that lexicographers depend on actual language use to formulate sentence examples, but also when it comes to evidence of actual use before adding new words to the dictionary.

In this regard, a study by Nossem (2018) reveals a heteronormative discourse present in dictionaries, for example in the sense that lexicographers tend to choose the singular masculine form in languages with flectional adjectives. According to Nossem (2018, p. 177), this reflects an arbitrary practice of decision making that comes from the dominant discourse in society that favorizes the male perspective and, in this context, the male forms. Nossem hereby states: "The lexicographer is not less a "harmless drudge" but a powerful (re-)producer of linguistic, cultural, and social power structures" (Nossem, 2018, p. 185).

A dictionary is in other words expected to provide reliable guidance for everyday linguistic problem solving, but at the same time it contributes to the maintenance of ideological structures by reinforcing conceptualizations of language and society. Recognizing this dual function is particularly relevant for the present study, as newly added adjectives may reflect or reinforce underlying gendered assumptions or norms. The next section describes the process of inclusion of new words in dictionaries and touches upon considerations regarding gender-inclusive language.

2.3 Gender-inclusive Language in Dictionaries

The inclusion of new lemmas in dictionaries is traditionally shaped both by empirical language use and by prevailing linguistic norms. Mugglestone (2013) analyses how early editions of the Oxford English Dictionary (OED) negotiated gender during a period of marked ideological and semantic change, focusing on the "Woman Question" and the contested meanings of terms such as *New Woman and suffragette*. She highlights how shifting editorial boundaries, omissions, and interpretive practices shaped the dictionary's treatment of gender across its history. Ten Hacken (2020) underlines how the inclusion process of new words in a dictionary depends on both empirical reality and language norms. Furthermore, he argues that dictionaries operate as gatekeepers in relation to neologisms, as users treat them as authoritative reference points for language use (ten Hacken, 2020, p. 147). Neologisms typically designate emerging concepts, drawing on processes such as compounding and derivation, semantic extension, and different types of borrowing (see e.g. Trap-Jensen, 2020a). Although lexicographers generally strive for a descriptive orientation, users frequently expect dictionaries to act as normative authorities (ten Hacken, 2020, p. 147; Hovmark & Svardal, 2024, p. 9).

During the last decades, the lexicographical practice of neologism inclusion has relied heavily on corpus and frequency-based methods. Thus, while users may expect dictionaries to indicate which words "exist" in the language, lexicographers often shift this responsibility onto corpus evidence (ten Hacken, 2020, p. 146). However, no corpus is value-neutral, and there might be an issue with for example gender biases in a corpus (Chen, 2019, p. 371). While lexicographers build the content of a dictionary based on corpus evidence, this might in fact result in a compromise with ideology. A frequency and corpus-based lemma selection can in this sense be seen as problematic because it disables the possibility to avoid power or ideology influencing the selection process (Fuertes-Olivera & Tarp, 2022, p. 112). Mainly because it can result in different asymmetries, such as gender bias or prioritizing written language over spoken language. Fuertes-Olivera and Tarps investigate the role of lexicographers regarding power imbalances in the relation between language and gender by studying gendered practices in general dictionaries of Spanish is analyzed. By adapting a CLDS framework, they investigate how professional titles in Spanish refer to male and female professionals in both headwords and definitions. Amongst the findings are that women are mostly absent from general dictionaries of Spanish (Fuertes-Olivera & Tarp, 2022, p. 120). Furthermore, the masculine form is the dominant lemma form for nouns, while the feminine version is only a form of this lemma (ibid). The study also indicates that Spanish lexicographers unintentionally contribute to the invisibility of women in the

dictionary as well as making them visible only by referring to them as "the wife of a professional man" (Fuertes-Olivera & Tarp, 2022, p. 128). Additionally, two other recent studies find that corpus-based dictionaries contribute to reproducing gender stereotypes due to the biases existing in the underlying corpora (Müller-Spitzer & Rüdiger, 2022; Müller-Spitzer, 2023). The dictionaries thereby reproduce existing gender biases, such as portraying men as active agents and women in more passive or relational roles, even though the lexicographic aim is to reflect actual language use. While most studies of neologisms rely on corpus–based methods for data collection, the study by Foubert (2023) uses crowdsourced dictionaries for identifying contemporary feminist neologisms. This study is theoretically within a *Feminist Linguistic Activism* (FLA) framework, with a focus on gender-fair language understood with women's structural subordination and men's structural dominance. Among the main findings is that many neologisms related to gender and sexuality both have shaped earlier feminist movements and still influence contemporary neologisms.

The lemma selection in dictionaries can also provoke reactions from dictionary users, which was the case in relation to a recent revision of *Den Danske Ordbog* (DDO) (Trap-Jensen, 2020b). This underscores that a corpus cannot be taken as a fully accurate representation of the language and that a specific gender critical approach in the lemma selection process might be crucial (Trap-Jensen, 2020b, p. 144). This is also stressed by Fjeld, who problematizes gender-related asymmetries in lemma lacunae (Fjeld, 2001, 2015). A gender-critical approach to Scandinavian dictionaries has also been adopted in several other studies (e.g. Hageberg, 1990; Nikula, 1997; Pearce, 2008; Hult, 2010; Norberg, 2016; Larsson, 2017; Sköldberg, 2020, 2022; Ölander Aadland, 2025). These studies reveal subtle forms of gender asymmetries, including the underrepresentation of lemmas related to traditionally female-coded domains and the prevalence of stereotypical gender portrayals in sentence examples. Although these studies do not primarily focus on neologisms or newly introduced entries, they nevertheless reveal gender-related asymmetries both in lemma selection and in illustrative examples. Additionally, there are other studies with a gender-inclusive language scope, as for example the recent study of the process concerning the inclusion of the gender-neutral pronoun ´hen´ in the two standard dictionaries for Norwegian (Bugge et al., 2026). The present study contributes to new empirical insights into gendered language in two national dictionaries of Norwegian and may thereby raise awareness of gender related representational dynamics in lexicographic practices.

3 Methodology and Data

In this section, the methodological approach for the present study is accounted for, consisting of socio-semantic categorization and concordance analysis. This is followed by a presentation of the data and necessary delimitations made for the purposes of the study.

3.1 Methodological Approach

The first part of the analysis in the study is conducted with a semantic categorization as a methodological approach, while the second part applies concordance analysis with an aim to investigate sentence examples for the new adjectives in the dictionary entries to detect any gender asymmetries. Semantic categorization involves classifying the newly added adjectives in line with their semantic meaning, which allows an investigation of the kinds of attributes and evaluations the adjectives construe as well as the social functions and contexts in which they operate. In the scope of the present study, the specific interest is in potential gendered language and asymmetries in gender representation. As earlier lexicographic research of both corpora and dictionaries reveal (see e.g. Pearce, 2008; Norberg, 2016; Hu et al., 2019; Flörke & Engberg-Pedersen, 2021; Müller-Spitzer & Rüdiger, 2022; Müller-Spitzer, 2023;

Holtås, 2023; Ölander Aadland, 2025; Klosa-Kückelhaus & Kernerman, 2020), there are gender stereotypical patterns particularly evident in collocations and example usage. These contribute to the reproduction of gender-stereotypical representations at a semantic level, where women typically are associated with adjectives from semantic fields describing marital status, relational roles, nationality or ethnic background, religious affiliation, reproductive characteristics, attractiveness, physical appearance and mood or emotional state. Male gender referents, by contrast, tend more often to co-occur with adjectives denoting equipment or clothing, sexual orientation, physical activity and strength, violent behavior and criminality, agency, moral evaluation, ability, behavioral traits and social status. In the context of the present study, these gender stereotypical tendencies are used as a backdrop to divide the semantic categories applied in the analysis into gender-typical semantic domains.

The categorization system used in the present study builds on a socio-semantic category system that includes the three major categories *Functionalization* (e.g. an occupation or role), *Identification* (e.g. interpersonal relations, ethnicity or physical appearance) and *Appraisement* (e.g. intellect, morality or behavioral traits). The backdrop for this category system is van Leeuwen's classification of social actors in text (van Leeuwen, 1996), and it has been applied in several studies investigating gender portrayal with adjectival use (see e.g. Caldas-Coulthard & Moon, 2010; Moon, 2014; Holtås, 2023; Ölander Aadland, forthcoming). Some necessary adjustments have been made for the present study, as shown in an overview in Tab. 1.

TABLE 1 The Semantic Categorization System and Associated Examples

Classifying	Physical and health	Appraisement
Sex (e.g. trans, *kjønnsdelt* 'gender-segregated')	Life and Death (e.g. *suicidal*, *knivdrept* 'stabbed to death')	Mood and Emotional State (e.g. *alvorstynget* 'grave', *lettrørt* 'easily moved')
Age (e.g. *yngre* 'younger')	Health (e.g. *aidssyk* 'AIDS sick', *febernedsetttende* 'fever reducing')	Intelligence (e.g. *reflektert* 'reflected', *streetsmart*)
Sexual Orientation (e.g. *panseksuell* 'pansexual', *queer**)	Pregnancy (e.g. *flergangsfødende* 'given birth several times')	Morality (e.g. ego, *hevntørst* 'vengeful')
Origin (e.g. *sentraleuropeisk* 'Central European', *vermlandsk* 'from Vermland')	Sexual Desire (e.g. *kinky*, *voyeuristisk* 'voyeuristic')	Ability (e.g. *hardtarbeidende* 'hardworking', *lesekyndig* 'literate')
Ethnicity/Nationality (e.g. *umesamisk* 'Ume Sami', *tatarisk* 'Tatar')	Appearance (e.g. *oppsvulmet/oppsvulma* 'swollen', *mellomblond* 'medium blonde')	Behavioral Traits (e.g. *medfølende* 'compassionate', *sjefete/sjefet* 'bossy')
Occupation (e.g. *agrar* 'agrarian', *korttidsansatt* 'temporary employee')	Attractiveness (e.g. *barmfager* 'buxom', *snasen* 'appealing')	Interest (e.g. *danseglad* 'dance-loving', *dedikert* 'dedicated')
Lifestyle (e.g. emo, *bohemaktig* 'bohemian')	Skin Colour[a] (e.g. *lyshudet/lyshuda* 'light-skinned', *melaninrik*** 'melanin-rich')	Personal Trait (e.g. *drittkul/dritkul* 'awesome', *speisa* 'spaced')
Religion and Belief (e.g. *kristenkonservativ* 'Christian conservative', *kantiansk* 'Kantian')	Equipment (e.g. *møblert* 'furnished', *væpnet/væpna* 'armed')	General Attribute (e.g. *illesmakende* 'tasting bad', *kontrastfylt* 'contrasted')

TABLE 1 (*CONTINUED*)

Classifying	Physical and health	Appraisement
Social Status (e.g. *høykulturell/høgkulturell* 'highly cultured', *verdensledende* 'world-leading')	**queer* is also categorized under "Sex" ***melaninrik* 'melanin-rich' is also categorized under "Health"	Relational (e.g. *nyskilt* 'recently divorced', *polyamorøs* 'polyamorous')
Politics and Society (e.g. *høyreradikal/høgreradikal* 'far-right', *kulturpolitisk* 'cultural policy')		
Colour (e.g. *rødgul* 'red-yellow', *sandfarget/sandfarga* 'sand-colored')		
General (e.g. *bakovervendt* 'rear-facing', *dagfersk* 'day-fresh')		

[a]Skin colour could also have been placed under "Classifying". However, such a categorization is more context-dependent, and I have chosen to assign each subcategory to only one main category for the sake of analytical clarity and simplicity.

As shown in Tab. 1, the three main categories used in the present study are *Classifying, Physical and Health and Appraisement*. Furthermore, these main categories have subcategories; 12 under *Classifying*, 8 under *Physical and Health*, and 9 under *Appraisement*. Several adjectives have different meanings, and these have therefore been categorized with several labels from both main categories and subcategories. For instance, the adjective *kjempestor* 'huge' has been assigned to both the subcategory *General under Classifying* and the subcategory *Appearance* under *Physical and Health*.

3.2 The Datasets and Adopted Delimitations

The extracted datasets, one for each dictionary, contained all new words added to the dictionaries in connection with the comprehensive revision. For the purposes of the present study, the datasets were limited to newly added adjectives, amounting to a total of 1,173 adjectives in NOB and 1,125 adjectives in BOB. The datasets were further delimited to adjectives marked as "published" to enable concordance analyses of sentence examples in the dictionary entries.[4] Furthermore, the semantic categorization was done manually in Excel, and during this analysis some errors in the data material were detected.[5] After these erroneous entries were excluded, the material comprises 926 new adjectives in BOB and 966 in NOB. As mentioned, the inclusion of new entries in both dictionaries was a part of the revision project, and the main aim with the new entries was to even differences both in lemma selection as well as the structure in the dictionaries (Rauset, 2025, p. 12). The list of new adjectives in each dictionary is therefore largely overlapping, but there are 97 unique adjectives in the dataset for NOB and 59 in BOB.

[4] As the lexicographers work on the dictionaries on an ongoing basis, changes may have occurred with respect to which entries are published after the dataset was extracted on the 3rd of February 2025.

[5] The following were removed from the datasets before the analysis: The noun *melkeskum/mjølkeskum* 'milk foam' in the BOB dataset and the noun *leikedyr* 'toy animal' in NOB. The suffixes *-fast* in both datasets and *-hjarta* in NOB were also removed, because they do not constitute independent adjectives. Moreover, *solariumbrun* 'tanning bed tan' occurs in both datasets, but this word does not appear in the published dictionaries, which only attest to the form *solariebrun. Solariumbrun* has therefore been excluded from the datasets. In addition, two cases of duplicates are merged in the datasets: *dominikansk* 'Dominican' in both datasets and *yrjande/yrande* 'bustling' in NOB (this adjective has one shared dictionary entry but appeared as two separate entries in the dataset).

4 Findings

This section opens with a presentation of the findings from the socio-semantic categorization. This is followed by a review of the findings from the concordance analysis.

4.1 Findings from the Semantic Categorization

Based on their semantic meaning, all new adjectives were first assigned to one or several of the three overarching semantic categories: *Classifying, Physical and Health*, and *Appraisement*. As shown in Tab. 2, the distribution across these categories reveals only minor differences between BOB and NOB. In both dictionaries, most adjectives fall within the main category *Classifying* (645 in NOB vs. 611 in BOB), followed by *Appraisement* (291 in NOB vs. 292 in BOB) and *Physical and Health* (171 in NOB vs. 169 in BOB). Given the highly general nature of these main categories, the subcategorization results provide a more informative basis for interpretation. The following section therefore presents the outcomes of this analysis.

TABLE 2 Frequencies for Each Main Category in NOB and BOB

Socio-Semantic Category	NOB	BOB
CLASSIFYING	645	611
APPRAISEMENT	291	292
PHYSICAL AND HEALTH	171	169

4.1.1 Findings from the Categorization in Socio-Semantic Subcategories

The semantic categorization of the adjectives in subcategories shows small differences between the two dictionaries, as illustrated in Fig. 1 and Fig. 2.

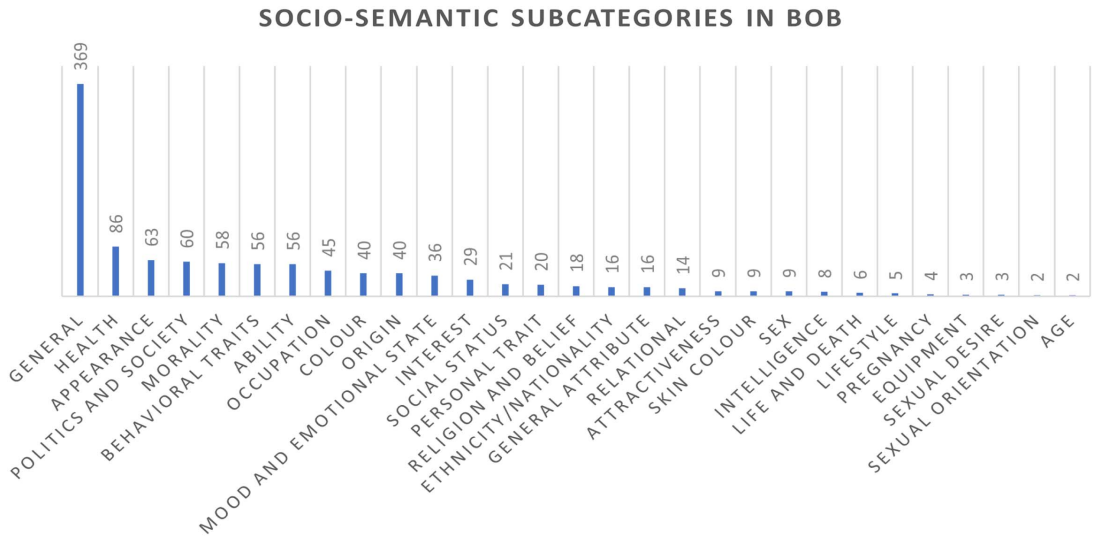


FIGURE 1 An Overview of Frequencies for the Socio-Semantic Subcategories in BOB

Among the subcategories that contain more than 50 adjectives, we find *General, Health, Politics and Society, Appearance, Behavioral Traits, Morality* and *Ability* in both dictionaries. As outlined in *Section 3.1*, findings from earlier studies on gender-stereotypical adjectival usage in corpora and dictionaries serve as a framework for categorizing gender-typical socio-semantic domains.

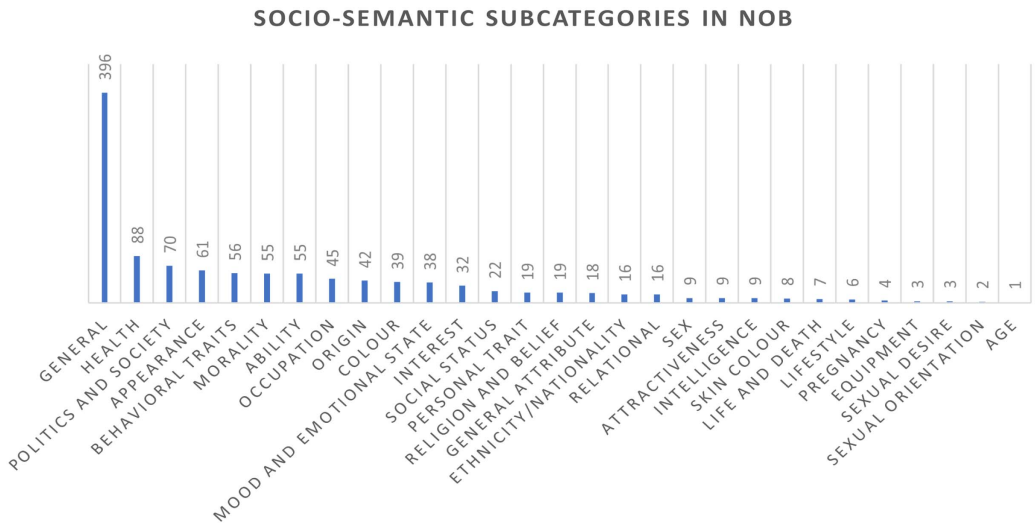


FIGURE 2 An Overview of Frequencies for the Socio-Semantic Subcategories in NOB

Tab. 3 shows that there are in total 10 subcategories considered more MASC-typical domains and 8 subcategories more FEM-typical domains, while the remaining 11 subcategories are considered neutral. Although the differences in number of subcategories assigned to each gender category are small, calculating the relative proportion of adjectives in total for each category indicates considerable differences between the subcategories labelled as MASC versus FEM in both dictionaries. Among the new adjectives, in total 29 % in BOB and NOB are labelled in MASC-typical semantic subcategories while 15 % in both BOB and NOB belong to FEM-typical semantic subcategories. The remaining adjectives assigned to NEUTRAL-typical subcategories are 56 % in both dictionaries. Overall, the findings thereby point to potential gender asymmetries in new adjectival entries in both dictionaries, reflected in a higher number of new adjectives assigned to typical MASC-domains than FEM-domains. However, since the analysis is limited to adjectives introduced during the revision process, it remains unclear whether this pattern is representative of the complete set of adjectival entries in the dictionaries. Furthermore, this does not tell us if the adjectives are used specifically in connection with either gender in the dictionary entries. To reveal further gendered asymmetries in each dictionary as well as the dictionaries in comparison, concordance analyses of sentence examples in the entries for the new adjectives have been conducted. The findings are presented in the following section.

TABLE 3 Gender-typical Tendencies for the Semantic Subcategories

Gender-category	Gender-typical Subcategories in BOB and NOB	Adjectives in relative proportion (BOB/NOB)
MASC	Politics and Society Morality Ability Behavioral Traits Occupation Social status General Attribute Intelligence Sexual Orientation Equipment	29 % BOB/29 % NOB

TABLE 3 (*CONTINUED*)

Gender-category	Gender-typical Subcategories in BOB and NOB	Adjectives in relative proportion (BOB/NOB)
FEM	Appearance Ethnicity/Nationality Relational Mood and Emotional State Attractiveness Pregnancy Sexual Desire Religion and Belief	15 % BOB/15 % NOB
NEUTRAL	General Health Origin Colour Interest Personal Trait Age Sex Skin Colour Life and Death Lifestyle	56 % BOB/56 % NOB

4.2 Findings from the Concordance Analysis of Gender-use in Sentence Examples

When conducting the concordance analysis, a set of predefined categorization criteria were applied. Firstly, when the entry for an adjective only has one sentence example and this lacks gender, it was labelled with 'Gender-neutral' (e.g. the adjective *stupfull* 'very drunk' in BOB has the sentence example "båtføreren var stupfull" 'the boat driver was very drunk'). Secondly, 'Gender-asymmetry' was used when only one gender occurs and the entry has only one sentence example (e.g. for the adjective *småfet/småfeit* 'chubby' we find "en småfet mann" 'a chubby man'). Thirdly, many of the entries for the new adjectives in BOB and NOB include several sentence examples, and when some of these examples include gender, they were labelled as 'Partially gender-neutral'. In some cases, the sentence examples in the same entry include both male and female, and these were therefore also given the same label (e.g. the adjective *kontrollerande* 'controlling' in NOB has the example "ho har ein kontrollerande far" 'she has a controlling father'). Moreover, the sentence examples labelled with 'Partially gender-neutral' show cases where the dictionaries together give a gender-neutral representation by having male in a sentence example for the specific entry in one dictionary and female in the other dictionary, as illustrated in Fig. 3 with the adjective *kontoransatt/kontortilsett* 'office employee' that has the example "den kontoransatte kvinnen" 'the office-employed woman' in BOB and "kontortilsette menn" 'office-employed men' in NOB. Lastly, the label 'Lacking examples' shows entries that lack sentence examples. To be able to calculate gender representation in total, the gender use in the sentence examples that are marked as both 'Partially gender-neutral' and 'Gender-asymmetry' were examined.[6] The relative frequency of each gender

[6] Note that sentence examples labelled as 'Partially gender-neutral' may have both genders included (e.g. the adjective *granskende* 'scrutinizing' in BOB has the sentence example "hun kastet et granskende blikk på ektemannen før hun gikk videre" 'she took a scrutinizing glance at her husband before moving on'), and these are accounted for.

category was calculated as the number of occurrences in the category divided by the total number of occurrences. The results from these concordance analyses are shown in Tab. 4.



FIGURE 3 An Example of 'Partially gender-neutral' in the two dictionaries

TABLE 4 Results from the Concordance Analysis of Gender-use in Sentence Examples

Gender-use	*Nynorskordboka* (NOB)	*Bokmålsordboka* (BOB)
Gender-neutral	68 %	68 %
Partially gender-neutral	7 %	5 %
Gender-asymmetry	3 %	4 %
Gender representation relatively	FEMALE 44%/MALE 56 %	FEMALE 47 %/MALE 53 %
Lacking sentence examples	22 %	23 %

As shown in Tab. 4, the use of gender across the examined sentence examples in the two dictionaries are broadly comparable. The gender-neutral sentence examples in both dictionaries are the largest group with a total of 630 (68 %) in BOB and 658 (67 %) in NOB. Furthermore, there are similar tendencies regarding partially gender-neutral examples in the two dictionaries; in BOB 50 (5 %) have this label and 68 (7 %) in NOB. The gender-asymmetry examples in BOB are in total 35 (4 %) and 27 (3 %) in total in NOB. In BOB the gender representation in 85 examples (partially gender-neutral + gender-asymmetry in total) is 48 times with male gender and 43 times with female. In NOB the gender representation in 95 examples (partially gender-neutral + gender-asymmetry in total) is 61 times with male gender and 47 times with female. The relative frequency of each gender in BOB is FEMALE 47 % and MALE 53 % and respectively in NOB is FEMALE 44 % and MALE 56 %. This indicates a weak tendency towards overrepresentation of the male gender in the sentence examples for new adjectives in both dictionaries, and the most pronounced deviation is in NOB. Lastly, there are in total 211 (23 %) entries in BOB and 213 (22 %) in NOB lacking sentence examples.

5 Concluding Discussion

The study presented in this article examines gender representation in new adjectival entries in the Norwegian dictionaries *Bokmålsordboka* (BOB) and *Nynorskordboka* (NOB). The findings in the socio-semantic categorization of new adjectives reveal only minor differences between the two dictionaries. This outcome is hardly surprising, as aligning the dictionaries was a central aspect of the revision work. However, the analysis shows that there are more MASC-labelled subcategories than FEM. Additionally, when comparing the relative proportions across the two

dictionaries, differences between MASC- and FEM-labelled subcategories are substantial (29 % versus 15 %). Taken together, the results point to gender asymmetries in both dictionaries, insofar as a greater number of the new adjectives are assigned to MASC-labelled subcategories. Furthermore, the concordance analysis shows that gender–neutral examples constitute the majority in both dictionaries (68% in BOB and 67% in NOB). Partially gender–neutral examples form a small share (5% in BOB and 7% in NOB), while explicitly gender–asymmetrical examples are comparatively rare (4% in BOB and 3% in NOB). A slight imbalance appears when considering relative gender distribution: in BOB, male and female referents account for 53% and 47% respectively, and in NOB the imbalance is somewhat stronger (56% versus 44%). This indicates a weak tendency toward male overrepresentation in sentence examples for the new adjectives, most clearly in NOB. A substantial proportion of entries lack sentence examples altogether (23% in BOB and 22% in NOB). Overall, the two dictionaries display largely comparable gender usage with only minor, systematic deviations.

Viewed through the lens of CLDS, the findings of the present study support the understanding of dictionaries not as neutral repositories of language, but as cultural artefacts that actively construct meaning (Chen, 2019). At the same time, the present study does not question the importance of corpus-based and descriptive lexicography. Rather, it examines how corpus evidence is interpreted, selected, and represented through lexicographic practice. Decisions regarding lemma selection, sense division, example choice, labelling, and revision necessarily involve interpretation. From a CLDS perspective, these interpretative processes constitute a relevant object of study. As authoritative reference works, dictionaries both reflect and shape social knowledge, and in doing so they may reproduce implicit assumptions and power-related ideologies, including gender norms. While dictionaries play an important role in preserving linguistic change, they also risk perpetuating established patterns of representation if such patterns are not critically examined. The introduction of new lemmas is particularly sensitive to such processes. As Mugglestone (2013, pp. 47–48) notes, lemma selection and representation inevitably interact with dominant ideologies and prevailing cultural norms. Additionally, the linguistic representation of gender in a dictionary that is corpus-based depends on the diversity of the empirical base (Müller-Spitzer & Rüdiger, 2022, p. 139). The present study shows that new adjectival entries in the two Norwegian dictionaries exhibit minor but systematic gender asymmetries. Although the revealed patterns are relatively small, they nonetheless indicate how lexicographic practices related to lemma selection may contribute to gender-related differences.

At the same time, the concordance analysis may reflect editorial choices that result in a more balanced distribution of gender representations, in line with the principles of descriptive lexicography (Mugglestone, 2013, pp. 49–50). The increasing use of multiple sentence examples per entry, while primarily motivated by the aim of illustrating variation in use, also has the effect of promoting a more balanced gender representation. Moreover, variation in gender usage across parallel entries, where one dictionary employs a masculine referent in an example and the other dictionary a feminine referent for the same lemma, further contributes to an increasing balance in gender representation. From a CLDS perspective, these practices can be seen as forms of lexicographic awareness that acknowledge the discursive power of dictionaries and their potential role in either reinforcing or counteracting gender asymmetry. Rather than signaling a fully neutral representation, the findings point to an ongoing negotiation between descriptive ideals and institutional authority, as well as an awareness of social norms. In this sense, the dictionaries examined illustrate how contemporary lexicography may both subtly reproduce and cautiously challenge gendered discourses, especially in the treatment of new entries. The purpose of the present study is therefore not to assess whether dictionary content should mirror social reality more or less faithfully, but to examine how gendered representations emerge within lexicographic discourse and how these representations change

across revisions. The issue, therefore, is not whether dictionaries can or should be neutral, but how they may balance the documentation of language use with an awareness of the ideological effects such use may entail. A fully balanced representation may, in some cases, risk obscuring empirically attested asymmetries, whereas an uncritical reproduction of these asymmetries may contribute to their perpetuation. The challenge thereby lies in developing a lexicographic practice that is both empirically grounded and critically informed.

References

Atkins, S. B. T. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press.

BOB = *Bokmålsordboka*. Retrieved May 2026, from Web version (1994–) https://ordbokene.no/ The Language Council of Norway and The University of Bergen.

Bugge, E., Bukve, T. & Myklebust, H. (2026). Labeling Gender Across and Beyond the Binary: Introduction of the Gender-Neutral Pronoun Hen in Norwegian Standard Language. *Journal of Language and Sexuality*, 15(1), 50–72. https://doi.org/10.1075/jls.24033.bug

Caldas-Coulthard, C. R. & Moon, R. (2010). 'Curvy, hunky, kinky': Using Corpora as Tools for Critical Analysis. *Discourse & Society*, 21(2), 99–133. https://doi.org/10.1177/0957926509353843

Chen, W. (2019). Towards a Discourse Approach to Critical Lexicography. *International Journal of Lexicography*, 32(3), 362–388. https://doi.org/10.1093/ijl/ecz003

Fairclough, N. (2001). *Language and Power* (2nd edn). Longman.

Fairclough, N. (2003). *Analysing Discourse: Textual Analysis for Social Research*. Routledge.

Fjeld, R. V. (2001). Ingenting å snakke om – Kvinneord og ordbøker. *Språknytt*, 29(4), 10–13.

Fjeld, R. V. (2015). Eksempler og stereotypisering av kjønn i noen nordiske ordbøker. *Perspektiv på lexikografi, grammatik och språkpolitik i Norden, Institutet för de inhemska språkens webbpublikationer*, 39, 35–65.

Flörke, I. & Engberg-Pedersen, E. (2021). Køn i korpus. Semantisk asymmetri i brugen af ordene kvinde og mand, pige og dreng i danske avistekster. *NyS, Nydanske Sprogstudier*, (60), 37–82. https://doi.org/10.7146/nys.v1i60.129349

Foubert, O. (2023). Neologisms in Contemporary Feminisms: For a Redefinition of Feminist Linguistic Activism. *Glad!*, 15, 1–26. https://doi.org/10.4000/glad.7722

Fuertes-Olivera, P. A. & Tarp, S. (2022). Critical Lexicography at Work: Reflections and Proposals for Eliminating Gender Bias in General Dictionaries of Spanish. *Lexikos*, 32(2), 105–132. https://doi.org/10.5788/32-2-1699

Hageberg, A. (1990). Kvinner og menn og ordbøker. *Norsk Lingvistisk Tidsskrift*, 8, 48–67.

Holtås, V. S. (2023). *Deilige damer, maskerte menn: Ei triangulerande undersøking av adjektivattributt til dame, kvinne og mann i Norsk aviskorpus (bokmål)* [Master thesis, University of Bergen]. https://bora.uib.no/bora-xmlui/handle/11250/3071222

Hovmark, H. & Svardal, T. (2024). Hvilket datamateriale bygger nordiske ordbøger på? Skævheder, udfordringer og løsninger. *LexicoNordica*, (31), 7–16. https://doi.org/10.7146/ln.v1i31.151282

Hu, H., Xu, H. & Hao, J. (2019). An SFL Approach to Gender Ideology in the Sentence Examples in the *Contemporary Chinese Dictionary*. *Lingua*, 220, 17–30. https://doi.org/10.1016/j.lingua.2018.12.004

Hult, A.-K. (2010). Kort och gott. Om idiomens språkprov i Svensk ordbok utgiven av Svenska Akademien (2009). In H. Lönnroth & K. Nikula, *Nordiska studier i lexikografi 10. Rapport från Konferensen om lexikografi i Norden* (pp. 209–222).

Kachru, B. B. & Kahane, H. (1995). *Cultures, Ideologies, and the Dictionary: Studies in Honor of Ladislav Zgusta*. De Gruyter.

Klosa-Kückelhaus, A. & Kernerman, I. (2020). Global Viewpoints on Lexicography and Neologisms: An Introduction. *Dictionaries: Journal of the Dictionary Society of North America*, 41(1), 1–9.

Kress, G. & Hodge, B. (1979). *Language as Ideology*. Routledge & Kegan Paul.

Larsson, L. (2017). Vilken plats hade kvinnan?: Om grammatiskt och sexuellt genus i SAOL:s sjunde upplaga 1900. *Språk och stil (Uppsala, Sweden: 1991)*, NF 27, 49–66.

Moon, R. (2014). Meanings, Ideologies, and Learners' Dictionaries. *Proceedings of the XVI EURALEX International Congress* (pp. 85-105). The User in Focus. 15-19 July 2014. Institute for Specialised Communication and Multilingualism.

Mugglestone, L. (2013). Acts of Representation: Writing the Woman Question in the Oxford English Dictionary. *Dictionaries: Journal of the Dictionary Society of North America*, 34(1), 39–65.

Müller-Spitzer, C. (2023). Gender Stereotypes in Dictionaries: The Challenge of Reconciling Usage-based Lexicography with the Role of Dictionaries as Social Agents. *Lexikos*, 33(2), 79–94. https://doi.org/10.5788/33-2-1843

Müller-Spitzer, C. & Rüdiger, J. O. (2022). The Influence of the Corpus on the Representation of Gender Stereotypes in the Dictionary. A Case Study of Corpus-Based Dictionaries of German. In *Proceedings of the XX EURALEX International Congress* (pp. 129–141).

Nikula, K. (1997). Språkprov i könsperspektiv. In *Fackspråk och översättningsteori. VAKKI-Symposium XVII* (pp. 194–206).

NOB = *Nynorskordboka*. Retrieved May 2026, from Web version (1994–) https://ordbokene.no/ The Language Council of Norway and The University of Bergen.

Norberg, C. (2016). Naughty Boys and Sexy Girls: The Representation of Young Individuals in a Web-Based Corpus of English. *Journal of English Linguistics*, 44(4), 291–317. https://doi.org/10.1177/0075424216665672

Nossem, E. (2018). Queering Lexicography: Balancing Power Relations in Dictionaries. In *Queering Translation, Translating the Queer* (1st edn, pp. 172–187). Routledge. https://doi.org/10.4324/9781315505978-13

Pearce, M. (2008). Investigating the Collocational Behaviour of Man and Woman in the BNC Using Sketch Engine. *Corpora, 3*(1), 1–29. https://doi.org/10.3366/E174950320800004X

Rauset, M. (2025). *Sluttrapport. Revisjon av Bokmålsordboka og Nynorskordboka – Revisjonsprosjektet*. Universitetet i Bergen. https://hdl.handle.net/11250/3187459

Selback, B., Sjo, K., Karlsen, K. E. & Berg-Olsen, S. (2025). Revisjonen av Bokmålsordboka og Nynorskordboka: Oppdatering og tilpasning av ordbøkene til en ny tid. *LexicoNordica*, (32), 131–153. https://doi.org/10.7146/ln.v1i32.160906

Sköldberg, E. (2020). En mer inkluderande lexikografi – bortom frågan om kvinnor och män. *Oslo Studies in Language*, 11(1), 7–30. https://doi.org/10.5617/osla.8165

Sköldberg, E. (2022). Andra upplagan av Svensk ordbok: Förutsättningar och redaktionella val. *LexicoNordica*, (29), 139–152. https://tidsskrift.dk/lexn/article/view/134779

Skrede, J. (2017). *Kritisk diskursanalyse*. Cappelen Damm akademisk.

ten Hacken, P. (2020). Norms, New Words, and Empirical Reality. *International Journal of Lexicography*, 33(2), 135–149. https://doi.org/10.1093/ijl/ecaa005

Trap-Jensen, L. (2020a). Language-Internal Neologisms and Anglicisms: Dealing with New Words and Expressions in The Danish Dictionary. *Dictionaries: Journal of the Dictionary Society of North America*, 41(1), 11–25.

Trap-Jensen, L. (2020b). Inklusion eller mindretalsdiktatur? Om politisk korrekthed, minoritetshensyn og leksikografisk deskriptivisme i Den Danske Ordbog. *LexicoNordica*, (27), 137–156. https://doi.org/10.7146/ln.vi27.123220

van Leeuwen, T. (1996). The Representation of Social Actors. In M. Coulthard & C. R. Caldas-Coulthard (Eds), *Texts and Practices: Readings in Critical Discourse Analysis* (pp. 32–70). Routledge.

Ölander Aadland, E. J. (2025). Kjønnsrepresentasjon og -fremstilling i Bokmålsordbokas brukseksempel. *LexicoNordica*, (32), 197–219. https://doi.org/10.7146/ln.v1i32.160909

Contact information

Emma Josefin Ölander Aadland
University of Bergen
emma.aadland@uib.no

Linda Prossliner

AI-SUPPORTED PHRASEOGRAPHY FOR YOUNG LEARNERS

Designing Definitions and Illustrations through AI – Opportunities and Challenges

Abstract This paper presents a case study exploring the potential of Large Language Models for learner-oriented phraseography through the generation of COBUILD-style Full-Sentence Definitions and AI-generated dual illustrations representing both the literal and phraseological meanings of idioms. The study is situated within *Phrasolino*, an electronic Italian-German phraseological learner's dictionary for German-speaking children aged 7–11 learning Italian as an L2. A corpus-based dataset of 100 Italian idioms containing the body-part noun *testa* ('head') was compiled from the Italian Web Corpus itTenTen20 and complementary lexicographic resources. ChatGPT-5.5 was prompted to generate child-oriented Full-Sentence Definitions and corresponding dual illustrations. The outputs were evaluated qualitatively with regard to semantic accuracy, phraseological adequacy and pedagogical suitability. The findings show that ChatGPT successfully reproduces the formal conventions of COBUILD-style definitions, although recurrent difficulties emerge in the semantic simplification and age-appropriate adaptation of phraseological meaning. The image-generation analysis further reveals both the potential and the limitations of AI-assisted multimodal phraseography, particularly regarding semantic disambiguation, literal interpretation and representational bias. Overall, the study highlights the opportunities of generative AI for learner-oriented lexicography while emphasizing the continuing necessity of lexicographic supervision and post-editing.

Keywords Phraseological learner's dictionary; L2 phraseography; phraseodidactics; digital lexicography; children's dictionaries; AI-assisted definition generation; AI-generated illustrations; Generative AI

1 The Changing Landscape of Learner Lexicography

The rapid development of artificial intelligence is redefining both the nature of dictionaries and the practice of lexicography. New generations of learner-oriented, multimodal and interactive dictionary resources are emerging through AI-based technologies and large language models (Lew, 2023, 2024; Li & Tarp, 2024). Consequently, the role of the lexicographer is increasingly shifting from manual compilation toward the curation, supervision and critical evaluation of AI-generated content (Lew, 2023, pp. 4–9). At the same time, lexicography has gradually moved beyond the description of isolated words toward a broader treatment of phraseological and multiword units (Lew, 2024, pp. 2–3). Learner's dictionaries, particularly British monolingual ones, played a pioneering role in this development by systematically incorporating idioms, collocations, lexical chunks and other multiword expressions into dictionary entries (Cowie, 1999, pp. 52–81). The transition to digital media further accelerated this shift by removing physical space constraints and enabling new forms of access to lexical information (Lew & de Schryver, 2014, pp. 349–351; Lew, 2024, p. 3). Against this background, the present paper introduces *Phrasolino*, a model for an electronic Italian-German phraseological learner's dictionary designed for children aged 7–11 with L1 German and L2 Italian. Section 2 outlines the main principles of the dictionary model, including its target group, organisational structure, microstructural foundation and the role of AI in the generation of definitions and illustrations. Section 3 presents a case study based on 100 Italian idioms containing the body-part noun (henceforth BPN) *testa* ('head') and investigates the use of ChatGPT-5.5 for generating

COBUILD-style Full-Sentence Definitions (henceforth FSDs) and dual illustrations representing both the literal and phraseological meanings of the idioms.[1]

2 Phrasolino: A phraseological learner's dictionary model for children

Phrasolino is a model for an electronic Italian-German phraseological learner's dictionary for German L1 children aged 7–11. Its microstructure draws inspiration from the Phrase-based Active Dictionary (PAD), a web-interface prototype of an electronic learner's dictionary for German, Italian and English currently developed by DiMuccio-Failla, Giacomini and their research team (DiMuccio-Failla & Giacomini, 2022). The PAD constitutes one of three components of the broader PhraseBase project, alongside a phrase-based WordNet and a construction inventory (Giacomini, 2025, pp. 15–19; DiMuccio-Failla, 2025).

Within this framework, Phrasolino focuses on the development of phraseological competence in children learning Italian as an L2. Idioms play a central role in natural language use, allowing speakers to express abstract concepts in a concrete, compact and highly expressive manner (Espinal & Mateu, 2019, p. 4). Moreover, the acquisition of multiword expressions —which, according to Jackendoff's (1995, pp. 133–165) expanded concept of the lexicon, constitute a substantial proportion of a speaker's lexical inventory— contributes significantly to communicative competence and overall L2 proficiency (Boers et al., 2006; Siyanova-Chanturia & Pellicer-Sánchez, 2019, p. 1).

Following Wiegand's (1987) user-oriented framework, Phrasolino is designed according to the following parameters: (a) target users: German L1 children learning Italian as an L2; (b) format: digital; (c) context of use: school and home study; (d–e) user age and usage span: 7–11 years, corresponding to the primary-school period; (f) location: classroom or home environment; (g–h) purpose: support in learning Italian idioms; and (i–j) learning goal: the development of phraseological competence.[2]

2.1 *Core principle 1: Semantic grouping*

The dictionary model adopts a semantic rather than alphabetical organisation of idioms. Following phraseographical approaches such as Weinberger's (2012) *Kroatisch/Serbisch–deutsches phraseologisches Wörterbuch* and Schemann's *Synonymwörterbuch der deutschen Redensarten* (2012), phraseological units are arranged into semantic groups and subgroups according to shared meanings. This organisation facilitates the comprehension and memorisation of idioms while making semantic relations such as partial synonymy and antonymy more transparent.

[1] Given the large amount of visual material generated, selected supplementary figures, including the PAD entry model and the proposed Phrasolino microstructure as well as an example of a full dictionary entry, are provided via hyperlinks rather than reproduced in full within the paper. Furthermore, twenty AI-generated illustrations examined in the second part of the case study are linked directly to their corresponding ChatGPT-generated images, allowing readers to consult the complete visual dataset while maintaining a manageable length of the paper.

[2] Regarding the last point of the list, in the opinion of the author of this paper, the goal must be an enhancement of general idiomatic competence, as the CEFR framework does not provide clear or consistent criteria for classifying idioms by level; it mainly references communicative functions and lexical range, not idiomatic mastery. In applied linguistics and lexicography, many scholars (e.g., Boers et al., 2006) note that idiomatic competence develops gradually and unevenly, making a strict CEFR mapping artificial.

It therefore provides a more associative and learner-oriented access to phraseological meaning (Weinberger, 2012, pp. XI–XII).

2.2 Core principle 2: Multimodality through AI-generated illustrations

Illustrations in dictionaries function as ostensive definitions supporting and clarifying verbal explanations (Landau, 2001; Svensén, 2009). Their value is particularly evident in the case of abstract, complex or culturally specific concepts, where visual representations can facilitate meaning construction, reduce ambiguity and improve information retrieval (Gangla-Birir, 2015; Stein, 1991). Studies further indicate that learners process textual and visual information complementarily and that illustrations can enhance the comprehension, memorisation and retention of idiomatic expressions (Lew et al., 2018; Boers et al., 2008; Szczepaniak & Lew, 2011; Dziemianko, 2022).

Each idiom in Phrasolino is accompanied by two illustrations: one representing the literal interpretation and one visualizing the phraseological meaning. The former aims to make the lexical structure of the idiom more transparent, whereas the latter conveys its figurative and holistic meaning (cf. Burger, 2015, p. 13). Together, the images are intended to support learners in relating literal wording to phraseological interpretation. The dual-illustration approach is also compatible with Dobrovol'skij and Piirainen's (2018) *Conventional Figurative Language Theory*. According to this framework, many idioms retain an image component that remains cognitively accessible and functions as a semantic bridge between the literal wording and the phraseological meaning. Presenting both interpretations visually may therefore help learners explore the relationship between the underlying mental image and the conventional figurative interpretation, while simultaneously increasing phraseological transparency and motivation. From a psycholinguistic perspective, the approach is furthermore consistent with Cacciari and Tabossi's (1988) *Configuration Hypothesis*, according to which idiomatic meaning is activated once sufficient information in the string allows the expression to be recognized as an idiom, at the so-called idiom key (Tabossi & Zardon, 1993, p. 147). For children aged 7–11, literal and figurative meanings may initially be processed simultaneously, while figurative competence develops gradually throughout primary school (Levorato & Cacciari, 1999, p. 52). Making the relationship between literal and figurative meaning visually explicit may therefore facilitate phraseological learning (Boers, 2000).

Particular attention is devoted to idioms characterized by a double semantic plane (see Sections 2.3 and 3.3), in which both literal and phraseological interpretations are semantically plausible. In such cases, contextual cues and dual illustrations are intended to help learners distinguish between the two readings while simultaneously fostering awareness of idiomatic polysemy. The illustrations are generated with ChatGPT-5.5 using the corresponding FSDs as the semantic basis for prompt generation and image creation (see Section 3.3.2).

2.3 Microstructural basis and its extension

The PAD model, which forms the microstructural basis of Phrasolino, builds on Sinclair's (1991) idiom principle, according to which natural language largely consists of semi-prefabricated expressions rather than isolated words. Following Hanks's (2004) thesis about the genesis of norms, it prioritizes typical and frequent usage patterns. Its microstructure is designed to reflect natural language use and promote effective language learning through a multi-tiered structure that progressively guides users toward more detailed semantic information. The PAD distinguishes four description levels: the **lexeme level**, the **syntactic-construction level**,

the **sense-field level** and the **lexical-unit level**. Together, these levels combine formal, semantic and pragmatic information, including definitions, semantic relations, collocations and usage examples (Di Muccio-Failla & Giacomini, 2022, pp. 488–490). To facilitate the understanding of the structure described above, a PAD entry for the verb *agree* is linked here[3].

Unlike PAD entries, which typically start with the lexeme level, a Phrasolino entry starts with the **semantic group** and subgroup assigned to the idiom (e.g., *E attitude towards fellow human beings; Eb affection–aversion*). This organisation reflects the semantic structuring principle proposed in Section 2.1 and enables learners to approach idioms conceptually before focusing on individual phraseological units. To facilitate comprehension for German L1 children, the semantic group is accompanied by a German translation. The **head idiom**, the term adopted in Phrasolino for the phraseological unit heading the entry, simultaneously replaces the syntactic-construction and lexical-unit levels of the PAD. A **German equivalent** appears next to the head idiom, preferably phraseological whenever a suitable correspondence exists; otherwise, a non-idiomatic German paraphrase is provided. German equivalents are taken primarily from the *Dizionario Idiomatico tedesco-italiano* by Fenati et al. (2009). The head idiom is directly linked to an **illustration** representing its phraseological meaning. Beneath it, a literal translation (marked with λ* whenever the literal translation is not semantically acceptable) links to a second illustration visualizing the lexical components of the idiom. In line with Dziemianko's (2024) findings on the pedagogical advantages of immediately visible illustrations, the proposed dictionary model displays both illustrations directly within the entry while allowing users to enlarge them by clicking. This approach seeks to maximise immediate visual support without sacrificing the interactive nature of the digital resource. In cases where the literal reading is semantically acceptable, the asterisk is omitted and the literal meaning is additionally treated in the **Special Cases** section, where a second sense field appears. As next step, the **sense field** provides a short Italian paraphrase of the phraseological meaning of the head idiom and is likewise accompanied by a German translation. **Definitions** follow the COBUILD FSD model (Sinclair, 1987), which has been shown to be particularly suitable for language learners. To further support L2 users at primary-school level, each FSD is immediately followed by its German translation, reducing potential difficulties arising from Italian metalanguage while preserving exposure to authentic Italian. Usage **examples** are primarily extracted through GDEX (Good Dictionary Examples) in Sketch Engine (Kilgarriff et al., 2008). Where examples contain vocabulary or structures that may be too complex for the target users, they are linguistically adapted while preserving the idiom's authentic usage and meaning. **Synonyms** and **antonyms** are represented by semantically related idioms, or, where no suitable idiomatic equivalent exists, by semantically related collocations, each accompanied by a short definition and an authentic usage example. Both synonymous and antonymous idioms are linked to their corresponding dictionary entries, thereby guiding learners through a semantic network of interconnected idioms. **Variations** include formally similar idioms differing only in one or two lexical items while preserving the same meaning as the head idiom; for these, only usage examples are provided. Finally, the **Special Cases** section further develops the treatment of idioms characterized by a double semantic plane, namely the coexistence of a literal and a figurative meaning. This phenomenon has been discussed extensively in phraseological research (Dobrovol'skij & Piirainen, 2005, 2018; Mellado Blanco, 2021). While the figurative interpretation constitutes the conventional phraseological meaning, many idioms retain traces of their literal imagery, which may remain cognitively accessible and contribute to motivation, transparency and interpretation. Following the PAD framework, *Special Cases* refer to lexical units that are formally identical to the main lexical unit of the entry but display a different

[3] A more extensive discussion of the relationship between the PAD microstructure and Phrasolino will be discussed in a forthcoming paper. For detailed descriptions of the PAD architecture and its theoretical foundations, see Di Muccio-Failla and Giacomini (2022), Giacomini (2025) and Di Muccio-Failla (2025).

meaning. In Phrasolino, this category comprises idioms that are formally identical to the head idiom while additionally allowing a semantically plausible literal interpretation alongside the phraseological one. The symbol ~ marks this correspondence. Within the present dataset, 22 idioms displayed both a conventional phraseological meaning and a semantically acceptable literal interpretation, including *alzare la testa or camminare a testa alta*. Some examples are presented below:

- *alzare la testa:* literal: physically raising one's head; phraseological: rebelling/standing up for oneself
- *camminare a testa alta:* literal: walking with one's head held high; phraseological: acting proudly/with dignity

In these cases, the compositional equivalent is not marked with an asterisk, and the contextualised example additionally links to the corresponding literal illustration, making both interpretations visually accessible. As in the main entry, the sense field and the FSD in the **Special Cases** section are likewise accompanied by German translations.

3 Case study: Prompting definitions and figurative illustrations with GPT-5

The aim of the study is to investigate to what extent a Large Language Model (LLM) can reproduce lexicographic tasks involved in the compilation of the Phrasolino entries. The focus lies on the generation of AI-supported FSDs in the COBUILD tradition (see Section 3.3.1) as well as on the creation of dual illustrations for idioms containing the BPN *testa* ('head').

3.1 Preparation phase

Corpus linguistics is widely recognized for its crucial contribution to identifying pre-patterned word combinations as a fundamental aspect of language. The integration of extensive electronic corpora with advanced computational methods has significantly broadened the traditional scope of phraseology, shedding light on the diverse realizations of Sinclair's (1991) idiom principle across multiple languages (cf. Granger, 2019, p. 228). Given this methodological framework, the present study adopts a corpus-based approach to compile and analyse a dataset of Italian idioms containing the BPN *testa* ('head').

3.1.1 Phraseme selection

BPNs play a central role in phraseology and frequently function as source domains for metaphorical conceptualization due to the embodied nature of human cognition (Johnson, 1987; Lakoff & Johnson, 1980). Because they are grounded in physical and sensory experience, BPNs are particularly suitable for the conceptualization of abstract meanings (Kövecses, 2000). Cross-linguistic studies further show that lemmas such as *body, hand, head* and *foot* belong to the most frequent BPNs in phraseological expressions across languages (Ganfi et al., 2023, p. 9). Against this background, the present case study focuses on Italian idioms containing the BPN *testa* ("head"), selected due to the high phraseological productivity associated with this lemma. The dataset consists of 100 idioms collected through corpus-driven analysis and complemented by lexicographic resources. The idioms were subsequently classified according to core phraseological features such as stability, polylexicality, reproducibility, idiomaticity and expressivity (Burger, 2015, pp. 11–29, 77–78).

3.1.2 Research questions

The present study is guided by the following research questions:

- To what extent is ChatGPT-5.5 capable of generating pedagogically adequate COBUILD-style FSDs for Italian idioms in Phrasolino?
- Which prompt characteristics, linguistic criteria and recurrent difficulties emerge in the generation of child-oriented FSDs for learners aged 7–11?
- To what extent can AI-generated FSDs serve as a semantic basis for the generation of dual illustrations representing both the literal and the phraseological meanings of idioms?
- To what extent does AI-assisted phraseography require lexicographic supervision and post-editing?

3.2 Material collection phase

For the present case study, a dataset of 100 Italian idioms containing the BPN testa ('head') was compiled and analysed. The Italian Web Corpus itTenTen20 (Kilgarriff et al., 2014) was used to identify lexical information, collocational patterns and recurrent phraseological constructions associated with the lemma *testa*. In particular, PoS-gram patterns (i.e., sequences of part-of-speech tags in a specific order, such as Verb + Noun or Noun + Preposition + Noun) were retrieved through concordance searches, advanced CQL queries and Word Sketch analyses. The corpus-driven results were subsequently compared with lexicographic resources such as the *Dizionario dei modi di dire* (Quartu, 2012) and the online *Treccani* (Treccani, n.d.) dictionary, which were additionally used to identify further idioms and compile a final dataset of 100 idioms.

3.3 Processing phase

The definitions and dual illustrations for the dictionary entries of the present case study were generated in two interactive sessions on May 24th and 25th via the paid web interface of ChatGPT based on the GPT-5.5 language model. In contrast to the free version, the subscription-based version provides access to more advanced language and multimodal generation features, faster response times, greater generation stability and extended usage limits, which are particularly relevant for larger lexicographic workflows and iterative prompt refinement.

3.3.1 Prompting definitions

Phrasolino adopts the FSD format introduced by the COBUILD project (Sinclair, 1987), as this defining style appears particularly suitable for young language learners. Unlike traditional substitutable definitions, COBUILD-style definitions combine meaning and usage within a more natural and contextualized explanatory structure (Sinclair, 1987; Hanks, 1987). Typically, the first part of the definition illustrates the syntactic and phraseological usage pattern, while the second paraphrases the meaning itself (Barnbrook, 2002). As noted by Lew (2023, p. 2), this approach breaks away from the 'substitutability orthodoxy' by adopting a more conversational explanatory style resembling authentic pedagogical discourse. This learner-oriented format is particularly relevant for children aged 7–11, as it presents phraseological meaning through contextualized explanations rather than abstract metalinguistic formulations.

3.3.1.1 Linguistic criteria

The initial prompt instructed the model to analyse a list of 100 Italian idioms with the BPN *testa* ('head') and to generate COBUILD-style FSDs for each idiom. The prompt specified that each definition should consist of two parts: (1) a conditional clause presenting the concrete usage pattern of the idiom in the second person singular, and (2) a main clause explaining the phraseological meaning of the expression. The use of the second person singular (tu) was deliberately chosen because the dictionary is intended for children, as this form allows a more direct, accessible and learner-oriented style of explanation and may address learners more personally. The target group, namely children aged 7–11 with German as L1 and Italian as L2, was explicitly specified in the prompt together with instructions to employ simple, everyday language. The *Nuovo Vocabolario di base della lingua italiana* (De Mauro, 2016) was indicated as the principal lexical reference resource, and the corresponding online PDF was provided within the prompt. To ensure structural consistency, three model examples illustrating the desired FSD format were also included. Since the present paper is written in English, the original Italian prompt is provided here in English translation.[4]

3.3.1.2 Results and lexicographic evaluation of the AI-generated FSDs

All 100 COBUILD-style FSDs generated by ChatGPT-5.5 were manually evaluated with regard to semantic accuracy, phraseological adequacy, linguistic simplicity and age-appropriateness for the target group. Interestingly, although 22 idioms in the dataset displayed both a phraseological and a semantically plausible literal interpretation, ChatGPT consistently generated definitions only for the phraseological meaning. This suggests that the model automatically recognised these expressions as conventional idioms and prioritised their figurative meaning. Since idioms displaying a double semantic plane constitute an important category within the proposed dictionary model, future prompts should explicitly request both phraseological and literal definitions. The absence of literal definitions was therefore not treated as an error. The generated definitions were then assessed using Lew's (2023, pp. 4–8) evaluative categories: *Bad, Wanting, Passable, Good* or *Great*. Although this framework was originally developed for single lexical items rather than idioms, it was adopted here as a general qualitative framework for assessing the pedagogical adequacy of AI-generated FSDs. The specific phraseological shortcomings emerging from the present dataset were subsequently analysed separately. The evaluation was conducted by the author, who also acted as dictionary compiler, since the study was exploratory rather than designed to establish inter-rater reliability. Within the present study, *Bad* refers to definitions conveying an incorrect idiomatic meaning; *Wanting* to definitions that capture the general meaning but display important semantic or pedagogical shortcomings; *Passable* to generally adequate definitions requiring noticeable lexicographic revision; *Good* to definitions with only minor stylistic or lexical issues; and *Great* to publication-ready definitions requiring no substantial revision.[5]

Eight definitions were rated as *Bad* because they conveyed an incorrect idiomatic meaning, while 15 were classified as *Wanting* due to recurring problems that can be grouped into three main categories. A further 15 definitions were evaluated as *Passable*, whereas the remaining

[4] Links to the original GPT-5.5 interaction and to a PDF containing the complete documentation of the case study, including prompts, generated outputs and evaluation notes, are also provided in the reference section.

[5] The distinction between *Wanting* and *Passable* inevitably involves a degree of qualitative judgement. In the present study, *Wanting* definitions substantially impaired semantic precision or pedagogical adequacy, whereas *Passable* definitions were generally suitable but required noticeable lexicographic revision.

62 definitions were classified as *Good* or *Great*. The distribution of the ratings is presented in Table 1.

TABLE 1 Qualitative evaluation of the generated COBUILD-style FSDs

Rating category	Description	Number of definitions	Percentage
Great	semantically precise, natural and fully age-appropriate	23	23 %
Good	minor stylistic or lexical issues only	39	39 %
Passable	generally correct but requiring noticeable revision	15	15 %
Wanting	problematic simplification or conceptual mismatch	15	15 %
Bad	semantically inadequate or pedagogically unsuitable/ wrong sense	8	8 %
Total	-	100	100 %

Of the 38 definitions not rated as *Good* or *Great*, 8 were classified as *Bad* because the model failed to identify the correct idiomatic meaning. The subsequent qualitative analysis focuses on the 15 definitions classified as *Wanting*, as they best illustrate the recurring lexicographic and pedagogical shortcomings observed in the dataset. These shortcomings can be grouped into three main categories: semantic oversimplification, explanatory vagueness/circularity, and hidden conceptual abstraction despite surface simplicity. Although the 15 *Passable* definitions also required revision, their shortcomings were generally less pronounced and are therefore not discussed separately.

1. Linguistic simplicity vs. semantic precision (loss of idiomatic nuance): One recurring difficulty concerned the balance between simplified learner-friendly language and the preservation of the idiomatic meaning. In several cases, the generated definitions tended to oversimplify the semantic content of the idiom, thereby reducing important pragmatic, emotional or evaluative nuances. For example, *avere qualcosa in testa* (lit. 'to have something in one's head'; phraseological meaning: 'to be obsessed with/constantly think about something') was paraphrased merely as 'thinking about something all the time', while the idiom may additionally imply fixation, persistent worries or intrusive thoughts. Similarly, *essere una testa d'uovo* (lit. 'to be an egghead'; 'to be an intellectual/overly academic person') was reduced to 'being very intelligent and studying a lot', omitting the slightly ironic or pejorative connotation often associated with the expression. Comparable reductions emerged in idioms such as *riempire la testa* (lit. 'to fill someone's head'; 'to influence or overwhelm someone mentally'), where the generated definition failed to capture the idea of verbal overload or negative influence, and *gettarsi a testa bassa* (lit. 'to throw oneself head first'; 'to act recklessly or impulsively'), whose explanation ('doing something without fear') weakened the idiom's central nuances of imprudence and excessive impulsiveness. In *chiedere la testa* (lit. 'to ask for someone's head'; 'to demand someone's punishment or removal'), the paraphrase 'wanting someone to be punished' additionally softened the stronger semantic implications of condemnation, ruin or symbolic elimination conveyed by the original idiom.

2. Restricted learner vocabulary vs. explanatory specificity (vagueness and circularity): A second issue emerged from the attempt to restrict the definitions to basic everyday vocabulary suitable for children. For instance, *fare la testa come un pallone* (lit. 'to make someone's head like a balloon'; 'to overwhelm someone with excessive talking') generated a paraphrase such as 'talking too much and making someone tired', which insufficiently expresses the repetitive and mentally exhausting verbal pressure implied by the expression. Similar tendencies appeared in *mettere in testa qualcosa a qualcuno* (lit. 'to put something into someone's head'; 'to convince or influence someone psychologically'), where 'making someone have an idea or thought' remained semantically underspecified, and in *fare qualcosa con la testa nel sacco* (lit. 'to do something with one's head in a sack'; 'to act in a confused or unaware state'), whose explanation only vaguely conveyed the notions of confusion, unawareness or impaired judgement associated with the idiom.

3. Surface simplicity vs. cognitive abstraction (hidden conceptual abstraction despite simple language): Despite the use of simplified language, the model sometimes introduced abstract conceptual vocabulary more characteristic of adult-oriented conceptualizations. This tendency suggests that LLMs may implicitly reproduce abstract metalinguistic patterns even when instructed to generate child-oriented definitions. Notably, such conceptual abstractions emerged despite the prompt explicitly requiring the use of the *Nuovo Vocabolario di base* (De Mauro, 2016) and simple everyday language suitable for children. The findings therefore suggest that lexical simplification alone does not necessarily prevent cognitively abstract conceptualizations. Expressions such as *avere la testa quadra* (lit. 'to have a square head'; 'to think in an excessively rigid way') generated explanations involving notions such as 'rigid reasoning', while *uscire con la testa rotta* (lit. 'to come out with a broken head'; 'to emerge badly damaged from a situation') produced formulations referring to being 'seriously damaged by a situation'. Although semantically accurate, such explanations remain cognitively dense due to implicit psychological and metaphorical abstraction. Similarly, *essere una testa calda* (lit. 'to be a hot head'; 'to be hot-tempered and impulsive') was paraphrased through concepts such as impulsive behaviour and emotional self-regulation, which may exceed the cognitive and lexical level expected for children aged 7–11. These cases illustrate that syntactic simplicity alone does not automatically guarantee child-oriented comprehensibility. Similarly to Lew's (2023, pp. 4–9) findings, which showed that many ChatGPT-generated COBUILD definitions were comparable in quality to those written by human lexicographers, the present evaluation demonstrates that the model was highly effective in reproducing the formal and structural conventions of COBUILD-style FSDs.[6] The principal limitations emerged not at the syntactic level, but in the semantic and pedagogical adaptation of phraseological meaning for young learners. The findings suggest that LLMs can serve as valuable tools for learner-oriented phraseography, while still requiring substantial lexicographic supervision to ensure semantic precision, cognitive accessibility and age-appropriate explanations.

Bad, Wanting and *Passable* definitions were not discarded but manually post-edited by the lexicographer before inclusion in the proposed dictionary. Consequently, the present evaluation should be understood not as a selection mechanism for accepting or rejecting AI output,

[6] Referring to ChatGPT's ability to generate COBUILD-style FSDs, De Schryver (2023, pp. 374–381) describes a similar experiment in which the model successfully reproduced COBUILD-style dictionary entries across different parts of speech and idiomatic expressions. The author argues that this effectiveness is closely linked to ChatGPT's capacity to emulate corpus-pattern-based mappings between word meaning and word use originally developed within Patrick Hanks' explanatory framework for COBUILD lexicography. In this context, De Schryver even suggests that Patrick Hanks may be regarded as the 'first ChatGPT engine in lexicography", since he pioneered the systematic mapping of meaning onto usage patterns that Large Language Models now automate.

but as a means of identifying the degree of human revision required prior to publication. The following section presents the results of the dual-illustration experiment, focusing on the representation of both literal and phraseological meanings, particularly in idioms displaying a double semantic plane.

3.3.2 Prompting illustrations for Phrasolino

3.3.2.1 Methodology

In order to explore the potential of AI-assisted illustration generation for phraseological learner's lexicography, a smaller qualitative case study was conducted on a subset of 10 Italian idioms containing the BPN *testa* ('head'). The idioms were selected from the categories previously evaluated as *Good* or *Great* in the definition-generation phase. This selection criterion was deliberately adopted in order to minimise potential distortions resulting from inadequate FSDs and to test whether high-quality AI-generated definitions could successfully function as a semantic basis for illustration prompting. For each idiom, two separate illustrations were generated: one representing the literal interpretation and one representing the phraseological meaning. In addition, three idioms displaying a double semantic plane, i.e., idioms allowing both a phraseological and a semantically plausible literal interpretation, were intentionally included in the sample. For these expressions, the prompt explicitly indicated that the idiom allowed both a literal and a phraseological reading (e.g., *Camminare a testa bassa* [both literal & phraseological readings are possible]).[7] The analysis did not primarily aim to evaluate the artistic quality of the generated illustrations or all 100 dictionary entries, but rather to examine whether generative AI can successfully operationalise the dual-illustration principle proposed for Phrasolino and to identify recurring tendencies in AI-assisted illustration generation under controlled conditions. The selected idioms were intended to represent different degrees of idiomaticity and semantic complexity, including expressions displaying a double semantic plane. Subsequent experimentation carried out during the further development of Phrasolino confirmed that illustration quality depends largely on prompt specificity and iterative prompt refinement rather than on the idiom itself.

3.3.2.2 Prompt design

The illustration prompts were based on the previously generated COBUILD-style FSDs, which provided the semantic basis for the image generation process. In order to ensure comparability across the dataset, the prompts followed a standardised structure. In addition to the FSDs, all prompts contained stylistic instructions specifying a child-friendly comic or illustrated style suitable for children aged 7–11. Particular attention was paid to the distinction between literal and figurative meaning. The original Italian prompt was translated into English and is retrievable here.

3.3.2.3 Evaluation criteria

The generated illustrations were analysed qualitatively with regard to their pedagogical and phraseological adequacy. The evaluation focused primarily on semantic clarity, comprehensibility and the successful visual representation of phraseological meaning rather than on aesthetic criteria. Particular attention was paid to semantic adequacy, the visual differentiation of

[7] The original FSDs generated in the previous case study were retained for the illustration prompts. No additional literal FSDs were provided; the model was informed only that the three idioms displayed both a phraseological and a semantically plausible literal interpretation.

literal and figurative meanings, child-oriented comprehensibility and age-appropriateness, potential sources of semantic misguidance, and the extent to which abstract phraseological meanings could be transformed into concrete visual representations.

3.3.2.4 Results

The sample consisted of ten Italian idioms containing the BPN *testa* ('head'):[8] camminare a testa bassa, grattarsi la testa, mettersi in testa qualcosa, essere/andare fuori di testa, rompersi la testa, avere grilli per la testa, mettere i piedi in testa a qualcuno, avere la testa altrove, perdere la testa, and giocarsi la testa. Several recurring tendencies could be observed in the AI-generated illustrations. First, ChatGPT systematically integrated written text into the images in order to support semantic disambiguation between the literal and the phraseological interpretation. The model frequently added labels such as *significato letterale* and *significato fraseologico*, often accompanied by short pedagogical explanations beginning with expressions such as *'significa che . . . '*. Interestingly, the model tended to reuse the original COBUILD-style FSDs mainly for the phraseological illustrations, whereas the literal illustrations were almost always accompanied by short substitutable-style definitions. Only in the cases of camminare a testa bassa (lit. 'to walk with one's head down'; 'to feel ashamed or sad') and grattarsi la testa (lit. 'to scratch one's head'; 'to think hardly, because one does not understand something') did the figurative illustrations also contain more traditional explanatory formulations rather than the original FSDs. This suggests that the model implicitly associates phraseological meaning with more contextualized explanatory discourse, while tending to explain literal meanings through classical definitional strategies.

Another notable aspect concerns idioms displaying a double semantic plane, i.e., expressions allowing both a phraseological and a semantically plausible literal interpretation. In total 3 idioms of this type were included in the dataset. In the analysed sample, this phenomenon emerged particularly clearly in camminare a testa bassa, grattarsi la testa and mettersi in testa qualcosa (lit. 'to put something on/in one's head'; 'to decide firmly to do something'). In these cases, the model generally succeeded in separating the two semantic readings visually. Especially in mettersi in testa qualcosa, the literal illustration depicting a child physically putting an object on their head remains cognitively plausible, while the figurative image successfully conveys determination and persistence. In rompersi la testa (lit. 'to break one's head'; 'to think very hard about a problem') Chat GPT identified a broad colloquial literal reading that remains semantically understandable, even though the phraseological meaning is clearly more salient in actual language use.

A further recurring phenomenon concerned the rationalisation of semantically implausible literal readings. Even idioms whose literal interpretation would normally be perceived as absurd or impossible were represented visually in a highly plausible and coherent way. In mettere i piedi in testa a qualcuno (lit. 'to put one's feet on someone's head'; 'to dominate someone'), for example, the literal image depicts a girl physically placing her foot on another child's head. Similarly, avere grilli per la testa/avere la testa piena di grilli (lit. 'to have crickets in one's head/to have one's head full of crickets'; 'to have strange or fanciful ideas') is represented through the image of a child whose head literally contains crickets. Combined with the accompanying explanation *'avere grilli per la testa significa avere dei grilli dentro la testa'*, the resulting illustration normalises the literal reading to such an extent that it may appear cognitively valid or realistic for children. From a pedagogical perspective, this suggests that learner-oriented phraseological illustrations may require additional visual markers signalling that the literal

[8] Each idiom in the following list is linked to the corresponding original ChatGPT-generated image.

reading does not correspond to an actual real-world situation, for example through symbols such as a red cross or other disambiguating elements.

The generated illustrations additionally revealed a noticeable representational bias. Across the first five prompts, the model repeatedly produced highly similar child characters, typically white boys with light brown hair. Only after an explicit request for diversity did variation in gender, ethnicity and appearance emerge, suggesting a reliance on stereotypical default representations. Another interesting tendency emerged in idioms displaying more than one phraseological interpretation. In such cases, ChatGPT frequently generated separate visual elements for the different figurative meanings, which proved to be a surprisingly effective strategy for semantic differentiation. In perdere la testa (lit. 'to lose one's head'; 'to lose control/to fall deeply in love'), for example, the figurative image simultaneously represents emotional loss of control and romantic infatuation through two parallel visual scenes. This demonstrates a relatively sophisticated handling of phraseological polysemy at the multimodal level.

Not all idioms, however, produced satisfactory results. Particularly problematic was the case of giocarsi la testa (lit. 'to gamble one's head'; 'to risk everything'). Here, the literal definition itself was incorrectly interpreted compositionally as 'to play with one own's head', resulting in a visually strange illustration depicting a child playing with a toy head worn on top of her own head. The figurative illustration was also only partially successful, as the generated scene focused primarily on physical danger — a child balancing on a rope above a river — rather than on the broader figurative notion of risking everything in a high-stakes situation. This example illustrates that highly abstract or semantically opaque idioms remain especially difficult for AI systems to visualize adequately, even when supported by FSDs.[9] Illustrations that failed to represent the intended meaning satisfactorily were not accepted unchanged. Instead, prompts were iteratively refined until a pedagogically adequate result was obtained. The subsequent development of Phrasolino further showed that increasingly specific prompts, combined with short explanatory text embedded directly in the illustrations, substantially improved semantic clarity and reduced ambiguity.

4 Outlook: Opportunities and challenges

Similarly to Lew's (2023, pp. 8–9) COBUILD-based study, the present case study suggests that ChatGPT-5.5 is capable of generating learner-oriented COBUILD-style FSDs. The model successfully reproduced the characteristic structure and explanatory style of COBUILD definitions for children aged 7–11, and the predominance of Good and Great ratings suggests that Large Language Models can already serve as valuable tools in learner-oriented lexicographic workflows. At the same time, the main difficulties emerged not at the structural level, but in the semantic and pedagogical adaptation of phraseological meaning. Recurrent problems included semantic oversimplification, explanatory vagueness and cognitively abstract conceptualisations despite superficially simple language. Notably, such abstractions occurred even though the prompt explicitly referred to the *Nuovo Vocabolario di base della lingua italiana* (De Mauro, 2016) and required child-oriented vocabulary, suggesting that lexical simplification alone does not guarantee cognitively accessible explanations. More generally, the findings highlight that the quality of AI-generated lexicographic content depends less on the language model itself than on careful prompt engineering and iterative human supervision. Output quality proved to be strongly influenced by precise prompting, a clearly defined target group and subsequent human refinement. Successful AI-assisted phraseography therefore requires not only technical familiarity but also lexicographic, pedagogical and phraseological expertise.

[9] All 20 generated illustrations, together with the corresponding direct links, are also included in the shared PDF documentation.

The second part of the case study demonstrated that AI-generated FSDs may also function as a semantic basis for multimodal illustration prompting. In many cases, the generated dual illustrations successfully differentiated between literal and phraseological meaning and transformed abstract idiomatic concepts into visually accessible representations suitable for children. Particularly interesting were idioms displaying a double semantic plane, where the model was generally capable of visually separating two coexisting semantic interpretations despite the formal identity of the idiom. At the same time, the multimodal analysis also revealed important limitations. One particularly significant tendency concerned the rationalisation of semantically implausible literal readings. Even highly unrealistic literal interpretations were frequently visualised in coherent and cognitively plausible ways, potentially creating semantic ambiguity for younger learners. Similarly, the generated illustrations revealed recurring representational biases, with the model repeatedly defaulting to highly similar child figures until explicit prompting instructions requiring greater diversity were introduced. These observations suggest that AI-generated phraseological illustrations may require additional pedagogical and visual disambiguation strategies including increasingly specific prompts and the integration of short explanatory text within the illustrations themselves, in order to avoid semantic misinterpretation and stereotypical representations.

The present study does not yet include user testing, but represents an exploratory step focusing on the generation and expert evaluation of AI-generated definitions and illustrations. The resulting dictionary entries are currently being implemented as an interactive Axure RP prototype, which will serve as the basis for future learner studies. Likewise, no systematic comparison between AI-generated and human-generated lexicographic content has yet been undertaken. Future research should therefore investigate both learner interaction with the proposed dictionary model and potential differences in the reception and effectiveness of AI-generated versus human-generated definitions and illustrations. Overall, the findings suggest that Large Language Models offer considerable potential for learner-oriented and multimodal phraseography, particularly in supporting time-intensive tasks such as the generation of preliminary definitions, pedagogical reformulations and illustrative teaching materials. At the same time, the study demonstrates that the greatest potential of generative AI lies not in replacing the lexicographer but in supporting human expertise throughout the lexicographic workflow. Human supervision remains essential to ensure semantic precision, pedagogical adequacy, multimodal coherence and age-appropriate design. In this respect, the present findings reinforce Li and Tarp's (2024, p. 415) argument that the effectiveness of generative AI in lexicography ultimately depends on critical human evaluation and prompt-engineering competence.

References

Barnbrook, G. (2002). *Defining language: A local grammar of definition sentences*. John Benjamins Publishing Company.

Boers, F. (2000). Metaphor awareness and vocabulary retention. *Applied Linguistics*, 21, 553–571.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H. & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research*, 10(3), 245–261. https://doi.org/10.1191/1362168806lr195oa

Boers, F., Eyckmans, J. & Stengers, H. (2008). Variables in the mnemonic effectiveness of pictorial elucidation. In F. Boers & S. Lindstromberg (Eds.), *Cognitive linguistic approaches to teaching vocabulary and phraseology (Applications of Cognitive Linguistics*, Vol. 6, pp. 189–216). De Gruyter Mouton.

Burger, H. (2015). *Phraseologie: Eine Einführung am Beispiel des Deutschen*. Erich Schmidt.

Cacciari, C. & Tabossi, P. (1988). The comprehension of idioms. *Journal of Memory and Language*, 27, 668–683. https://doi.org/10.1016/0749-596X(88)90014-9

Cacciari, C. & Glucksberg, S. (1995). Imaging idiomatic expressions: Literal or figurative meaning? In M. Everaert, E.-J. van der Linden, A. Schenk & R. Schreuder (Eds.), *Idioms: Structural and psychological perspectives* (pp. 43–56). Lawrence Erlbaum Associates.

Chat-GPT 5.5 Chat (2026). *Prompting FSD in COUBUILD style & prompting dual illustrations for Phrasolino*. Retrieved May 24, 2026 from https://chatgpt.com/share/6a13332c-44a0-83eb-ab09-8399ad52764e

Cowie, A. P. (1999). *English dictionaries for foreign learners: A history*. Clarendon Press.

DiMuccio-Failla, P. V. & Giacomini, L. (2022). A proposed microstructure for a new kind of active learner's dictionary. *Lexicographica*, 38(1), 475–499. https://doi.org/10.1515/lex-2022-0016

DiMuccio-Failla, P. V. (2025). A theory for a usage-based cognitive lexicography. In L. Giacomini & V. Piunno (Eds.), *Patterns of meaning in lexicography and lexicology (Lexicographica Series Maior*, Vol. 167, pp. 19–90). De Gruyter.

De Mauro, T. (2016). *Il Nuovo vocabolario di base della lingua italiana*. Internazionale. Retrieved May 17, 2026, from https://www.internazionale.it/opinione/tullio-de-mauro/2016/12/23/il-nuovo-vocabolario-di-base-della-lingua-italiana

De Schryver, G.-M. (2023). Generative AI and lexicography: The current state of the art using ChatGPT. *International Journal of Lexicography*. Advance online publication. https://doi.org/10.1093/ijl/ecad021

Dobrovol'skij, D. O., & Piirainen, E. (2005). *Figurative Language: Cross-Cultural and Cross-Linguistic Perspectives*. Elsevier.

Dobrovol'skij, D., & Piirainen, E. (2018). Conventional Figurative Language Theory and Idiom Motivation. *Yearbook of Phraseology*, 9, 5–30. https://doi.org/ 10.1515/phras-2018-0003

Documentation of Case Study, original link (2026). Retrieved May 24, 2026 from https://drive.google.com/file/d/1c7n4CJGQj77QnOXLp8EuC1pKafkfsPHk/view?usp=sharing

Dziemianko, A. (2022). The usefulness of graphic illustrations in online dictionaries. *ReCALL*, 34(2), 218–233. https://doi: 10.1017/S0958344021000264

Dziemianko, A. (2024). Pictures in online dictionaries: Shall we see them? *GEMA Online® Journal of Language Studies*, 24(2), 31–53. https://doi.org/10.17576/gema-2024-2402-03

Espinal, M. T. & Mateu Fontanals, J. (2019). Idioms and phraseology. *Oxford Research Encyclopedia of Linguistics*, 1–37. https://doi.org/10.1093/acrefore/9780199384655.013.51

Fenati, B., Rovere, G., & Schemann, H., con la collaborazione di L. Giacoma. (2009). *Dizionario idiomatico tedesco-italiano*. Zanichelli.

Ganfi, V., Piunno, V. & Mereu, L. (2023). Body part metaphors in phraseological expressions: A comparative survey on Italian, Spanish, French and English. *Languages in Contrast*, 23(1), 1–32.

Gangla-Birir, L. (2015). The use of pictorial illustrations in African language dictionaries. *Lexikos*, 15, 38–51. https://doi.org/10.5788/15-0-3

Granger, S. (2019). Formulaic sequences in learner corpora. In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding formulaic language: A second language acquisition perspective* (pp. 228–245). Routledge.

Giacomini, L. (2025). Introduction to the PhraseBase project. In L. Giacomini & V. Piunno (Eds.), *Patterns of meaning in lexicography and lexicology (Lexicographica Series Maior*, Vol. 167, pp. 15–18). De Gruyter.

Gibbs, R. & O'Brien, J. (1990). Idioms and mental imagery: The metaphorical motivation for idiomatic meaning. *Cognition*, 36(1), 35–68.

Hanks, P. (1987). Definitions and explanations. In J. Sinclair (Ed.), *Looking up: An account of the COBUILD project in lexical computing* (pp. 116–136). Collins.

Hanks, P. (2004). Corpus pattern analysis. In G. Williams & S. Vessier (Eds.), *Proceedings of the 11th EURALEX International Congress* (pp. 87–98). Université de Bretagne Sud.

Jackendoff, R. S. (1995). The boundaries of the lexicon. In M. Everaert, E.-J. van der Linden, A. Schenk & R. Schreuder (Eds.), *Idioms: Structural and psychological perspectives* (pp. 133–165). Lawrence Erlbaum Associates.

Johnson, M. (1987). *The body in the mind: The bodily basis of meaning, imagination, and reason*. University of Chicago Press.

Kilgarriff A., Husak M., McAdam K., Rundell M., Rychlý P. (2008) GDEX: automatically finding good dictionary examples in a corpus. In E. Bernal, J. DeCesaris (Eds), *Proceedings of the XIII EURALEX International Congress*. Universitat Pompeu Fabra, Barcelona, pp. 425–432.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, 1, 7–36.

Kövecses, Z. (2000). *Metaphor and emotion: Language, culture, and body in human feeling*. Cambridge University Press.

Lakoff, G. & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.

Landau, S. I. (2001). *Dictionaries: The art and craft of lexicography* (2nd ed.). Cambridge University Press.

Levorato, M. C. (1999). Idiom comprehension in children: Are the effects of semantic analysability and context separable? *European Journal of Cognitive Psychology*, 11(1), 51–66.

Lew, R. & de Schryver, G.-M. (2014). Dictionary users in the digital revolution. *International Journal of Lexicography*, 27, 341–359. https://doi.org/10.1093/ijl/ecu011

Lew, R. et al. (2018). Competition of definition and pictorial illustration for dictionary users' attention: An eye-tracking study. *International Journal of Lexicography*, 31(1), 53–77. https://doi.org/10.1093/ijl/ecx002

Lew, R. (2023). ChatGPT as a COBUILD lexicographer. *Humanities and Social Sciences Communications*, 10, Article 704. https://doi.org/10.1057/s41599-023-02119-6

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, 11, Article 389. https://doi.org/10.1057/s41599-024-02889-7

Li, Q. & Tarp, S. (2024). Using generative AI to provide high-quality lexicographic assistance. *Lexikos*, 34, 397–418. https://doi.org/10.5788/34-1-1944

Mellado Blanco, C. (2021). Phraseology, Patterns and Construction Grammar: An Introduction. In N. Filatkina, K. Steyer & S. Stumpf (Eds.), *Productive Patterns in Phraseology and Construction Grammar* (pp. 1–25). De Gruyter.

Schemann, H. (2012). *Synonymwörterbuch der deutschen Redensarten*. De Gruyter. (SDR)

Sinclair, J. (Ed.). (1987). *Looking up: An account of the COBUILD project in lexical computing and the development of the Collins COBUILD English language dictionary*. Collins ELT.

Sinclair, J. M. (1991). *Corpus concordance collocation*. Oxford University Press.

Siyanova-Chanturia, A. & Pellicer-Sánchez, A. (2019). Formulaic language: Setting the scene. In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding formulaic language: A second language acquisition perspective* (pp. 1–15). Routledge.

Svensén, B. (2009). A handbook of lexicography: *The theory and practice of dictionary-making*. Cambridge University Press.

Szczepaniak, R. & Lew, R. (2011). The role of imagery in dictionaries of idioms. *Applied Linguistics*, 32(3), 323–347. https://doi.org/10.1093/applin/amr001

Stein, G. (1991). Illustrations in dictionaries. *International Journal of Lexicography*, 4(2), 99–127.

Tabossi, P. & Zardon, F. (1993). The activation of idiomatic meaning in spoken language comprehension. In C. Cacciari & P. Tabossi (Eds.), *Idioms: Processing, structure and interpretation* (pp. 145–162). Lawrence Erlbaum Associates.

Tomasello, M. (2001). First steps toward a usage-based theory of language acquisition. *Cognitive Linguistics*, 11(1), 61–82. https://doi.org/10.1515/cogl.2001.012

Tomasello, M. (2015). The usage-based theory of language acquisition. In E. L. Bavin & L. R. E. Naigles (Eds.), *The Cambridge handbook of child language* (2nd ed., pp. 89–106). Cambridge University Press.

Treccani (n. d.). *Testa*. Retrieved May 17, 2026, from https://www.treccani.it/vocabolario/testa1/

Varriale, S. (2023). La fraseologia nell'apprendimento dell'italiano LS in contesti tedescofoni/ Phraseology in learning Italian as a foreign language in German-speaking contexts. DIDIT. Didattica dell'Italiano. *Studi applicati di lingua e letteratura*, 3, 95–116.

Weinberger, H. (2012). *Kroatisch/Serbisch – deutsches phraseologisches Wörterbuch angeordnet nach semantischen Gruppen*. Facultas.

Wiegand, H. E. (1987). Zur handlungstheoretischen Grundlegung der Wörterbuchbenutzungsforschung. *Lexicographica*, 3, 178–227.

Quartu, B. M. (2012). *Dizionario dei modi di dire della lingua italiana*. Hoepli. Retrieved May 17, 2026, from https://dizionari.corriere.it/dizionario-modi-di-dire/a.shtml

Contact information

Linda Prossliner
Institute of Translation Studies, University of Innsbruck, Austria
Linda.Prossliner@uibk.ac.at

Hana Skoumalová

EXPANDING THE MWE DATABASE WITH VARIANTS

Abstract This paper presents recent extensions to the Czech Lemur MWE database, focusing on the systematic generation of syntactic transformations and morphological derivations to improve phraseological corpus annotation. The database currently contains over 29,500 entries, each enriched with information on positional variability, optional components, and synonymy, enabling the mwe_tagger to identify a wide range of attested variants. We describe a rule–based approach to generating periphrastic passives, reflexive passives, and deagentive constructions, as well as selected derivational variants supported by MorfFlex dictionary and DeriNet tool. While simple derivations (e.g., nominalizations or participial adjectives) can be safely automated, more complex cases require manual validation due to semantic drift or structural constraints. To evaluate the impact of these expansions, we compare annotations in SYNv14 and two versions of the Frap corpus. Although transformed or derived MWEs remain relatively infrequent (2–3.5% of all MWE tokens), their inclusion increases annotation coverage and reveals additional phraseological behavior. These results demonstrate that controlled automatic expansion is both feasible and beneficial for large–scale MWE annotation.

Keywords Corpus annotation; phraseology; database expansions

1 Introduction

In this paper, we introduce the MWE database for Czech—Lemur[1], and explain how we enhance it with transformations and derivations. The database is designed to serve both for corpus annotation purposes and as a lexicographic resource for human users (Vondřička, 2019; Skoumalová et al., 2024). Unlike other projects that focus on annotating MWEs in corpora (e.g., PARSEME; Savary et al., 2026), and use manually annotated corpora to train a language model to recognize MWEs, our project is based on building a lexical database of MWEs, which is used to search for and tag MWEs in texts. The database currently contains approximately 29,500 entries (the working version contains over 30,000 entries); however, the entries also include possible variants (such as synonyms in certain positions, optional components of MWEs, etc.), so the total number of MWEs recorded in the database is higher.

The phraseological annotation of the corpus consists of an MWE lemma (mwe_lemma attribute) and an MWE tag (mwe_tag attribute). The mwe_tag attribute has ten positions—MWE usage type, syntactic type, derivation or transformation type, and six additional positions containing labels for the idiomaticity of the whole MWE (lexical, morphological, syntactic, semantic, pragmatic, and statistical). The last field is of a more technical nature—exactly one member of the MWE is marked with the letter C so that each occurrence of the MWE may be found exactly once and so that statistics can be calculated.

The corpora are annotated using the mwe_tagger program, which utilizes a compiled version of the dictionary. During compilation, all possible synonyms for individual positions in a phrase are used, word order constraints are applied, options for inserting optional expansions are considered, and possible transformations and derivations are also applied.

[1] https://db.korpus.cz/search/lemur.

2 The Lemur database

The Lemur database serves both lexicographers and other human users, as well as a data source for the phraseological annotation of corpora. The annotation program mwe_tagger searches the text (within sentence boundaries) for potential MWEs and annotates them. Figure 1 shows the phraseme *jít svou cestou* 'go one's own way' found in a corpus. We can see that various variants of the expression have been identified in the corpus—such as *jít svou cestou* 'go one's way' (line 2), *jít vlastní cestou* 'go own way' (line 1), *jít svou vlastní cestou* 'go one's own way' (line 9), *kráčet svou cestou* 'walk one's way' (line 11), etc.—as well as different word-order variants, like '*vlastní cestou jde autor filmu lit.* 'own way.INS goes author.NOM of the film' (line 12).

1	budou USA trvat na tom , že	půjdou/jít_svou_cestou	vlastní cestou , a zpochybňovat základní linii
2	jsem to respektoval . </s><s> A on si	šel/jít_svou_cestou	svou cestou a já jsem pokračoval v
3	a Erdogan zdůrazňuje , že Turecko si	půjde/jít_svou_cestou	svojí cestou , už méně nahlas říká
4	, jako je paní Merkelová , a	půjdeme/jít_svou_cestou	vlastní cestou . </s><s> Máme s kým jít
5	1991 nastalo rozdělení vesnic a každá si	šla/jít_svou_cestou	vlastní cestou . </s><s> Tělovýchovu jsem nechal spojenou
6	s Legem , ani jej kopírovat ,	jdeme/jít_svou_cestou	svou cestou . </s><s> Nechceme děti svazovat návody
7	spekulantů a makléřů odmítly socialistický klam a	kráčejí/jít_svou_cestou	vlastní cestou . </s><s> Cestou svobodného , kapitalistického
8	to provokace Demokratů . </s><s> Ale vylhaná hysterie	šla/jít_svou_cestou	už svojí cestou . </s><s> Neustále běžely zprávy
9	kolegům sdělil , že každá země může	jít/jít_svou_cestou	nyní svou vlastní cestou . </s><s> Zároveň ujistil
10	neustále využíváni z kterékoliv světové strany . </s> <s>	Pojďme/jít_svou_cestou	svojí vlastní cestou , Česká republika si
11	více upevňují a zdokonalují a s jistotou	kráčejí/jít_svou_cestou	svou cestou . " </s><s> Zvláštní úloha v
12	těší mě to . " </s><s> Vlastní cestou	jde/jít_svou_cestou	autor filmu i ve formě propagace . </s>
13	Laky Royal . </s><s> Oba expartneři si tak	jdou/jít_svou_cestou	dál životem každý vlastní cestou . </s><s> Přitom

FIGURE 1 Phraseme *jít svou cestou* found in a corpus

Figure 2 shows the entry in the database. The figure shows the slots for the individual positions in the phrase, which are color-coded according to the properties defined for each slot. The first slot is highlighted in blue, which means it is required. However, it contains two synonyms, *jít* 'go' and *kráčet* 'walk,' either of which can be used in the first position of the phrase. The second slot, containing the word *si* 'self.DAT' (dative of convenience), is highlighted in gray, which means it is optional. In Figure 1, this variant is used on line 2. Slots 3 and 4 are highlighted in green, which means that at least one of them must be used in the expression. Possible variations are *jít (si) svou cestou*, *jít (si) vlastní cestou*, or *jít (si) svou vlastní cestou*, but not *jít (si) cestou* 'go way.INS.' This variant is not considered a phraseme.



FIGURE 2 Phraseme *jít svou cestou* in the working version of the database

In addition to the variability described above, other words can also be inserted into the expressions, such as in line 13 of Figure 1: *oba expartneři* ***si*** *tak* ***jdou*** *dál životem každý* ***vlastní cestou*** 'both ex-partners thus go on with their lives, each following their own path.'

3 Compilation of the database

All the possibilities and restrictions described so far are stored in the database, and when the database is compiled into a machine-readable dictionary, they are used to generate all possible variations of a given expression. This yields a list of expressions:

- *jít svou cestou*
- *jít vlastní cestou*
- *jít svou vlastní cestou*
- *jít si svou cestou*
- *jít si vlastní cestou*
- *jít si svou vlastní cestou*
- . . . and similarly for *kráčet*

When the program searches the text for the MWEs, variations in word order and the insertion of additional words are also permitted. In principle, all grammatically permitted variations are allowed; only variants that are impossible are explicitly restricted in the database.

4 Transformations and derivations

While reviewing the results in the annotated corpora, we realized that some expressions may appear in syntactically transformed forms, or that certain words may be used in morphologically derived forms. Trying to enter all these items into the database manually would be both time-consuming and unreliable—and, above all, unnecessary, since it can be done automatically during the database compilation process.

4.1 Transformations

We speak of transformation when there is a change in syntactic structure, but the POS of all words contained in the MWE remains unchanged. Typically, these are periphrastic passive constructions, reflexive passives, and reflexive deagentive constructions. An example of the periphrastic passive is the sentence: *Smlouvy byly uzavírány na desítky let.* 'Contracts were concluded for decades.', in which the active construction *uzavřít smlouvu* 'to conclude a contract' has been changed to the passive.

The conversion to the periphrastic passive is performed using lexical rules (see Rosen & Skoumalová, 2018):

1. If the MWE contains a noun phrase in the accusative case, it is converted to a subject in the nominative case, and the original subject is deleted (if it is a member of MWE). The verb in the derived MWE must be in the passive participle form and must agree morphologically with the new subject.

1. házet flintu do žita → flinta je hozena do žita
to throw rifle.ACC to rye → rifle.NOM is thrown to rye
'to give up'

2. If the MWE does not contain a noun phrase in the accusative case, then only the original subject is omitted (if present). The verb must be in the passive participle form and agrees with the neuter singular.

2. dosáhnout úspěchu → úspěchu bylo dosaženo
to reach success.GEN → success.GEN was reached

For reflexive passives similar rules apply. The original object in the accusative case becomes the subject in the nominative case, the verb takes the reflexive particle *se*, and it can only appear in finite forms in the text (ex. 3). If the verb does not have an accusative object, a deagentive construction is formed without a subject but with the reflexive particle *se*. The verb must be in a finite form and agrees with the neuter singular (ex. 4).

3. házet flintu do žita → flinta se hodí do žita
to throw rifle.ACC to rye → rifle.NOM SE throws to rye

4. kálet do vlastního hnízda → kálí se do vlastního hnízda
to defecate to own nest → defecates SE to own nest
'to foul one's own nest'

We also include constructions in which the subject of the verb phrase becomes the head of a noun phrase and the original predicate turns into a relative clause, like *jablko nepadlo daleko od stromu—jablko, které nepadlo daleko od stromu* 'apple did not fall far from tree—apple that did not fall far from tree.' This transformation applies only to proverbs and sentence-like MWEs. In the standard word order, the subject (member of the MWE in nominative) precedes the verb, and they usually follow one another immediately; therefore, it is sufficient to insert a comma between the subject and the verb, followed by the pronoun *který* 'which,' which agrees in gender and number with the subject. This change applies to sentence phrases and proverbs.

4.2 Derivations

Derivations involve morphological changes to the head of the phrase and sometimes also changes in part of speech. Some derivations are relatively simple and regular, and they are included in the morphological dictionary MorfFlex we use (Hajič, 2004; Hajič et al., 2024).[2]

Simple derivations involve nominalizing verbs using the suffixes *-ní/tí*, e.g., *kopat* → *kopání* 'to dig → (the) digging' or *lít* → *lití* 'to pour → (the) pouring,' or adjectivizing verbs, e.g. *volat* → *volaný* 'to call → called,' or *lít* → *litý* 'to pour → poured.'

For more complex derivations, we use the DeriNet tool (Ševčíková & Žabokrtský, 2014). DeriNet functions as a network that links words with the same root. It does not specify the direction of derivation (it does not identify the base word and its derivatives) nor does it provide information on whether two words are semantically synonymous. That is why we must be careful when using this tool.

[2] In fact, we use a modified version of the MorfFlex dictionary, which has not yet been published. However, simple derivations work the same way in it.

Relatively safe derivations are feminine forms derived from masculine nouns, such as *ministr financí* → *ministryně financí* 'minister of finance' → 'female minister of finance.' We need to be more careful with diminutives. For example, *zdravotní sestřička*, derived from *zdravotní sestra*, has the same meaning, 'medical nurse.' In contrast, the word *panna* has several meanings—virgin, girl (archaic), or doll—and the MWE entry *stará panna* means 'old maid,' while *stará panenka* means 'old doll.' The situation is even more complicated with words like *hra* '(the) play,' which is considered to be a derivative of the verb *hrát* '(to) play' in DeriNet. This derivation of the verb *hrát* is possible only when the verb does not take an accusative object:

5. hrát do karet → hra do karet
to play to the cards → (the) play to the cards
'play into someone's hands'

6. hrát mrtvého brouka → *hra mrtvého brouka
to play a dead beetle → (the) play dead beetle.GEN/ACC
'to play possum'

In the case shown in ex. 6, it would be possible to change the derived MWE to *hra na mrtvého brouka*, '(the) play of dead beetle', but it is not always possible to insert a preposition before nouns derived this way from verbs, and even when it is possible, different prepositions are used for different words. Another difficult case involves MWE consisting of a verb and an adverb. In this case, the adverb must be converted into an adjective and the word order changed. It should be noted, however, that even the regular derivation *hrát* → *hraní* must undergo this transformation.

7. hrát férově → férová/é hra/hraní
to play fair → (the) fair play/playing

Generally speaking, only certain derivations are safe enough (and the associated transformations simple enough) that we can use them automatically without concern. The rest need to be checked, and selected ones added to the database manually.

5 Corpus annotation and results

The first corpus annotated using the Lemur database was SYNv14 (Křen et al., 2025). This corpus contains more than 6.6 billion positions and 131,293,719 MWEs were found in it, which corresponds to a frequency of 19,843.51 i.p.m. (instances per million). Beside the SYNv14 corpus, we also annotated the Frap corpus, which is a part of an ongoing project, and we compared the annotation with SYNv14 corpus. In Frap corpus we used only the PUB part (journalistic texts from SYNv13) and we annotated it both in development and production mode. In production mode, the annotation program only annotates those transformations and derivations that we consider safe. In development mode, we annotate all derivations offered by DeriNet, and then we go through the annotated MWEs and decide whether to include the derived word in the database among the synonyms of the original word or to exclude it.

The results of the comparison are shown in Table 1, where the abbreviation i.p.m. stands for instances per million (for easier comparison of corpora of different sizes), and T&D stands for transformations and derivation. The T&D column for Frap-prod corpus shows transformations and derivations that are considered safe, while the same column for Frap-dev corpus shows all T&Ds including those that may have different meaning than the original MWE. The last column shows the percentage of T&Ds from all found MWEs.

TABLE 1 Annotated MWES in three corpora

	Positions	MWEs	i.p.m.	T&D	i.p.m.	% of T&D
SYNv14	6,616,455,821	131,293,719	19,843.51	2,942,683	444.75	2.24
Frap-prod	76,025,911	2,107,772	27,724.39	57,857	761.02	2.74
Frap-dev	76,025,911	2,123,454	27,930.66	74,751	983.23	3.52

From the table we can read several things. First, MWEs that undergo a transformation or derivation are not very frequent in corpora. The difference in i.p.m.s between SYNv14 and Frap-prod shows the amount of work that has been done on the database since last year—the difference in i.p.m. cannot be explained only by differences in the texts, since Frap is a random selection of SYNv14. And the difference between Frap-prod and Frap-dev shows the number of non-safe T&Ds, which have to be checked and either ignored or added to the database manually. The study therefore demonstrates that controlled automatic expansion is not only feasible but also yields measurable benefits for large–scale MWE annotation.

References

Hajič, J. (2004). *Disambiguation of Rich Inflection (Computational Morphology of Czech)*. Karolinum.

Hajič, J., Hlaváčová, J., Mikulová, M., Straka, M., Štěpánková, B. (2004). MorfFlex CZ 2.1. Data/software, LINDAT-CLARIAH. https://hdl.handle.net/11234/1-5833.

Křen, M., Čapka, T., Hnátková, M., Jelínek, T., Křivan, J., Petkevič, V., Skoumalová, H., Vondřička, P. (2025). SYN v14: corpus of contemporary written Czech. https://www.korpus.cz/kontext.

Rosen, A., Skoumalová, H. (2018). No way to have your say out of the frame: Specifying valency of multi-word expressions. *Prace Filologiczne*, LXXII(1), 301–320.

Savary, A., Scholivet, M., Ramisch, C., Nakamura, T., Bilinski, E., Stymne, S., Giouli, V., Markantonatou, S., Pais, V., Mitrofan, M., Estève, L., Guillaume, B., Barbu Mititelu, V., Čibej, J., Hernández, R. D., Fendel, V., Gantar, P., Kanishcheva, O., Krstev, C., Liebeskind, C., Lobzhanidze, I., Marković, A., Nešpore-Bērzkalne, G., Pagano, A., Shamsfard, M., Stankovic, R., Tajalli, V., Tiberius, C., Padhye, A. (2026). PARSEME 2.0 Multilingual Corpus of Multiword Expressions. In N. Calzolari, K. Choukri, A. Gangemi, B. Maegaard, J. Mariani, J. Odijk & D. Tapias (Eds.), *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)* (pp. 4819–4834), Palma, Spain. European Language Resources Association (ELRA).

Skoumalová, H., Kopřivová, M., Petkevič, V., Jelínek, T., Rosen, A., Vondřička, P., Hnátková, M. (2024): Lemur: A lexicon of Czech multiword expressions. In V. Giouli & V. Barbu Mititelu (Eds.), *Multiword Expressions in Lexical Resources: Linguistic, Lexicographic, and Computational Perspectives* (pp. 1-37) Language Science Press.

Ševčíková, M., Žabokrtský, Z. (2014). Word-Formation Network for Czech. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (Eds.), *Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014)* (pp. 1087–1093).

Vondřička, P. (2019): Design of a multiword expressions database. *Prague Bulletin of Mathematical Linguistics*, 112, 83–101.

Contact information

Hana Skoumalová
Charles University, Faculty of Arts
hana.skoumalova@ff.cuni.cz

Acknowledgements

This work was supported by the Czech Science Foundation as part of Project No. 24-10254S *Czech phraseology and changes in its use in temporal and genre contexts*.

Petra Storjohann

THE INTERPLAY BETWEEN INDIVIDUAL NEW LEXICAL ITEMS AND MULTI-LAYERED DISCOURSE REPRESENTATIONS OF NEOLOGISMS

Abstract This paper examines how discourse lexicography can be extended to the study of contemporary neologisms, particularly those emerging from major societal crises such as COVID-19, climate change, energy shortages, migration, inflation and the war in Ukraine. Traditional discourse lexicography has mainly focused on historically controversial terms and has typically relied on selected textual sources rather than large-scale corpora. In contrast, the paper argues for a corpus-assisted discourse-analytic approach that systematically captures the semantic, evaluative and ideological dimensions of newly coined and borrowed lexical items. Particular attention is paid to features such as semantic prosody, affective meaning, stance-taking, politicisation and power relations, which often become entrenched in neologisms and shape their use in public discourse. The paper introduces the German neologism resource IDS *Neo*$^{2020+}$, which combines two complementary levels of consultation: detailed lexical entries for individual neologisms and broader discourse or cluster entries that trace thematic developments and discourse networks over time. Both entry types are interconnected. This dual structure allows users to explore both the meanings of specific new words/expressions and their wider discursive functions and relations with other new terms within public communication. The paper also outlines the corpus-linguistic and computational methods used in the project, including word embeddings.

Keywords Controversial neologisms; CADS; discourse lexicography; discourse entries; IDS *Neo*$^{2020+}$

1 Introduction

German discourse lexicography (Kämper, 2006) has largely focused on documenting historically contested terms, while neologisms have received comparatively little attention. Moreover, semantic analyses of discourse-relevant historical keywords have typically relied on selected sources, such as parliamentary speeches or diaries, rather than large-scale corpora. By contrast, corpus-assisted approaches to critical discourse analysis (CADS), as proposed by Baker (2006) and Partington et al. (2013), are well established for examining controversial lexical items. These approaches combine corpus-linguistic methods with statistically robust analyses across extensive text collections. However, their potential remains underexplored in both discourse lexicography and the lexicographic treatment of neologisms. Applying CADS to neologisms appears particularly promising (Storjohann, in press), especially as many recent lexical innovations have emerged in response to crises such as COVID-19, inflation, the energy and cost-of-living crises, climate crisis, skilled labour shortage, migration and the war in Ukraine. This paper argues that the lexicographic treatment of contemporary neologisms must systematically incorporate their discursive dimensions to adequately capture the meanings of newly coined, crisis-related terms. Because such neologisms typically emerge and evolve within dynamic public debates, their semantic profiles cannot be fully represented through decontextualised definitions alone. Instead, their meanings are shaped by evaluative stances, ideological positions and the broader discourse contexts in which they occur. These often include emotionalisation, politicisation, marginalisation and power dynamics, such as those involved in exclusion or discrimination (e.g. Baker et al., 2013; Lewandowska-Tomaszczyk & Wilson, 2022). These elements become increasingly entrenched in the meanings of many neologisms (Schmid, 2018), they can be detected in corpora and discourse lexicography offers a robust framework for their systematic documentation and representation.

The new (soon to be launched) German neologism resource IDS *Neo*$^{2020+}$ provides dual-level access to new words within critical discourses. The resource provides headword entries with fine-grained information on individual neologisms, encompassing their semantic, evaluative and functional properties. The inclusion of neologisms in the resource involves an automatic detection procedure followed by editorial review based on lexicographic criteria. Using the IDS written corpora as digital resources, previously unattested lexical items can be detected automatically (Klosa &Üngen, 2018). Each year, this process generates several thousand candidate neologisms, which are initially ranked by frequency and distribution. The automatically generated list is subsequently refined through editorial review. Non-lexical items, such as proper names, spelling errors and orthographic variants, are removed, while inflected forms, synonyms and word-formation variants are grouped and analysed. Additional frequency analyses are then used to determine the appropriate headword for dictionary inclusion (see also Heylen et al., in prep.). Among the items excluded from the candidate list are proper names. Spelling errors occur particularly in the case of compound words borrowed from English, which are often adopted without being adapted to the conventions of German orthography. The editorial evaluation of the automatically generated candidate list also includes the identification of synonyms (e.g. *Micro-Retirement/Mini-Rente*) and morphological variants of the same lexeme (e.g. *D-Ticket* and *Deutschlandticket*). In a subsequent step, coexisting orthographic variants are consolidated (e.g. *Deutschlandticket* vs. *Deutschland-Ticket*), and derivational forms of the neologism are identified (e.g. *Bürgergeldempfänger, Bürgergeldbezieher* and *Bürgergeldbezug* derived from *Bürgergeld*).

Each entry is accompanied by descriptive explanations, incorporating domain- and discourse-specific information, as outlined in Chapter 3. Discursive facets that have become entrenched in word meanings are captured in individual entries within a dictionary component dedicated to domains and critical discourse. These can document a range of features associated with each neologism, including evaluative and affective dimensions, stance-taking properties and semantic prosodies (e.g. for headwords such as *Wokismus* 'wokism', *Klimakleber* 'climate activist glueing him/herself onto surfaces'). In some cases, discursive features are related to text genre, linguistic varieties or the medium of communication. In addition, the resource includes cluster or discourse-level entries, which can be accessed via a search interface or alternatively, through the dashboard (see the top right corner of Figure 1).



FIGURE 1 Overview of different entry types, including a navigation menu and a dashboard

These entries compile neologisms constitutive of a given (critical) discourse or domain where they explicate their discursive functions in context. The discourse entries situate neologisms within complex discourse structures that capture intertextual relations and diachronic patterns. They further illustrate the historical emergence and evolution of new lexical items, thereby mapping discourse trajectories and tracing thematic tendencies or shifts in the salience of societal discourses (e.g. CLIMATE CRISIS, ENERGY CRISIS). The two modes of consultation are closely interconnected, see Figure 1. They function as complementary levels, allowing users to explore both the fine-grained properties of individual neologisms and the broader discursive networks in which they are embedded. Chapter 2 highlights the discourse-constitutive roles of neologisms and underscores the need for corpus-assisted approaches to discourse lexicography focussed on neologisms. Chapter 3 shows how these levels intersect and inform one another, highlighting how lexical items participate in, reflect and shape the discursive formations in which they circulate. In Chapter 4, the methods used to obtain the data for each entry type are outlined, with emphasis on word-embedding techniques.

2 Neologisms in Societal Discourses

Newly coined words or expressions play a significant role in societal discourse as they emerge in response to changing social, political, technological and cultural conditions. In this sense, neologisms are not only linguistic innovations but also indicators of social transformation. The majority of German neologisms that have emerged since 2020 can be situated within broader discursive formations addressing a set of partially interrelated societal challenges. Some of these discourses are novel in origin and have generated distinct neo-discursive fields (COVID-19 pandemic, the war in Ukraine). Others represent the continuation or renewed intensification of pre-existing debates, notably those surrounding migration and the climate crisis. The discourse surrounding the energy crisis shows substantial lexical overlap with related domains, including the war in Ukraine and climate discourse. More recently, it has also intersected with discussions of the war in Iran, the disruption of oil trade due to the blockade of the Strait of Hormuz and the destruction of some refinery infrastructure in the Gulf region. Within German political discourse, these fields are further intertwined with discussions on political decisions regarding the energy transition (*Energiewende*), i.e. the shift from fossil fuels and nuclear energy to renewable energy sources. Moreover, the COVID-19 pandemic precipitated significant disruptions to global supply chains, contributing to economic stagnation and recessionary pressures, which in turn culminated in inflationary developments between 2021 and 2024.

Neologisms not only denote but also actively construct a new socio-political reality. The term *Long Covid*, for instance, refers to persistent symptoms following a COVID-19 infection. It remains difficult to conceptualise and manage, as it lies at the intersection of multiple forms of uncertainty: biological, clinical and social. Similarly, *Zeitenwende* ('turning point' or 'watershed moment') has become a politically charged designation for a perceived historical rupture, particularly since the Russian invasion of Ukraine. It implies a fundamental epochal shift, especially in relation to military defence and energy policy. In a comparable vein, *Freiheitsenergie* ('freedom energy') reflects efforts to reduce dependence on imported fossil fuels, particularly from Russia, and to secure energy autonomy through diversification and renewable sources. This emotive and ideologically inflected term frames energy policy not merely in economic or infrastructural terms, but as a matter of liberty and sovereignty. Another illustrative example is *Shrinkflation*, an emotionally charged neologism describing the phenomenon whereby prices remain stable while product quantity or quality diminishes. This framing produces a perception of deception among consumers, who experience inflation as a breach of trust, thereby introducing a moral dimension into what might otherwise be a purely economic concept. A new term may also highlight urgency, suggest innovation or signal political alignment.

Energiewende ('energy transition') illustrates how a single word can encapsulate a broad political agenda, combining technological change, environmental goals and societal restructuring. Similarly, *grüne Transformation* ('green transition') functions as shorthand for complex processes, making them accessible in public communication. Here, neologisms also play a persuasive role. By framing issues in novel terms, they can influence how problems are perceived and debated. However, neologisms can also generate ambiguity or contestation, especially when their meanings are interpreted differently by various social groups (e.g. *Tradwife*, *Wokeness*).

Neologisms are not merely additions to vocabulary; they actively structure perception, frame communication and shape public reactions. Many carry strong emotional charge, generating urgency or anxiety that can mobilise collective responses, but also contribute to fatigue or pessimism over time. Frequently adopted by media, political actors and academia for their expressive power, such terms go on to shape public debate, headlines and policy discourse in more sustained ways (cf. von Dijk 1995).

2.1 Approach to Corpus-Assisted Discourse Lexicography

Discourse lexicography is a specialised dictionary branch typically devoted to contested historical concepts in German (Kämper, 2008, 2012). Its aim is to represent discourse in terms of its "complexity and coherence" (Kämper, 2006, p. 334) by showing how lexical elements contribute to the construction of linguistic and extralinguistic realities. Such dictionaries thus seek to elucidate both the structure of a discourse and the linguistic means through which it is realised. Discourse, in this sense, is a collectively produced, thematically coherent communicative process enacted by a definable set of participants. It unfolds across time through diverse textual forms, fulfils specific functions and is characterised by a distinctive lexical repertoire. It can be analysed on multiple levels, including argumentation structures, lexical patterns, pragmastylistic markers and the perspectives of different speaker groups. These provide access to shared attitudes, emotions, positions and representations of power and (de)legitimisation in public communication. Accordingly, discourse-specific lexical items are inherently relational, forming semantic networks that structure the discourse's subject matter, participants, historical embedding and functional orientation (cf. Kämper, 2006). Capturing and systematising these interrelations is the central aim of discourse lexicography.

To date, discourse dictionaries have neither been based on large-scale corpora nor focused on neologisms. Against this backdrop, the new resource seeks to integrate emerging lexical items into an established discourse-lexicographic framework while extending it through the inclusion of usage-based innovations and the systematic use of large corpora (DeReKo). Corpus-Assisted Discourse Studies (Baker, 2006; Gilling, Mautner & Baker, 2023) combine corpus linguistics with qualitative discourse analysis to examine patterns of language use. By identifying recurring features such as keywords and collocations, CADS enables their interpretation within broader social and ideological contexts. This approach provides systematic, empirically grounded insights into explicit and implicit meanings including the construction of sentiment, attitudes, social realities, identities and power relations.

The extension of this framework to the study of neologisms is motivated by findings in Storjohann (2024), which show that neologisms in emerging discourses occupy specific positions and contribute to the formation of distinct discursive dimensions. Clusters of neologisms construct particular facets of discourse by introducing new conceptual strands, extending existing ones or recalibrating evaluative and argumentative trajectories through novel forms of positioning. They can also encode processes of (de)legitimisation, reflect shifts in power relations and convey new emotional or evaluative orientations. In addition, they may articulate

previously unexpressed attitudes adopted by specific groups within a linguistic community and segment thematic focal points across a discourse's temporal development (cf. Storjohann & Cimander, 2023). On the one hand, analysis of discourse structure shows that neologisms are interrelated and perform discourse-constitutive functions (see 3.1). On the other hand, discourse-relevant features become entrenched in lexical meaning as the semantic content of a new word is continuously negotiated within the speech community. In such cases, they must be explicitly captured in a neologism resource insofar as they shape usage and meaning (see 3.2).

2.2 Documenting new words in discourses

Since 2020, the COVID-19 discourse (CORONAKRISE) has constituted the most prominent global neo-discourse. In the German context, however, several additional discourses have also gained prominence, including political debates surrounding the AMPELKOALITION ('traffic light coalition', 2021–2024), the energy crisis (ENERGIEKRISE 2022–2024, 2026) or migration debates (FLÜCHTLINGSKRISE). Since 2022, the shortage of skilled workers (FACHKRÄFTEKRISE), the rise of the right-wing party AfD (DEMOKRATIEKRISE) and the war in Ukraine (UKRAINEKRIEG) have dominated public communication. A distinct neo-discourse has also emerged around contested aspects of artificial intelligence (KÜNSTLICHE INTELLIGENZ), while the discourse on climate change (KLIMAKRISE), though long established, has taken on new dimensions, particularly following the Fridays for Future movement after 2019. Earlier lexicographic resources (e.g. *Neologismenwörterbuch*, 2006ff.) did not allow for the systematic tracing of neologisms across discourses. This limitation has been addressed with the development of a new resource. Figure 2 presents the most salient discursive fields in media since 2020, highlighting the

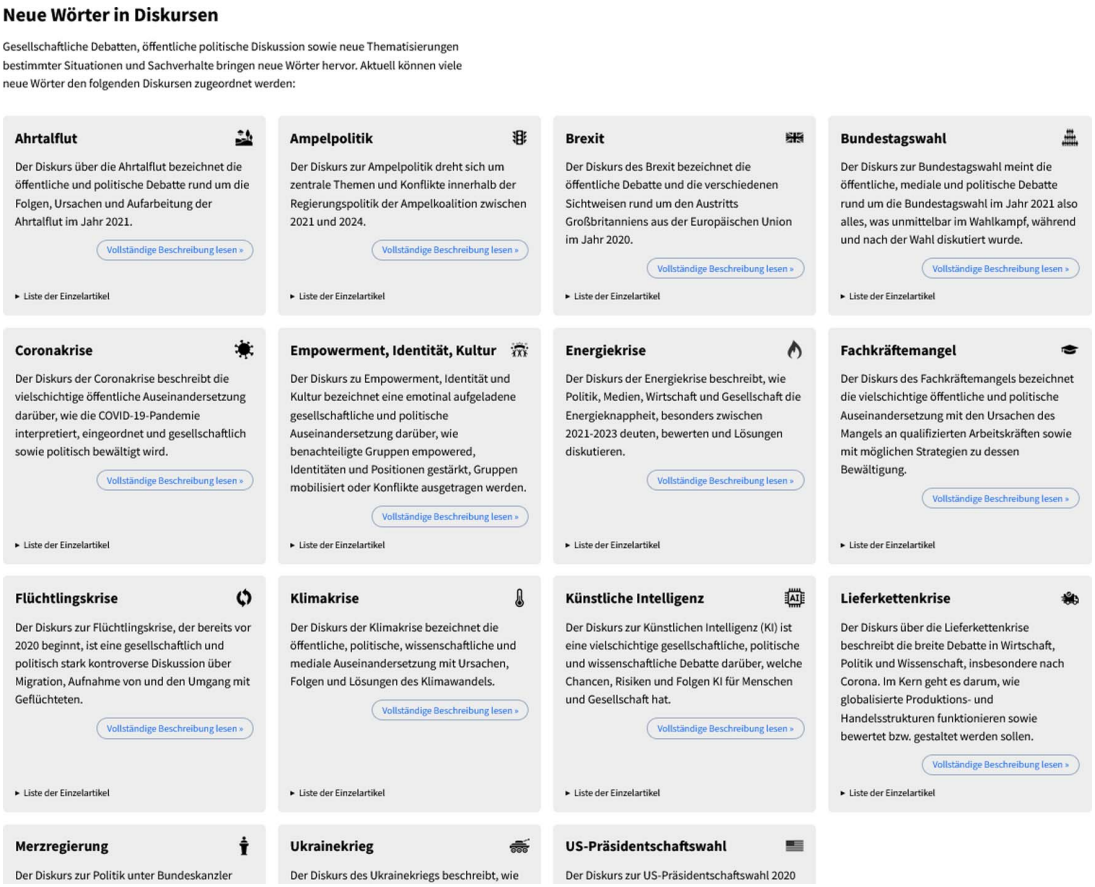


FIGURE 2 Overview of selected (neo-)discourses with relevant neologisms since 2020

central role of neologisms, which can now be systematically explored. The cards/tiles in the grid layout are clickable entry points for exploring and describing discourse properties. They provide access to a more discursive, conceptual and onomasiological mapping of neologisms in communicative use.

This section currently offers concise accounts of contested public debates and will be expanded to better capture the complexity of selected discourses and the role of neologisms within them. Key conceptual aspects are outlined in Section 3.1.

3 One Neologism Resource with Two Distinct but Related Levels

3.1 Documentation Discourse and Cluster Entries

When neologisms arise within a particular discourse, they can occupy distinct positions and serve various communicative and functional purposes (Storjohann, 2024). In particular, neologisms embedded in critical discourses are often integrated into established lexico-semantic relations and can exert a significant influence on the formation of new discursive configurations, i.e. neo-discourses. As Mattfeld, Warnke & Herford (2025, p. 122) observe, certain appellative nouns designating objects or conceptual entities appear to possess particular discursive force[1]. One of the most salient examples of such a rapidly developing neo-discourse is the COVID-19 crisis, which was characterised by a substantial influx of neologisms (Klosa-Kückelhaus, 2022).

> Ein Neodiskurs ist eine Aussagenformation, die ereignisbezogen (ungeplant) auftritt, eine mittelfristig wirksame Dynamik transtextueller Vernetzung mit additiver Tendenz hat sowie referentiell an vorgängige Diskurse gebunden ist, diese transformiert und potentiell selbst neue Diskurse hervorbringt. Ein Merkmal von Neodiskursen sind Neologismen. (Mattfeld, Warnke & Herford, 2025, p. 126)[2]

Neo-discourses have, to date, not constituted a central focus of lexicographic description. Nevertheless, their significance lies in their capacity to provide a more comprehensive perspective on broader societal debates, including patterns of lexical development and processes of lexical competition. They also make visible dominant and marginal sub-discursive structures as well as diachronic shifts within discourse configurations.

Currently, the description of each discourse comprises three components: an alphabetical list of all neologism headwords, each of which is documented as an individual entry (see 3.2), a concise overview of the discourse and its development, including key thematic focal points and the introduction of new conceptual elements, and an illustration of the collocational embedding of the discourse object itself (see Figure 3). Users can navigate to individual single words, as each new expression functions as a hyperlink to its corresponding headword.

[1] Mattfeld, Warnke & Herford (2025, p. 122) characterise such nouns as “diskursmächtig”.

[2] A neo-discourse is a formation of statements that emerges in an event-driven (unplanned) way, develops a medium-term dynamic of transtextual networking with an additive tendency, and is referentially tied to preceding discourses, transforming them and potentially giving rise to new discourses itself. A characteristic feature of neo-discourses is neologisms. (translation by author)



FIGURE 3 Discourse entry of the discourse KLIMAKRISE ('climate crisis')

In a subsequent phase starting in 2028, these discourse entries will be progressively expanded, e.g. by incorporating a navigational structure with a dropdown menu. This enables users to display neologisms in a chronological sequence along a timeline, as well as to organise them according to specific discourse dimensions and/or frequency patterns. In essence, this provides an overview of a data-driven analysis of how neologisms map onto shared conceptual fields within a given discourse. An examination of neologisms over time enables the identification of diachronic shifts in discourse, as illustrated in Figure 4.

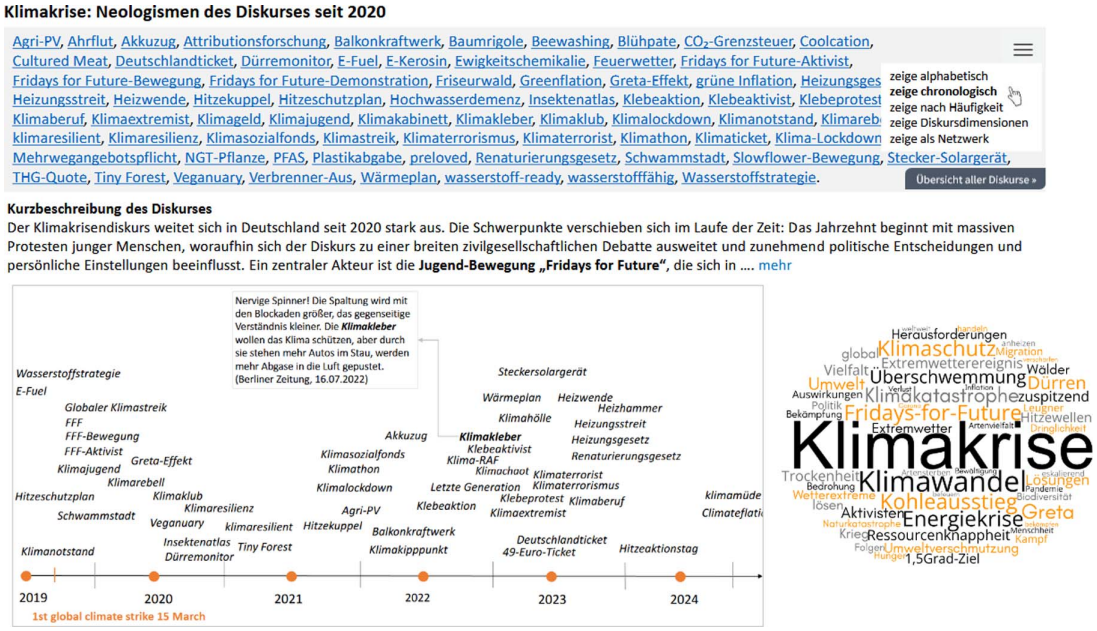


FIGURE 4 Diachronic evolution and emergence of neologisms within climate crisis discourse

Figure 4 indicates that in 2019 the climate discourse was marked by a surge of neologisms associated with global strikes and the Fridays for Future movement. In 2020, the focus shifts toward more academic and monitoring-oriented terminology, including tools tracking environmental conditions such as insect populations (*Insektenatlas*) and drought (*Dürremonitor*). By 2021, attention moves toward climate resilience, with adaptation increasingly emphasised alongside mitigation. In 2022, more radical environmental activism gains visibility, particularly through "Letzte Generation"[3] and its controversial protest tactics (*Klebeaktion*), such as glueing themselves to artworks or road surfaces. This period also sees intensified political debates under the Ampelkoalition (government coalition of the SPD, Grüne and FDP), including contested legislation like the heating law (*Heizungsgesetz*), alongside mobility initiatives such as the nationwide public transport ticket (*Deutschlandticket*). From 2024 onward, the discourse appears to lose momentum as political crises, especially the collapse of the governing coalition dominate attention. Subsequent elections and the formation of a new, more economically oriented coalition (under chancellor F. Merz (CDU)) may further explain the reduced emergence of neologisms in 2025.

Overall, the discourse appears to move from prevention toward consequence, and from mobilisation and optimism to a tone increasingly shaped by urgency and fatigue, as reflected in terms such as *klimamüde* ('climate fatigued'). Systematic analysis of neologisms across distinct discursive dimensions is therefore crucial. As constitutive elements of discourse, neologisms can be grouped into thematic clusters that capture salient patterns, including political measures on climate change, key actors in climate action, representations of the climate crisis and its economic and social impacts, as well as scientific approaches to mitigation and adaptation. Mapping these dimensions helps identify dominant narrative structures and reveals how discourse frames social and environmental realities. In the future, users will be able to browse through such facets for each discourse.

In addition to discourse entries that span multiple headwords within a given discourse, the resource will distinguish further types of cluster entries. For instance, the search interface and dashboard (Figure 1) both provide access to morphologically based clusters grouping lexemes formed through derivation and/or borrowing with recurring bases such as *-Inflation, -Influencer* or *-Washing*. These clusters reflect key aspects of discourse morphology (cf. Michel, 2025; Polanjar, 2024), where compounds and blends encode evaluative and affective stances. Their systematic documentation will be further expanded in future work. A major advantage of clustering formations is the increased visibility of productive word-formation patterns. Grouping items like *Bookfluencer, Finfluencer* or *Reise-Influencer* ('travel influencer') makes it possible to observe how a lexical base is extended across domains through regular morphological processes, rendering word formation productivity analytically accessible. At the same time, these clusters highlight semantic variation and pragmatic differentiation: the shared base *influencer* acquires different meanings depending on its modifiers, ranging from descriptive uses (*Internet-Influencer*) to more evaluative or critical formations (*Brexit-Influencer*). This enables a fine-grained analysis of how discourse communities reshape central concepts across social, economic and cultural contexts. From a discourse-analytical perspective, clustering also increases the transparency of ideological and evaluative dimensions, as many formations carry implicit, often ironic, critical or normative stances that become especially salient in comparison.

[3] "Letzte Generation" is part of a broader ecosystem of climate activism in Europe and is often discussed alongside several related UK-based organisations known for high-visibility civil disobedience such as "Just Stop Oil".

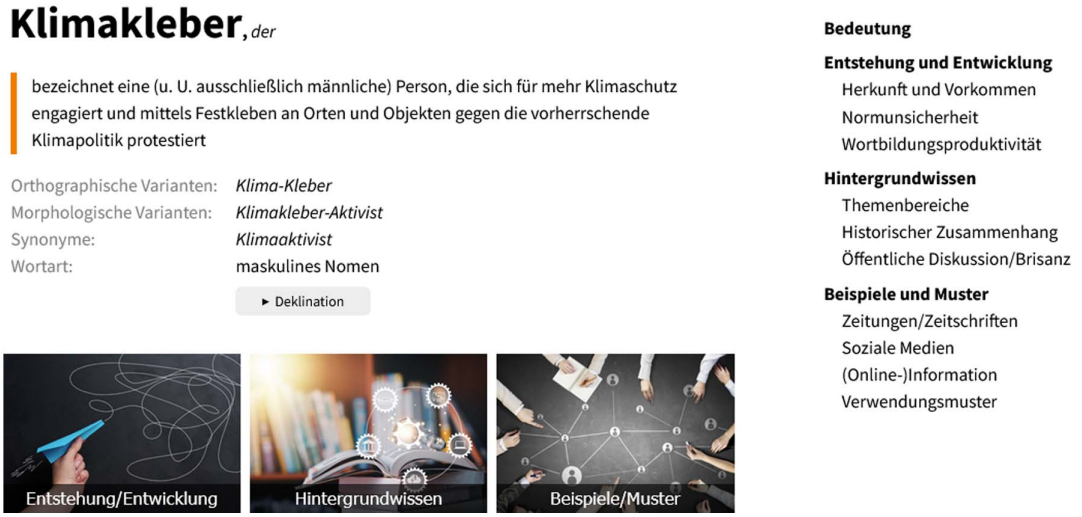


FIGURE 5 Entry for single headword *Klimakleber* and its general architecture/sections

3.2 Documentation of Single Neologisms

In 2022, a range of neologisms such as *Klimakleber, Klimachaot, Klimaextremist, Klima-RAF* and *Klimarebell* emerged to designate activists associated with the activist group "Letzte Generation". Their protest practices became highly visible forms of civil disobedience and included blocking roads by gluing themselves to pavements or attaching themselves to museum artworks. These lexical items are strongly emotive and evaluative, and are predominantly used by critics of climate activism or of these specific protest tactics. They form part of a broader discourse that intensified as demonstrations across Germany caused widespread disruption to traffic. The ways in which speech communities refer to *Klimakleber* directly shaped its semantic development through processes of entrenchment (Schmidt, 2018), whereby meaning is gradually consolidated through repeated discursive use and opposing interpretations among participants, as observed through contextual patterns (collocates).

In contrast, general-purpose dictionaries often lack this contextual dimension, with evaluative and discourse-specific meanings only partially captured (as, for instance, in Duden Online). The new resource therefore seeks to reflect the interaction between abstract lexical semantics and concrete discursive usage.

The sections vary in the extent to which they integrate linguistic and extralinguistic information. In particular, the middle section reflects the discourse-oriented design of the project, as it incorporates encyclopaedic and historical context that helps explain processes of lexicalisation and entrenchment (see Figure 6).

The information presented in Figure 6 (left) covers the domains in which *Klimakleber* is typically used, including politics, climate change, traffic, climate protest and energy policy. The entry also situates the lexeme within the broader climate crisis discourse and provides a direct link to the corresponding discourse entry. Thematically related terms are listed as well, alongside corpus examples illustrating typical usage contexts. In addition, relevant observations on text type and genre are documented within the same section (see Figure 6, left). The historical background shown in Figure 6 (right) provides a concise overview together with key events that have influenced the use of the item, presented in chronological order.

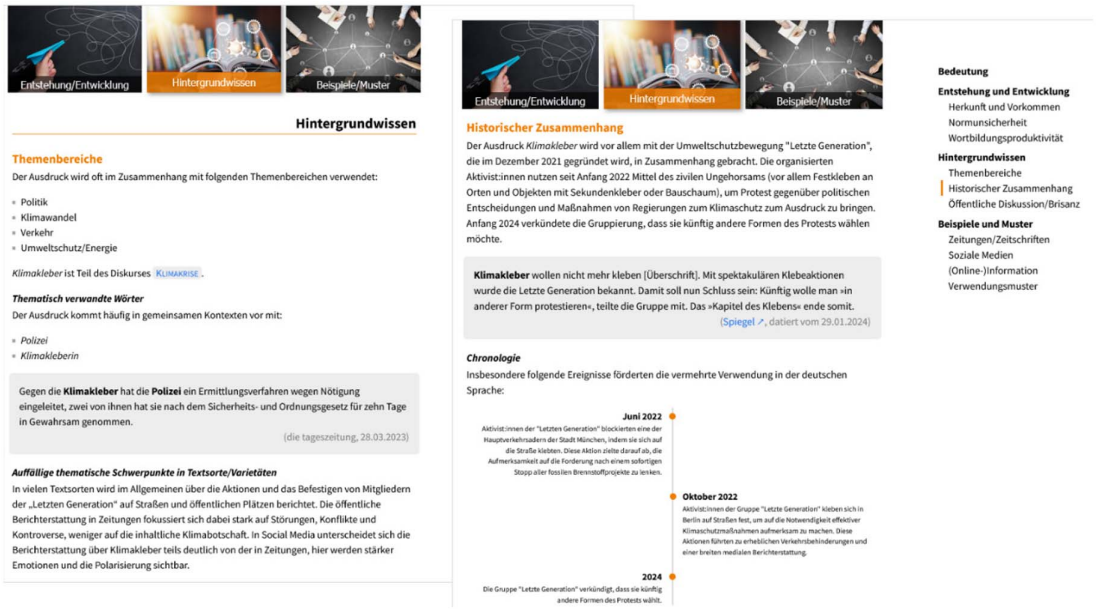


FIGURE 6 Details related to domains/discourse (left) and historical background (right)

Figure 7 documents the public debate on *Klimakleber* by juxtaposing opposing viewpoints. This section most clearly foregrounds the evaluative stances of different speakers, which shape both the groups that adopt the term and the range of affect it conveys. It also reveals the conflicts embedded in public discourse, capturing the core dynamics of political controversy. By presenting multiple perspectives, the discussion emerges not as a unified narrative but as a contested space in which consensus, disagreement and key tensions become visible.

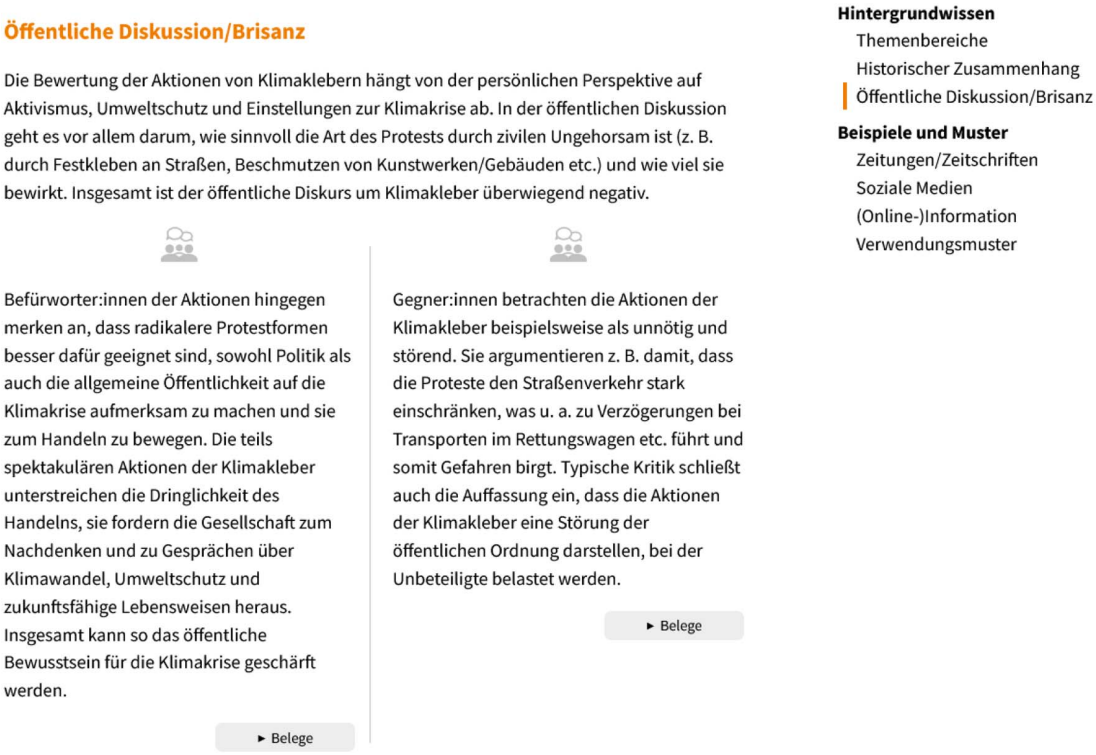


FIGURE 7 Opposing emotions and positions taken in public debates about *Klimakleber*

Such contrasts encourage readers to reflect on their own positioning, whether they align with particular viewpoints and whether they would adopt the terms under debate. This comparison also supports a critical reading of arguments by exposing underlying assumptions and implicit ideologies. Moreover, the differing perspectives typically correspond to distinct social actors such as activists, politicians, industry representatives or the general public, making visible whose voices are represented and whose interests are at stake. Taken together, this constitutes essential semantic and discursive knowledge for participation in real-world communication.

4 Tools and Methods

Baker (2006) shows how corpora are exploited to examine linguistic phenomena relevant to critical discourse studies. Collocation and concordance analysis are established methods for disambiguating and analysing words in authentic communication. The evaluative meaning of the item *Klimakleber* emerges through its typical collocations. Using the German Reference Corpus (DeReKo)[4], the analysis shows that *Klimakleber* is strongly associated with disruption, illegality and social disturbance. Frequent verbal collocates such as *blockieren* ('block'), *stören* ('disturb'), *gefährden* ('endanger') and *lahmlegen* ('paralyse') portray activists as disruptive agents, while nominal collocates including *Blockaden* ('blockades'), *Nötigung* ('coercion'), *Straßenblockaden* ('road blockades') and *Sachbeschädigung* ('property damage') reinforce a frame of criminalisation. References to *Polizeieinsatz* ('police intervention'), *Strafen* ('punishments'), *Präventivhaft* ('preventive detention') and *Knast* ('jail') further position climate protest as a law-enforcement issue rather than political activism. Only a small number of collocates, notably *Sympathie* ('sympathy'), *Verständnis* ('understanding') and *Klimaschutz* ('climate protection'), reflect support for activists or their goals. Qualitative analysis likewise reveals frequent pejorative labels such as *Klimaterroristen* ('climate terrorists'), *Straftäter* ('criminals') and *Denkmalschmierer* ('statue vandals'), alongside negative adjectives including *kriminell* ('criminal'), *rücksichtslos* ('ruthless'), *radikalisiert* ('radicalised') and *militant* ('militant'). Positive evaluative collocates are virtually absent. Overall, the collocational profile demonstrates that media discourse overwhelmingly frames *Klimakleber* as disruptive, deviant and illegitimate. Through repeated use, these evaluative associations become conventionalised, contributing to the semantic prosody of the neologism and reinforcing hostile public perceptions (Kerremans, 2015).

More recently, word embeddings have gained prominence in discourse studies (cf. Bubenhofer, 2020; Feldmüller, 2026). They rely on distributional analysis of collocation patterns between related lexical items and position them in a vector space in proximity to a given search term. Word embeddings are powerful tools because they operationalise one of the core assumptions of CADS: that meaning is relational, contextual, ideological and shaped by patterns of use. With respect to German, the open tool DeReKoVecs explores such patterns on the basis of DeReKo. It enables researchers to model how discourses structure meaning by analysing semantic neighbourhoods, tracing shifts and oppositions, and identifying patterns of framing. Through the visualisation of semantic proximity, implicit ideological patterns become visible by showing how words cluster, polarise or shift in meaning across contexts. In our case, word embeddings have been used in the analysis of *Klimakleber*, a term characterised by polarised evaluations. At least four distinct clusters can be identified in its semantic vicinity and grouped into topics (see Table 1).

[4] DeReKo is a continuously growing corpus of written German texts from a wide range of genres, including newspapers, literary, academic, legal, political and web texts, from which researchers can create custom subcorpora.

TABLE 1 Thematic word clusters identified through word embeddings (DeReKoVecs)

general expressions for climate activists	offensive expressions for people with generally socially unaccepted behaviour	expressions for people with socially unaccepted behaviour in traffic	expressions for people belonging to extreme ends of the political scale
Umweltaktivist, Klimaaktivist, Aktivist, Klimaschützer, . . .	*Chaoten, Störer, Pöbler, Provokateure, Störenfriede, Steinewerfer, . . .*	*Wildparker, Falschparker, Drängler, Autoposer, Raser, Verkehrsrowdy, . . .*	*Neo-Nazis, Faschos, Antifa, AfDler, Querdenker, Linksradikale, . . .*

Many of the words that exhibit strong semantic proximity are inherently offensive or overtly evaluative, often carrying emotional weight and signalling particular political or ideological positions. Numerous items frequently originate from domains such as POLITICS, TRAFFIC and CLIMATE CHANGE. This information is relevant to the descriptions of typical to contextual environments, as shown in Figure 6 (left). Likewise, the ideological orientation of such domains can be inferred through the analysis of collocational patterns. Consequently, word embeddings can serve as a valuable tool to support or complement insights obtained through other corpus–linguistic methods such as collocations or key words (for more details also see Storjohann, in press).

5 Summary

Describing neologisms within their discursive environments is indispensable for capturing essential semantic nuances of recently emerging terms. Given that many recent borrowings and coinages emerge in the context of crises, analysing their contextual behaviour often necessitates attention to stance-taking, evaluative dimensions, processes of politicisation, the enactment of power and legitimisation as well as sentiment. These aspects are central to Critical Discourse Studies, which makes such an approach particularly well suited for integration into the discourse–oriented lexicography of neologisms.

IDS *Neo*$^{2020+}$ is not conceived as a discourse dictionary in the strict sense. Several such resources already exist for historical keywords in German. Rather, its aim is to bridge a methodological gap by integrating a corpus–assisted approach to discourse lexicography with a critical discourse–analytic perspective on neologisms. In doing so, it combines systematic, data–driven lexical descriptions with an examination of the socio-political and ideological dimensions that shape the emergence and use of new lexical items. To achieve this, different interlinked entry types help inform users about meaning, contextual use, domains and various discourse dimensions. IDS *Neo*$^{2020+}$ is not a complete resource but will undergo major updates and expansion in the coming years. Ongoing development will progressively extend its coverage, refine existing entries and integrate new data and analytical perspectives in response to emerging linguistic and discursive phenomena.

References

Baker, P. (2006). *Using Corpora in Discourse Analysis*. Continuum.

Baker, P., Gabrielatos, C. & McEnery, T. (2013). *Discourse Analysis and Media Attitudes: The Representation of Islam in the British Press*. Cambridge University Press.

Bubenhofer, N. (2020). Semantische Äquivalenz in Geburtserzählungen: Anwendung von Word Embeddings. *Zeitschrift für germanistische Linguistik*, 48(3), 562–589.

van Dijk, Teun A. (1993). Principles of Critical Discourse Analysis. *Discourse & Society*, 4(2), 249–283.

DeReKo: *Deutsches Referenzkorpus*. Retrieved July 08, 2026, from https://www.ids-mannheim.de/digspra/pb-s1/projekte/korpora/

DeReKoVecs: Retrieved July 08, 2026, from https://corpora.ids-mannheim.de/derekovecs

Duden Online. Retrieved July 08, 2026, from https://www.duden.de/

Feldmüller, T. (2026). *Das Framing von Extremismusvarianten korpuslinguistisch explorieren und visualisieren: Distributionell-semantische Zugänge als Chance für die Kulturstudien*. Language Science Press.

Gilling, M., Mautner, G. & Baker, P. (2023). *Corpus-Assisted Discourse Studies*. Cambridge University Press.

Heylen, K., Kernerman, I., Freixa, J., Heinonen, T., Jung, H.-Y., Lorentzen, H., Nam, K., Rouxel, B., Salgado, A., Sánchez Ibáñez, M., Storjohann, P. & L. Trap-Jensen (in prep.). The ENEOLI matrix of neologism selection criteria: A comparative framework for lexicographic practice. *International Journal of Lexicography*.

Kämper, H. (2006). Diskurs und Diskurslexikographie. Zur Konzeption eines Wörterbuchs des Nachkriegsdiskurses. *Deutsche Sprache*, 4, 334–353.

Kämper, H. (2007). *Opfer – Täter – Nichttäter: Ein Wörterbuch zum Schulddiskurs 1945–1955*. IDS Mannheim, Retrieved July 08, 2026, from https://www.owid.de/wb/disk45/einleitung.html

Kämper, H. (2012). *Wörterbuch zum Protestdiskurs 1967/68*. DS Mannheim, Retrieved July 08, 2026, from https://www.owid.de/wb/disk68/start.html

Kerremans, D. (2015). *A Web of New Words. A Corpus-based Study of the Conventionalization Process of English Neologisms* (= *English Corpus Linguistics* 15). Peter Lang.

Klosa-Kückelhaus, A. (2022). German Corona-related neologisms and their lexicographic representation. In A. Klosa-Kückelhaus & I. Kernerman (Eds.), *Lexicography of Coronavirus-related Neologisms* (= *Lexicographica*, *Series Maior* 163) (pp. 27–42). De Gruyter.

Klosa, A. & H. Lüngen (2018). New German words: Detection and description. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (Eds.), *Proceedings of EURALEX 2018: Lexicography in Global Contexts* (pp. 559–569). Ljubljana University Press.

Michel, S. (2025). *Diskursmorphologie: Ansätze und Fallstudien zur Schnittstelle zwischen Morphologie und Diskurslinguistik*. De Gruyter.

Lewandowska-Tomaszczyk, B. & Wilson, P. (2022). Morphology and Emotion. In G.L. Schiewer, J. Altarriba & B. Chin Ng (Eds.), *Language and Emotions. An International Handbook*, 1 *(= Handbooks of Linguistics and Communication Science,* 46/1) (pp. 423–441). De Gruyter.

Partington, A./Duguid, A./Taylor, Ch. (2013). *Patterns and meanings in discourse: theory and practice in corpus-assisted discourse studies (CADS)*. John Benjamins.

Polajnar Lenarčič, J. (2024). Greenwashing, Pinkwashing, Whitewashing . . . Der kritische Metasprachdiskurs um X-Washing. Eine korpusgestützte diskurslinguistische Analyse. *Linguistica*, 1, 41–77.

Schmid, H.-J. (2018). *The dynamics of the linguistic system. Usage, conventionalization, and entrenchment*. Oxford University Press.

Storjohann, P. (in press). Neologisms and Emotionalisation in Contemporary Discourse: A Lexical Approach to Meaning, Context and Perspective. In B. Lewandowska-Tomaszczyk & M. Trojszczak (Eds.), *Opinion in Language, Media & Education*. Springer.

Contact information

Petra Storjohann
Leibniz-Institut für Deutsche Sprache
storjohann@ids-mannheim.de

Silga Sviķe

A LEXICOGRAPHICAL ANALYSIS OF NEOLOGISM DICTIONARIES BASED ON THE MULTILINGUAL DICTIONARY OF NEW WORDS (2010–2024)

Abstract This paper examines the mega-, macro-, and microstructure of all eight editions (2010–2024) of the Multilingual Dictionary of New Words (MDNW), published by the European Parliament's Directorate-General for Translation, focusing on the treatment of Latvian and Lithuanian neologisms. Combining qualitative source analysis with quantitative and qualitative analysis of the extracted material, the study compiled a dataset of 184 neologism occurrences, corresponding to 87 unique lexical units (69 Latvian, 18 Lithuanian). The analysis shows that the dictionary's macrostructure evolved from an early, hierarchically nested entry format toward a flatter, integrated one, and that its editorial model shifted from a cumulative index that retained every Latvian headword extracted from the 4th through the 7th editions (2013–2019) to a non-cumulative index in the 8th edition (2024) that included only the most recently attested words. These findings indicate that the MDNW applies no single, stable criterion of lexical "novelty" throughout its publication history, raising the broader question of how lexicographers should define the temporal scope of a "new word" and communicate it to dictionary users.

Keywords Neologism; new word; dictionary of new words; lexicography

1 Introduction

Neologisms are one of the categories of words that have long attracted the interest not only of the general public but also lexicographer. A *neologism*, or lexical innovation, refers to a recently created word or an existing word used in a novel sense, neither of which may yet be widely accepted by all speakers within a particular speech community (Cluver, 1989, p. 50). A *new word* is a broader term. It can refer to any word that is treated as new within a specific context, time period, or dictionary edition, even if it has already passed beyond the initial stage of neologism. A neologism specifically emphasises the process of a word's creation and its initial acceptance by society, whereas the term new word encompasses any recent addition to the lexicon. In linguistic research, the concept of 'recent' in relation to neologisms is inherently relative and historically bound. Different authors propose various criteria for defining this novelty, ranging from subjective perception to strict chronological frameworks. In Latvian linguistics, the dominant view is held by Alīse Laua (1969, p. 189), who states that words are considered neologisms only as long as they are perceived as new by the language users. This "feeling of novelty" is subjective (Baldunčiks, 2024 [2016], p. 265); once users become accustomed to the word and it assimilates, it ceases to be a neologism. Andrejs Veisbergs and his co-authors suggest that the journey from an occasionalism (an individual coinage) to the status of a neologism usually takes at least three to five years (Veisbergs et al., 2009, pp. 211–223). During this period, it is observed whether the word becomes established in general usage. Researchers at the Leibniz Institute for the German Language (IDS), such as Petra Storjohann and Luisa Cimander, often use a 10-year-period in neologism lexicography (Storjohann & Cimander, 2022, pp. 25–50). Neologisms are not "one-day flies" (German – *Eintagsfliegen*) but units that have entered the general language and become established in it within a specific decade. (Steffens, 2007, p. 148). Ladislav Zgusta emphasises that a neologism is a relative unit that can only be identified within the context of a specific period (Zgusta, 1971, p. 41). What was a neologism yesterday may be an ordinary word today and become obsolete tomorrow. Many authors

(e.g., Baldunčiks, 2024 [2016], pp. 259–275; Ozola & Helviga, 2020, pp. 290–306) define a word as "new" if it is not recorded in general explanatory dictionaries of previous periods. This could be an objective and practical criterion for researching neologisms in general; however, it cannot be applied when studying specialised dictionary of neologisms.

In this study, both terms — *neologism* and *new word* — are understood in their broadest sense. The information summarised above shows that there is no single, universally accepted definition of what constitutes a "new word." Determining the "recency" of neologisms in research is relative, as various authors define it on the basis of either a subjective feeling of novelty, a fixed chronological framework (usually ranging from 3 to 10 years), a significant historical turning point, or the objective lexicographic criterion — the absence of the word from general dictionaries published in previous periods. A word may be considered new even after a specific period of time has elapsed — for example, following the onset of the COVID-19 pandemic or a particular historical era. This makes the task all the more challenging for lexicographers who compile such dictionaries. While usage frequency is one of the criteria for including words in general-purpose dictionaries, the same criterion cannot be applied to neologisms. Boris Kern focuses on the statistical analysis of words and their frequency in corpora. Kern concludes that word frequency in a corpus systematically decreases as the derivational stage of a word within a word-formation chain increases, and this finding can serve as a practical criterion for deciding on inclusion in a dictionary (Kern, 2024, pp. 463–472). The views discussed concerning the nature of neologisms and their inclusion in dictionaries are contradictory. On the one hand, a word must be relatively new, but the criterion for "newness" is not unambiguous; on the other hand, in order for such a word to be included in a dictionary, it must also be sufficiently widespread; that is, the word must have been fully integrated into the language. In addition to a structural analysis of the *Multilingual Dictionary of New Words* (MDNW), this study examined the inclusion of Latvian and Lithuanian neologisms by comparing eight editions.

2 Theoretical Background and Methodology

The selection of entry words has not been the subject of extensive study in either general or specialised dictionaries; nevertheless, selection principles constitute some of the most critical criteria for evaluating dictionary quality (Baldunčiks, 2012, pp. 116–117; Bergenholtz, 2001, pp. 12–14; Peina, 2023, p. 53; Sviķe, 2016). This observation applies to research on general-purpose, bilingual, and specialised dictionaries alike. Henning Bergenholtz (Bergenholtz 2001, pp. 12–14) analysed 252 dictionary reviews and found that 22% of the remarks were related to the selection of entry words. Here, this selection criterion applies to specialised dictionaries of new words. In particular, neologisms are of interest to lexicographers because they raise the question of whether and how to include these words in a dictionary. With regard to dictionaries of neologisms, a different consideration emerges: a word's inclusion in this category is determined by its "freshness". That is, it remains a neologism only until its usage becomes established and widespread — a threshold that is itself rather subjective, since establishing a word's actual currency with precision would require a broader study of its use within the period concerned. Furthermore, the question remains whether a word retains its status as "new" once it has been included in a dictionary. For example, in the field of information technology, the coinage of new words is particularly common, and these newly created words quickly become established in the language. Thus, electronic dictionaries can promptly include them with a specific label and change it just as quickly once the words are widely known and established. However, the compilation of a specialised dictionary of neologisms raises several methodological questions that need to be addressed. The aim of this research is to analyse the lexicographical structure of the MDNW (2010–2024) and trace the inclusion and stability of Latvian neologisms across its eight editions. The research questions are as follows: To what extent do the selected Latvian lexical units exhibit chronological stability across subsequent

editions, and what linguistic or contextual factors characterise the neologisms that continue to appear across several consecutive editions? Chronological stability in this paper refers to the timeframe during which an individual coinage (occasionalism) transitions into a neologism and becomes established in the language. This process typically requires 3 to 5 years or an observation period of up to 10 years to ensure the word is not merely a short-lived coinage.

This study analyses eight editions of the MDNW (2010, 2011, 2012, 2013, 2014, 2016, 2019, 2024) published by the European Parliament's Directorate-General for Translation. The paper analyses the mega-, macro-, and microstructure of all eight editions, as well as Latvian and Lithuanian neologisms (referred to as "new words" in the dictionary) that have been collected and included in these editions. Latvian and Lithuanian are the only two living languages of the Baltic language family, which is why Lithuanian neologisms were included in the analysis as the closest available substitute for the period of 2010–2012 before the Latvian-language section was added to the MDNW project. This approach makes it possible to compare trends across all editions of the MDNW. The study uses the method of a qualitative analysis of theoretical sources. Data collection involves extracting neologisms from all eight editions, while data analysis employs quantitative and qualitative methods. The dataset comprises 149 occurrences of Latvian neologisms across editions 4–8 (counting repeated occurrences of the same lexeme in successive editions), corresponding to 69 unique Latvian lexical units. Lithuanian neologisms, included only in editions 1–3 (before the Latvian language section was included in the project), amount to 35 occurrences/18 unique units. In total, the dataset comprises 184 occurrences and 87 unique lexical units across both languages (see Table 1).

TABLE 1 Distribution of Lithuanian and Latvian neologisms across MDNW editions

Edition	Total extracted neologisms in edition (incl. repeated units)	New (unique) words	Cumulative unique words
MDNW 2010	5	5	5
MDNW 2011	12	7	12
MDNW 2012	18	6	18
Subtotal (Lithuanian)	**35**	**18**	**18**
MDNW 2013	20	20	20
MDNW 2014	32	12	32
MDNW 2016	33	9	41
MDNW 2019	43	10	51
MDNW 2024	21	18	69
Subtotal (Latvian)	**149**	**69**	**69**

3 Analysis of the Dictionary's Structural Levels

Before analysing the dictionary editions and the extracted language material, it is necessary to provide an overview of the dictionary's structure. At the megastructural level, all components of the dictionary texts — including the preface, usage notes, and appendices — are identified and classified by analysing their interrelationships, sequence, and functions. Macrostructural analysis involves examining the organised list of entries and evaluating the selection criteria for main and secondary entries, the principles of alphabetical ordering, and the approach to representing dictionary entries. At the microstructural level, the content and layout of individual

entries are studied in detail by analysing the formal information provided (spelling, pronunciation, grammar) and the semantic explanation, such as definitions, equivalents, and usage examples. All claims in this paper regarding the structure, scope, and editorial principles of the MDNW are based on the forewords and the content of the respective editions and are cited by year of publication, but full bibliographic details of all eight MDNW editions consulted (2010–2024) are provided in the paper's reference list. The number of languages represented in the MDNW gradually increased from 13 in the 1st edition (2010) to all 24 official languages of the European Union by the 6th edition (2016), with Latvian first appearing in the 4th edition (2013) and Romanian — the last language to be included — in the 6th edition (2016).

3.1 Megastructure of the Dictionary

The megastructure of a dictionary integrates the central wordlist (macrostructure) with the components of its front, middle, and back matter (Hartmann, 2001, p. 61; Hartmann & James, 1998, p. 93). The analytical framework developed by Karpinska (2012, p. 64) divides the outside-matter components — following Svensén's (2009) functional classification — into five basic categories: A) metafunctional components providing details about the dictionary itself and its use (the preface and the user's guide); B) components providing encyclopaedic information; C) components providing linguistic information; D) miscellaneous components without a clear function; and E) components belonging to the access structure (the table of contents, running heads, and thumb indices). The aim of the present analysis is to establish the scope of the outside matter components, their typical composition, and their position in the front matter or back matter, as well as to evaluate the content of the most relevant components: the preface and the user's guide. An examination of all eight editions of the MDNW using the megastructural analytical framework proposed by Karpinska (2012) — drawing on Hartmann (2001) and Svensén (2009) — reveals a consistently minimalist outer-text architecture (see Table 2).

TABLE 2 Presence of outer-text components across MDNW editions

Components	1st–5th ed. (2010–2014)	6th–8th ed. (2016–2024)
Preface (A)	✓	✓
User's guide (A)	✗	✗
Encyclopaedic information (B)	✗	✗
List of language abbreviations (C)	✓	✓
General/language-sorted index (E)	✗	✓

Across all editions, the foreword (Category A) constitutes the sole recurring metafunctional component in the front matter, while the user's guide — theoretically regarded as the only obligatory outer text — is systematically absent. Categories B (encyclopaedic components, referring in this context to "outside matter" that supplements the list of main entries with factual information) and D (miscellaneous components) are unattested in any edition. According to Karpinska, "miscellaneous components without a clear function" are outside-matter elements within a dictionary's megastructure (such as tables of weights and measures or mathematical symbols) that provide factual information with an unclear link to the dictionary's primary linguistic purpose, which modern lexicography suggests should be omitted or integrated directly into the headword list to avoid redundancy (Karpinska, 2012, pp. 64–65). No such elements were found in the editions analysed. Linguistic information (Category C) is reduced to language-code lists and abbreviation keys, features that appear consistently only

since the sixth edition (2016) onwards. The exception is the access structure (Category E), which becomes the functional centre of the megastructure from the sixth edition (2016): a general index and a language-sorted index in the back matter address the retrieval challenges inherent in the integrated multilingual new word list. All eight MDNW editions lack a user's guide—the only outer-text component considered obligatory in lexicographic theory — forcing users to infer entry microstructure, grammatical labels, language codes, and transcription conventions independently, which is particularly demanding across 24 languages and multiple writing systems. While the dictionary's intended audience — European Parliament translators — may interpret its conventions intuitively, external users receive no guidance on new word selection criteria or entry decoding. From the sixth edition (2016), a dual back-matter index partially facilitates retrieval, yet the absence of a user's guide remains the most significant weakness of the megastructure.

2.2 Macrostructure of the Dictionary

Compared with microstructure, the macrostructure of dictionaries has traditionally received less attention in academic research (Karpinska, 2012, p. 54), even though it forms the foundation of a dictionary. In dictionary design, the macrostructure denotes "the overall list structure which allows the compiler and the user to locate information" (Hartmann & James, 1998, p. 91). Most Western dictionaries adopt an alphabetical wordlist as their central component, though headwords may also be ordered thematically, chronologically, or by frequency, and the list may be supplemented by front, middle, or back matter. Constructing a dictionary's headword list requires the application of three sets of criteria — lexical type (single items vs multiword units), morphological type (inclusion of derived or inflected forms), and semantic type — while editorial decisions must also distinguish between criteria governing the headword list itself and those governing the entry structure, including the presence or absence of secondary headwords (Atkins, 2008, pp. 39–40). Concerning the arrangement of headwords, a strictly alphabetical macrostructure yields flat entries only, whereas the incorporation of secondary headwords produces a tiered structure organised according to morphological, grammatical, or orthographical criteria for main headwords, and morphological or lexical criteria for secondary headwords (Atkins, 2008, p. 40).

Across the analysed editions of the MDNW (2010–2024), an alphabetical macrostructure consistently serves as the primary organisational axis. Early editions (2010–2013) display divergent structural arrangements, alternating between a unified multilingual neologism list and their separation into language blocks. This macrostructural orientation stabilised from the 5th edition (2014) onwards, where separate language sections effectively avoided cross-linguistic script conflicts while maintaining a highly complex, tiered nesting system for secondary headwords. This structural shift can be illustrated concretely. In MDNW 2013, the entry *draugošanās* ('being active on a social network') is followed by a bare secondary form, *draugoties* ('to be active on a social network'; *-ojos*, *-jies*, *-jas*, past: *-ojos*), listed without its own gloss or translation—a tiered, tightly nested arrangement. This exact configuration persists in MDNW 2014, where both entries retain the same wording, definitions, and [2013] dating label, confirming that the nested microstructure remained stable across the 4th and 5th editions. In MDNW 2016, however, *draugoties* appears instead as a fully independent headword, complete with a transcription, a register label, Latvian and English glosses, and an etymological note, while draugošanās is no longer included in the macrostructure. A similar case is *pikšķerēšana* ('phishing ') in both MDNW 2013 and MDNW 2014; the entry nests two unglossed secondary forms, *pikšķerēt* ('to phish') and *pikšķerētājs* ('phisher'); in MDNW 2016, only the fully specified *pikšķerēšana* remains as a flat entry, and the secondary derivatives are no longer itemised. The persistence of this nested arrangement through 2014, followed by its disappearance in 2016,

indicates that the shift to a flatter microstructure occurred specifically between the 5th and 6th editions rather than gradually over the 2013–2016 period. These examples indicate that later editions replaced hierarchical nesting with a flatter, more uniformly structured entry format. The entry-selection criteria encompass both single-word lexemes and multiword units. Initially, the morphological coverage and the system of secondary headwords lacked editorial uniformity across language teams, which resulted in an inconsistent tiered structure. In the later editions (2016–2024), the macrostructure stabilised as an integrated multilingual A–Z headword list underpinned by an extensive outside-matter component, particularly through language-specific indices. Across successive editions, the dictionary evolved from an overtly hierarchical, nested entry layout towards a visually flatter macrostructure that integrated secondary derivatives directly into the microstructure.

3.3 Microstructure of the Dictionary

The term *microstructure* refers to the internal design and format of an individual dictionary entry, detailing how various categories of information about a headword are structured and presented for the benefit of the user (Hartmann & James, 1998, p. 94). According to Karpinska (2012, pp. 99–102), the analysis of dictionary microstructure consists of five stages: (1) Structural division: The entry is divided into two mandatory components: the formal structure of the left core (headword) and the semantic structure of the right core (explanation and translation); (2) Left core analysis: The headword's formal data are examined, including spelling variants, phonetic transcription, part-of-speech labels, and grammatical descriptions; (3) Right core analysis: Semantic and pragmatic aspects are investigated, focusing on meaning, usage markers, and context indicators; (4) Contextual analysis: The integration of illustrative examples and multi-word units is evaluated; (5) Typological classification: Entries are classified using Hausmann and Wiegand's ([1989] 2003) concept by example placement (integrated, non-integrated, or partially integrated microstructures) to determine the dictionary's macrostructural or microstructural orientation.

For the analysis of the dictionary's microstructure, a purposive sample of thirteen entries was drawn from two continuity-based pools. For Lithuanian, five entries — *brukalas* (EN: 'spam; an unsolicited email sent to a large number of addresses'), *išmanùsis* ('a smartphone'), *príegloba* ('hosting'), *riẽdis* ('Segway Personal Transporter'), and *tinklãraštis* ('a blog') — were selected as the only lexemes attested in all three early editions (MDNW 2010–2012). For Latvian, five entries were randomly sampled from the seven neologisms shown in Table 4 (see below) that remained in the dictionary's cumulative index from MDNW 2013 to MDNW 2019: *glābējsilīte* ('a place where people, typically mothers, can bring babies, usually newborn, and leave them anonymously in a safe place to be found and cared for'), *draugoties* ('to be active on the Latvian online social network draugiem'), *mākoņdatošana* ('data storage and processing on remotely accessed servers'), *ēnstrādnieks* ('an illegally employed worker, one of the elements of the shadow economy') and *bezstils* ('behaviour or fashion sense that does not comply with society's perception of good taste; a chaotic and unsuccessful mixture of different fashion styles'). To extend the comparison to the post-2019 period, three entries retained from MDNW 2019 to MDNW 2024 — *iedabisks* ('inherent'), *kaceklis* ('a teaser; short video clip about an upcoming film or TV show, intended to stimulate interest'), and *slimstrāde* ('presenteeism; working while sick') — were also added to the sample, as they constitute the only lexemes exhibiting continuity across the later edition. The selected examples are examined in all editions in which they appear, and only their most significant features are described below. However, these entries offer only limited insight into broader structural trends, and they cannot be considered statistically representative of the dictionary's entire lexical corpus. Given the

limited scope of this article, only a few illustrative examples are provided below, which highlight some trends that nevertheless characterise the dynamics of microstructures.

Across all eight editions, the structural division demonstrates a transition from rigid, paper-based typographic constraints to screen-optimised digital layouts, with the three most recent editions (2016, 2019, 2024) available online. The earliest editions of the MDNW (2010, 2011, 2012) maintain strict visual boundaries with hard paragraph breaks separating formal parameters from semantic data, e.g., the language code, the part-of-speech label — "*brùkalas (LT) n. nepageidaujami elektroniniai laiškai; EN spam; an unsolicited email sent to a large number of addresses. Brukalų siuntėjas, kuriam skirta didžiausia istorijoje bauda, gali išvengti bausmės, nes gyvena Kanadoje, o bauda buvo skirta JAV. brukãlius*"; but from MDNW 2013 onwards the entries include indications of the grammatical number and the year the new word came into use – "*brùkalas (lt) n. [2010] nepageidaujami elektroniniai laiškai; EN spam; an unsolicited email sent to a large number of addresses. Brukalų siuntėjas, kuriam skirta didžiausia istorijoje bauda, gali išvengti bausmės, nes gyvena Kanadoje, o bauda buvo skirta JAV. brukãlius*". From this edition (2013) onwards, the entry also includes inflected forms and past tense forms — "*draugoties (LV) -ojos, -jies, -jas; past: -ojos*". The left core includes an explanation in Latvian (Lithuanian) and English, as well as a usage example for the respective new word, followed — starting with the 2013 edition — by a noun reference for an entry that is a verb and a verb reference for an entry that is a noun (with an indication of number and grammatical gender). The right core significantly expands its deployment of usage labels by steadily replacing general stylistic and pragmatic markers — such as internet slang, ironical, youth language, and informal — "*bezstils (LV) vsk., v. [2013] darbības, rīcības, izturēšanās, runas, ģērbšanās u.c. īpatnību kopums, kas neatbilst sabiedrības izpratnei par labu gaumi, haotisks dažādu stilu sajaukums; en. such a behaviour or fashion sense which doesn't comply with a society's perception of a good taste, chaotic and unsuccessful mixture of different fashion styles; lack of fashion sense. Glamūrs, stils un bezstils "Jaunajā vilnī 2012""* (MDNW 2013, p. 21; MDNW 2014, p. 124) and "*bezstils ['bɛsstiłs] n. m., informal. LV: izturēšanās un ģērbšanās īpatnību kopums, kas neatbilst sabiedrības izpratnei par labu gaumi; haotisks dažādu stilu sajaukums. Dzeltenā prese pievērš ārkārtīgi lielu uzmanību bezstila problēmai. EN: behaviour or fashion sense that does not comply with society's perception of good taste; chaotic and unsuccessful mixture of different fashion styles (< bez + stils = without + style)*" (MDNW 2016, p. 22). Starting with MDNW 2016, the English translations of definitions are supplemented by an explanation of the word's morphology; parts of the text are shown in different colours (e.g., the entry word and the explanation in English). Contextual examples can be found in all editions, but no sources are cited for them, which would be very useful, for example, for verifying the authenticity of these examples. It is important to note that the microstructure of MDNW 2024 includes an audio pronunciation feature, which can be accessed by clicking on the speaker icon.

4 Description and Analysis of the Dataset: Latvian and Lithuanian Neologisms

The data show that, although the analysed dictionary is multilingual, neologisms in all 24 languages are only included from the 6th edition onwards. The dataset consists of data extracted from new-word entries in Lithuanian and Latvian in all eight editions of the MDMW (see Table 3). Table 3 lists 184 total occurrences: 87 unique lexical units, of which 18 are Lithuanian and 69 Latvian; 97 repeated occurrences across all analysed editions.

TABLE 3 Neologisms extracted from each MDNW edition, with English equivalents (184 total mentions: 87 unique lexical units, of which 18 are Lithuanian and 69 Latvian; 97 repeated occurrences across editions)

Edition	Extracted neologisms (English equivalent)
MDNW 2010 (LT)	*brukalas* – spam; *išmanùsis* – smart, e.g., smartphone; *príegloba* – hosting; *riẽdis* – Segway; *tinklāraštis* – blog
MDNW 2011 (LT)	*atrankà* – casting; *brukalas* – spam; *budùlis* – boor, uncultured person; *išmanùsis* – smart, e.g., smartphone; *plitùsis vaĩzdo įrašas* – viral video; *póvakaris* – afterparty; *príegloba* – hosting; *riẽdis* – Segway; *sąspietis* – flashmob; *sė́dmaišis* – bean bag chair; *tinklãlaidė* – podcast; *tinklāraštis* – blog
MDNW 2012 (LT)	*atrankà* – casting; *brukalas* – spam; *budùlis* – boor, uncultured person; *dokumentíena* – mass of sensational documents; *draugišius* – obsessive online friend-seeker, pej.; *išmanùsis* – smart, e.g., smartphone; *núožvalga* – face control; opinion; *pãsvetainė* – subsite; *plitùsis vaĩzdo įrašas* – viral video; *póvakaris* – afterparty; *príegloba* – hosting; *riẽdis* – Segway; *sąspietis* – flashmob; *sė́dmaišis* – bean bag chair; *seimáuti* – to dawdle as an MP; *tinkìnti* – to customise; *tinklãlaidė* – podcast; *tinklāraštis* – blog
MDNW 2013 (LV)	*bezstils* – lack of fashion sense; *draugošanās* – being active on a social network; *draugoties* – to be active on a social network.; *ēnstrādnieks* – illegally employed worker; *glābējsilīte* – a place where people (typically mothers) can bring babies, usually newborn, and leave them anonymously in a safe place to be found and cared for; *hipsteris* – hipster; *[Jūrmal]-geita* – -gate scandal; *mākoņdatošana* – cloud computing; *pikšķerēšana* – phishing; *pikšķerēt* – to phish; *pikšķerētājs* – phisher; *retrīts* – retreat; *stulbings* – planking-type phenomenon; *talkot* – to do voluntary community work; *talka* – voluntary community work; *tvikšanās* – tweetup; *tviteris* – Twitter; *tvitosfēra* – Twittersphere; *urla* – member of an aggressive urban subculture; *zibakcija* – flashmob
MDNW 2014 (LV)	*bezstils* – lack of fashion sense; *bitmonēta* – bitcoin; *čivināt* – to tweet; *draugošanās* – being active on a social network; *draugoties* – to be active on a social network.; *ēnstrādnieks* – illegally employed worker; *glābējsilīte* – a place where people (typically mothers) can bring babies, usually newborn, and leave them anonymously in a safe place to be found and cared for; *gūglēt* – to google; *hipsteris* – hipster; *[Jūrmal]-geita* – -gate scandal; *koučings* – coaching; *koučsērfings* – couchsurfing; *mākoņdatošana* – cloud computing; *mēms* – meme; *pašiņš* – selfie; *pešmerga* – peshmerga; *pikšķerēšana* – phishing; *pikšķerēt* – to phish; *pikšķerētājs* – phisher; *planšetdators* – tablet computer; *retrīts* – retreat; *spaiss* – synthetic cannabis; *stulbings* – planking-type phenomenon; *talkot* – to do voluntary community work; *talka* – voluntary community work; *tvikšanās* – tweetup; *tviteris* – Twitter; *tvitosfēra* – Twittersphere; *urla* – member of an aggressive urban subculture; *vebinārs* – webinar; *vīterot* – to tweet; *zibakcija* – flashmob
MDNW 2016 (LV)	*bezstils* – lack of fashion sense; *bitmonēta* – bitcoin; *čivināt* – to tweet; *draugoties* – to be active on a social network; *ēnstrādnieks* – illegally employed worker; *glābējsilīte* – a place where people (typically mothers) can bring babies, usually newborn, and leave them anonymously in a safe place to be found and cared for; *gūglēt* – to google; *hibrīdkarš* – hybrid war; *hipsteris* – hipster; *kebabnīca* – kebab fast-food restaurant; *koučings* – coaching; *koučsērfings* – couchsurfing; *mākoņdatošana* – cloud computing; *medijpratība* – media literacy; *mēms* – meme; *pakomāts* – parcel locker; *pašiņš* – selfie; *pešmerga* – peshmerga; *pikšķerēšana* – phishing; *planšetdators* – tablet computer; *retrīts* – retreat; *spaiss* – synthetic cannabis; *stulbings* – planking-type phenomenon; *supošana* – stand-up paddleboarding; *trollēt* – to troll; *tvikšanās* – tweetup; *urla* – member of an aggressive urban subculture; *vebinārs* – webinar; *velorikša* – cycle rickshaw; *vidisks* – environmental; *vīterot* – to tweet; *zaļināšana* – greening; *zibakcija* – flashmob

TABLE 3 (*CONTINUED*)

Edition	Extracted neologisms (English equivalent)
MDNW 2019 (LV)	*amatalus* – craft beer; *bezstils* – lack of fashion sense; *bitmonēta* – bitcoin; *čivināt* – to tweet; *draugoties* – to be active on a social network; *dzērājšoferis* – drunk driver; *ēnstrādnieks* – illegally employed worker; *glābējsilīte* – a place where people (typically mothers) can bring babies, usually newborn, and leave them anonymously in a safe place to be found and cared for; *gūglēt* – to google; *hibrīdkarš* – hybrid war; *hipsteris* – hipster; *iedabisks* – inherent; *kaceklis* – teaser video; *kebabnīca* – kebab fast-food restaurant; *koučings* – coaching; *koučsērfings* – couchsurfing; *mākoņdatošana* – cloud computing; *medijpratība* – media literacy; *mēms* – meme; *netiķete* – netiquette; *pakomāts* – parcel locker; *pašiņš* – selfie; *pešmerga* – peshmerga; *pikšķerēšana* – phishing; *planšetdators* – tablet computer; *retrīts* – retreat; *slimstrāde* – presenteeism; *spaiss* – synthetic cannabis; *stulbings* – planking-type phenomenon; *supošana* – stand-up paddleboarding; *trampampiņš* – Trump-like politician, pej.; *trampisms* – Trumpism; *trollēt* – to troll; *tvikšanās* – tweetup; *urla* – member of an aggressive urban subculture; *vebinārs* – webinar; *velorikša* – cycle rickshaw; *vidisks* – environmental; *vīterot* – to tweet; *zaļināšana* – greening; *zaļmaldināšana* – greenwashing; *zibakcija* – flashmob; *zibmaksājums* – instant payment
MDNW 2024 (LV)	*brīvkrāns* – public drinking fountain; *domvedis* – influencer; *ieauši* – in-ear headphones; *iedabisks* – inherent; *iespējkapitāls* – venture capital; *ietekmelis* – influencer, pej.; *kaceklis* – teaser video; *klupne* – error, bug; *kovidiots* – covidiot; *mājsēde* – lockdown; *okupeklis* – Soviet-era monument/mindset, pej.; *orkistāna* – Russia, pej., lit. 'Orkland'; *rašisms* – Russian ultranationalist ideology; *slimstrāde* – presenteeism; *stāvdrāzis* – e-scooter; *sveškauns* – second-hand embarrassment; *tālsaiste* – remote connection; *tankijs* – tankie; radical communism supporter; *tīkšķis* – the 'Like' button; *tviterkonvojs* – Twitter aid convoy; *vatņiks* – fervent follower of Russian propaganda

From the 4th edition onwards, the data extracted for this research comprises only Latvian new words, as Latvian was first included in the MDNW in this edition. Several new words appear in more than one edition; for example, MDNW 2011 includes five Lithuanian words that were already present in MDNW 2010. A total of 11 Lithuanian new words are repeated across editions 1–3. In total, 56 Latvian new words appear in the MDNW editions published from 2014 to 2024. Seven neologisms (see Table 4) from the Latvian neologism dataset remained included in four editions of the dictionary from the 4th edition (MDNW 2013) to the 7th edition (MDNW 2019) — *bezstils* (EN: behaviour or fashion sense that does not comply with society's perception of good taste; chaotic and unsuccessful mixture of different fashion styles (< *bez* + *stils* = without + style) (MDNW 2019, p. 24)), *draugoties* (EN: to be active on the Latvian social network draugiem.lv or on any other social network (< *draugs* = friend) (MDNW 2019, p. 53)), *ēnstrādnieks* (EN: illegally employed worker, one of the elements of the shadow economy (< *ēna* + *strādnieks* = shadow + worker) (MDNW 2019, p. 60)), *glābējsilīte* (EN: a safe place where people can bring their babies, usually newborns, and leave them anonymously to be found and cared for (< *glābējs* + *silīte* = saviour + crib) (MDNW 2019, p. 81)), *mākoņdatošana* (EN: data storage and processing on remote servers accessed via the internet (< *mākonis* + *datošana* = cloud + computing) (MDNW 2019, p. 131)), *urla* (EN: a member of an urban subculture characterised by a specific fashion sense (typically for sports clothing) and by a lifestyle involving aggressive and often criminal behaviour (MDNW 2019, p. 225)) and *zibakcija* (EN: flashmob; a group of people who assemble suddenly in a public place, perform an unusual act for a brief time, and quickly disperse, typically for purposes of entertainment, satire, or artistic expression (< *zibēt* + *akcija* = to flash + event) (MDNW 2019, p. 241)). All seven items were first attested in MDNW 2013 (4th edition) and continued to appear for six years, through MDNW 2019 (7th edition).

TABLE 4 Latvian neologisms appearing in all four MDNW editions (2013, 2014, 2016, and 2019)

Neologism (part of speech)	English equivalent	Word-formation
bezstils (n., m.)	behaviour or fashion sense that does not comply with society's perception of good taste; chaotic and unsuccessful mixture of different fashion styles	< bez + stils = without + style
draugoties (v.)	to be active on the Latvian social network draugiem.lv or on any other social network	< draugs = friend
ēnstrādnieks (n., m.)	illegally employed worker, one of the elements of the shadow economy	< ēna + strādnieks = shadow + worker
glābējsilīte (n., f.)	a safe place where people can bring their babies, usually newborns, and leave them anonymously to be found and cared for	< glābējs + silīte = saviour + crib
mākoņdatošana (n., f.)	data storage and processing on remote servers accessed via the internet	< mākonis + datošana = cloud + computing
urla (n., m.)	a member of an urban subculture characterised by a specific fashion sense (typically for sports clothing) and by a lifestyle involving aggressive and often criminal behaviour	— (no compound structure given in the dictionary)
zibakcija (n., f.)	flashmob; a group of people who assemble suddenly in a public place, perform an unusual act for a brief time, and quickly disperse, typically for purposes of entertainment, satire, or artistic expression	< zibēt + akcija = to flash + event

All seven items were first included in MDNW 2013 (4th edition) and remained in the index through MDNW 2019 (7th edition). None of these seven items — nor any of the other 36 Latvian words listed in MDNW 2019 — reappear in the MDNW 2024 (8th edition) index, which instead lists only 18 entries newly introduced in that edition. This indicates that the dictionary's cumulative editorial model, stable from 2013 to 2019, was discontinued in the 2024 edition in favour of a non-cumulative model in which only the most recently attested neologisms are included. This raises the question of how lexicographers select words for specific editions of a dictionary, with particular emphasis on the novelty of these words. The definition of the new word in the dictionary itself should also be clarified and defined for dictionary users. This finding is lexicographically significant for two reasons: first, it suggests that the MDNW functions not as a strictly synchronic "snapshot" dictionary of new words but as a cumulative, ever-growing repository lacking a clear mechanism for retiring or removing entries once a word is no longer new; second, it reveals a tension between the dictionary's self-designation as a "dictionary of new words" and its actual editorial practice, pointing to the need for a clearer definition of the term "new word" for dictionary users.

5 Conclusions

The MDNW grew from 70 pages in its inaugural 2010 edition to 272 pages in 2014, before fluctuating in length across subsequent editions to reach 243 pages in 2024, reflecting an overall near-fourfold expansion across the series despite periodic reductions in page count. Latvian was absent from the first three editions (2010–2012) despite Latvia's status as an EU member state, being included only in the fourth edition (2013) alongside Czech and Irish. Although the

fact that Latvia is a member of the EU does not necessarily mean that this dictionary must include Latvian words, it is nevertheless important for dictionary users to be aware of this. Following its inclusion, the number of Latvian entries grew steadily from 20 (2013) to a peak of 43 (2019), before declining sharply to 21 in the eighth edition (2024), suggesting possible institutional or editorial changes within the Latvian translation unit. Moreover, the cumulative entry counts overstate actual lexical growth, as the majority of entries in each edition are carried over from previous editions rather than newly added neologisms. Megastructural analysis reveals that all eight editions consistently lack a user's guide — the only outer-text component regarded as obligatory in lexicographic theory — while functional access structures such as general and language-sorted indices appear only from the sixth edition (2016) onwards, revealing a persistent gap between institutional practice and theoretical standards. Starting with the 2016 edition, the dictionary displays a complete redesign of its microstructure, introducing colour coding, phonetic transcription, and indices, while the year-of-first-inclusion markers were omitted. The microstructural evolution across the eight editions traces a trajectory from rigid typographic layouts with isolated formal and semantic cores towards fully integrated, digitally optimised entries enriched with phonetic transcription, morphological commentary, pragmatic markers, and embedded contextual examples, with the most substantial redesign occurring in the sixth edition.

Of all the extracted Latvian neologisms, seven lexical units—*bezstils* ('lack of fashion sense'), *draugoties* ('to be active on the Latvian social network draugiem.lv or on any other social network'), *ēnstrādnieks* ('illegally employed worker, one of the elements of the shadow economy'), *glābējsilīte* ('a safe place where people can bring their babies, usually newborns, and leave them anonymously to be found and cared for'), *mākoņdatošana* ('data storage and processing on remote servers accessed'), *urla* ('a member of an urban subculture characterised by a specific fashion sense (typically for sports clothing) and by a lifestyle involving aggressive and often criminal behaviour'), and *zibakcija* ('flashmob; a group of people who assemble suddenly in a public place, perform an unusual act for a brief time, and quickly disperse, typically for purposes of entertainment, satire, or artistic expression')—demonstrated cross-edition stability by remaining included across four editions from the fourth (2013) through the seventh (2019), indicating that their status as new words remained unchanged across four consecutive editions spanning six years. Taken in isolation, this pattern might suggest that "new word" was being used institutionally rather than temporally: once admitted, a neologism appears to have retained its place in the dictionary regardless of whether it still met a genuine synchronic criterion of novelty, since a word attested continuously since 2013 can hardly be considered "new" by 2019 in any linguistic sense. This reading is reinforced by the fact that none of the seven items shows an updated date label, a register change, or renewed justification across the four cumulative editions — their retention appears to be a matter of editorial inertia rather than a reassessed judgement of ongoing novelty. However, the 2024 edition complicates this interpretation. Rather than continuing to expand the cumulative list, MDNW 2024 discontinued it altogether, replacing it with a non-cumulative index of only 18 entries newly introduced in that period; none of the words retained since 2013 — including the seven discussed above — reappear in either the index or the body text of the 2024 edition. This discontinuity suggests that long-term inclusion did not, in the end, reflect a stable institutional redefinition of "new word" as a permanent lexicographic category. This reading is supported by the dictionary's own forewords, which consistently present the MDNW as an additive, ever-growing project—the 8th edition even celebrates a "quantitative leap" of 512 new entries. Yet no foreword ever mentions removing or reassessing previously included words, nor defines what makes a word "new" beyond the year of its first inclusion. Notably, the 2024 foreword says nothing about the disappearance of the 43 Latvian entries carried since 2019. This silence suggests that the shift documented here was not a deliberate, announced policy but an unremarked practice discontinuity — the

forewords narrate cumulative growth while the 2024 macrostructure quietly contradicts it, underscoring the need for the MDNW to define the category "new word" explicitly for its users.

References

Atkins, B. T. S. ([1992/1993] 2008). Theoretical Lexicography and Its Relation to Dictionary-making. In T. Fontenelle (Ed.), *Practical Lexicography. A Reader* (pp. 31–50). Oxford University Press.

Baldunčiks, J. (2012). *Vārdnīcu izstrāde Latvijā* (1991–2010). Pētījums J. Baldunčika vadībā. Rīga: Latviešu valodas aģentūra.

Baldunčiks, J. (2024 [2016]). Jaunvārdu reģistrēšanas problemātika "Mīlenbaha-Endzelīna vārdnīcā". In Baldunčiks, J. *Zinātniskie raksti. Viedokļraksti. Intervijas.* (pp. 259–275). Zinātne.

Bergenholtz, H. J. (2001). Lexikografie ist Selektion ist Selektion ist Selektion. In J. Korhonen (Hrsg.) *Von der mono- zur bilingualen Lexikografie für das Deutsche*. (pp. 11–30). Peter Lang.

Cluver, A. D. de V. (1989). *Dictionary of Terminography. Terminografie-woordeboek*. Human Sciences Research Council.

Hartmann, R. R. K., & James, G. (1998). *Dictionary of Lexicography*. Routledge.

Hartmann, R. R. K. (2001). *Teaching and Researching Lexicography*. Pearson Education Limited.

Hausmann F. J. & Wiegand H. E. ([1989] 2003). Component Parts and Structures of General Monolingual Dictionaries: A survey. In R. R. K. Hartmann (Ed.), *Lexicography. Critical Concepts. Vol. 3. Lexicography: Lexicography, Metalexicography and Reference Science*. (pp. 207–254). Routledge.

Karpinska, L. (2012). *Critical analysis of English-Latvian lexicographic tradition* [Doctoral dissertation, University of Latvia]. University of Latvia.

Kern, B. (2024). Considering Word Formation in Compiling Dictionaries. In K. Š. Despot, A. Ostroški Anić & I. Brač, (Eds.). (2024). *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress*. (pp. 463–472). Institute for the Croatian Language.

Laua, A. (1969). *Latviešu leksikoloģija*. Zvaigzne.

MDNW (2010). *Multilingual Dictionary of New Words*. European Parliament Directorate-General for Translation.

MDNW (2011). *Multilingual Dictionary of New Words*. Second Edition. European Parliament Directorate-General for Translation.

MDNW (2012). *Multilingual Dictionary of New Words*. Third Edition. European Parliament Directorate-General for Translation.

MDNW (2013). *Multilingual Dictionary of New Words*. Fourth Edition. European Parliament Directorate-General for Translation.

MDNW (2014). *Multilingual Dictionary of New Words*. Fifth Edition. European Parliament Directorate-General for Translation.

MDNW (2016). *Multilingual Dictionary of New Words*. Sixth Edition. European Parliament Directorate-General for Translation. Publications Office of the European Union.

MDNW (2019). *Multilingual Dictionary of New Words*. Seventh Edition. European Parliament Directorate-General for Translation. Publications Office of the European Union.

MDNW (2024). *Multilingual Dictionary of New Words*. Eighth Edition. European Parliament Directorate-General for Translation. European Union.

Ozola, E. & Helviga, A. (2020). Jaunvārdu pētniecības problemātika mūsdienu leksikoloģijā un leksikogrāfijā: Latvijas un Lietuvas pieredze. In *Res humanitariae*. (pp. 290–306). Klaipeda University Press.

Peina, E. (2023). *Mūsdienu latviešu terminogrāfijas teorētiska un pragmatiska analīze /Theoretical and Pragmatic Analysis of Modern Latvian Terminography*. [Doctoral dissertation, Liepāja University]. Liepāja University.

Steffens, D. (2007). Von "Aquajogging" bis "Zickenalarm". Neuer Wortschatz im Deutschen seit den 90er Jahren im Spiegel des ersten größeren Neologismuswörterbuches. In *Der Sprachdienst* 51 (4). (pp. 146–159).

Storjohann, P. & Cimander, L. (2022). Annäherung an den Coronadiskurs in der öffentlichen Kommunikation mithilfe von Neologismen./Approaching the Covid-19 Discourse Through Neologisms in Public Communication. In M. Jakosz & M. Kalasznik (Hg.). Corona-Pandemie: Diverse Zugänge zu einem aktuellen Superdiskurs. 1. Auflage. (pp. 25–50). V & R unipress | Brill Deutschland GmbH.

Svensén, B. (2009). *A handbook of lexicography: The theory and practice of dictionary-making*. Cambridge University Press.

Sviķe, S. (2016). *Speciālā leksika vispārīgajās divvalodu tulkojošajās vārdnīcās: augu nosaukumi/Special lexis in general bilingual dictionaries: plant names*. [Doctoral dissertation, Ventspils University of Applied Sciences]. Ventspils University of Applied Sciences.

Veisbergs, A., Brēde, M., Ločmele, G., Šalme, A. & Žīgure, V. (2009). Jaundarinājumi latviešu valodā 21. gadsimtā. In *Letonikas trešā kongresa zinātniskie raksti*. (pp. 211–223). Latvijas Zinātņu akadēmija.

Zgusta, L. (1971). *Manual of Lexicography*. Academia. Publishing House of the Czechoslovak Academy of Sciences.

Contact information

Silga Sviķe
Ventspils University of Applied Sciences
silga.svike@gmail.com

Carole Tiberius*, Martin Kroon*, Lut Colman & Dalí Chirino

RANKING COMBINATIONS IN *WOORDCOMBINATIES*

Abstract This paper describes efforts to extract frequency information from corpora and to rank the collocates in *Woordcombinaties*, a Dutch phraseological lexicographic resource, in a uniform and systematic way, taking various factors such as the size and composition of the corpus, linguistic processing and the mathematical properties of various association measures into account. A small sample of 2032 combinations, divided over 19 verbs and 19 nouns, was used, and their frequencies were extracted from the Corpus Contemporary Dutch, using surface co-occurrence, textual co-occurrence, and syntactic co-occurrence. Results were verified manually. It was found that: 1) corpus size increases coverage; 2) syntactic co-occurrence leads to loss of coverage; 3) combination of a narrow window span and within-sentence context seems to have the highest coverage; 4) separate treatment is needed to deal with multi-word collocates. However, given that surface/textual co-occurrence still produces a substantial amount of false positives, we intend to use syntactic co-occurrence to retrieve the frequency information for the combinations in *Woordcombinaties*, because it has a higher precision and offers the greatest scope for improvement.

Keywords Collocation measures; Dutch; lexicographic resource

1 Introduction

Woordcombinaties (Word Combinations) is a project of the Dutch Language Institute developing an integrated lexicographic resource combining a collocation and idiom dictionary with a pattern dictionary (Colman & Tiberius, 2018). It merges a word-in-context and collocation tool, following the example of SKELL (Sketch Engine for language learning; Kilgarriff et al., 2004; 2014), and a pattern dictionary, following examples such as the Pattern Dictionary of English Verbs (PDEV) and T-PAS for Italian (Ježek et al., 2014). The term 'word combinations' should thus be understood in a broad sense covering all types of multiword expressions, ranging from collocations to idioms, conversational routines as well as patterns (identified following the technique of Corpus Pattern Analysis developed by Hanks (2013)). The project focuses on combinations for nouns and verbs, and offers an online interface (as illustrated in Figure 1) through which the different types of combinations can be easily accessed.

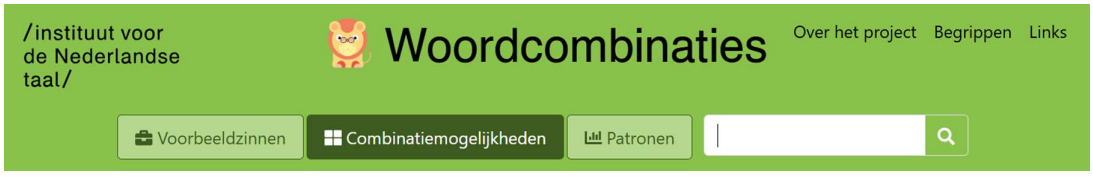


FIGURE 1 Part of the online interface of *Woordcombinaties* with the three search options

Going from left to right, the search options are arranged according to their increasing degree of complexity, ranging from a simple search of a word in example sentences (*voorbeeldzinnen*), over combinations (*combinatiemogelijkheden*), to pattern-meaning pairs (*patronen*). Two further search options are foreseen for idioms and conversational routines such that it will, for example, be possible to search for conversational routines linked to a specific speech act, such as 'greeting' or 'apologizing'. Currently idioms and conversational routines are encoded as special instances among the collocations and the patterns.

* Equal contribution

In this paper, we focus specifically on the combinations (*combinatiemogelijkheden*) in the second tab as we would like to find out what is the best way to rank the collocates from a lexicographic perspective. Collocates are retrieved from a subcorpus of approx. 200 million tokens of the Corpus Contemporary Dutch (CHN) using a customised sketch grammar in Sketch Engine. The grammatical relations in the sketch grammar closely match the relations shown in the online application. The collocates are manually validated before inclusion in *Woordcombinaties*.

At the time of writing, *Woordcombinaties* contains more than 100.000 combinations in 42 different grammatical relations for approx. 1500 entries. New entries and new combinations are added on a daily basis. Currently, the collocates are shown in alphabetical order in the online application as can be seen in Figure 2.

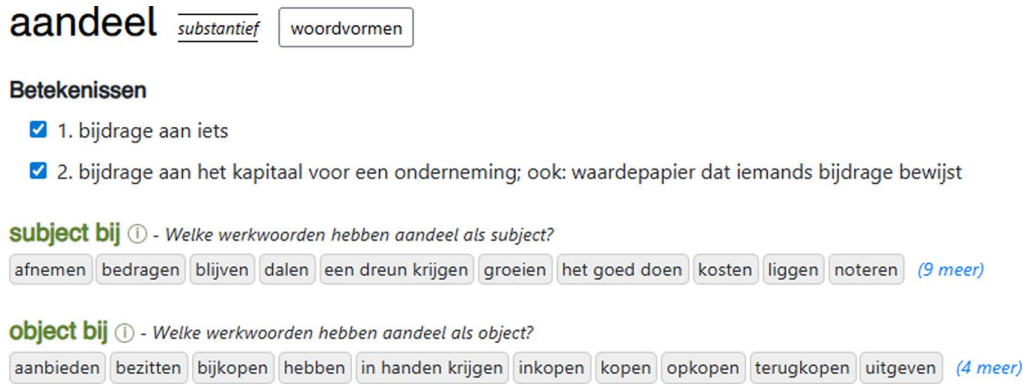


FIGURE 2 Extract of combinations with aandeel ('share') showing collocations in alphabetical order

However, to do justice to the fact that these collocates are included because they form meaningful, typically recurring combinations with the headword, we would like to add an option to rank them based on a combination of their frequency of co-occurrence and strength of association.

In theory, frequency information and salience scores can be retrieved from Sketch Engine, but for various reasons we do not want to use the numbers from the word sketches without further consideration. First of all, errors in the automatic annotation affect the numbers. For example, *pannenkoeken* ('pancakes') occurs both in the subject and object relation in the word sketch for the verb *bakken* ('to bake'), but should of course only be counted as an object (ideally adding the instances where it is incorrectly put in the subject relation to the total object count). Secondly, sets of semantically similar collocates tend to be grouped together in *Woordcombinaties*, e.g. *er [niets, niks, niet veel, weinig, amper iets] van bakken* (lit. 'to bake [nothing, not a lot, little, barely something] of it'; 'to make a complete mess of it') and it is not entirely clear what frequency information and salience score can be associated with the whole set. The same applies to combinations that contain optional elements such as the article in *marcheren in (het) gelid* ('to march in formation').

The previous two examples also highlight another issue, i.e. the fact that in *Woordcombinaties* both collocates and headwords can consist of more than one word. This deviates from the 'default' definition of a collocation as a combination of two components (e.g. Evert, 2009) raising additional difficulties for determining their frequency and salience[1], and leads to the following research question: *What is the best method for extracting frequency information and measuring association strength for ranking the collocates in Woordcombinaties in a uniform and*

[1] This is also the reason why we prefer to use the term *combination*, rather than the more strict *collocation*.

systematic way, taking factors such as the size and composition of the corpus, linguistic processing and the mathematical properties of various association measures into account?

The remainder of the paper is structured as follows. Section 2 summarises related work. Section 3 describes our experiment, while Section 4 presents the results. Section 5 concludes the paper.

2 Related work

Collocations constitute one of the most heavily researched areas in corpus linguistics, computational linguistics and lexicography. However, most of the work has focused on English and large-scale studies dealing with collocations in Dutch are, to the best of our knowledge, still missing.

Furthermore, more recent studies note that our understanding of the properties of association measures (which play an important role in traditional approaches dealing with collocations) and of other parameters such as the size of the co-occurrence context is still incomplete (Evert et al., 2017, p. 531) and that we should not underestimate the complexity and multidimensionality of the notion of co-occurrence (Gries & Durrant, 2020, p. 156).

To this end, Evert et al. (2017) present a large-scale evaluation study of association measures for collocation identification in English taking into account size and composition of source corpora, size of co-occurrence contexts and frequency thresholds. Their results show that a) small co-occurrence spans are better than larger spans and that syntactic co-occurrence is almost always the best choice; b) carefully sampled corpora outperform comparably-sized messy web corpora, but sufficiently large web corpora prove to be a valid substitute for the British National Corpus; c) frequency thresholds seem to have little effect in most situations; d) there is little interaction between the choice of association measure and the other parameters.

Building on this, Uhrig et al. (2018) focus specifically on using syntactic co-occurrence for collocation candidate selection. They find that the performance of parsers and the precision of association measures vary considerably between relations, but that log-likelihood is the best association measure considering all six grammatical relations in their study.

Ljubešić et al. (2021) take a slightly different approach and concentrate on the task of collocation ranking for Slovene lexicography. They find that lexicographers consider the listings of word embeddings similar to the rankings based on logDice (a popular measure in lexicography). They note that not only the composition of the corpus, but also the monosemy or polysemy of selected headwords seem to influence the results.

Köhler et al. (2025) address the question of which type of corpora should form the basis of lexicographic work for compiling (collocation) dictionaries for German. Their findings support the conclusions of Evert et al. (2017) that well-curated corpora perform better than messy web corpora of equivalent size. They specifically note that well-curated newspaper text corpora outperform web corpora of the same size. This is important as lexicographic resources (including *Woordcombinaties*) are often based on newspaper data.

Unlike most of the related work, our focus is not on identifying collocations, but on determining the best method for associating frequency information with the already validated combinations in *Woordcombinaties* and to rank them. From the parameters mentioned in the literature, we took corpus size and composition, size and type of co-occurrence context, and a range of association measures into account in our experiment.

3 Set up of the Case Study

3.1 Test data

For our case study we selected 19 nouns and 19 verbs semi-randomly from *Woordcombinaties* together with their combinations for 18 grammatical relations (see Appendix A), leaving out some low-frequent relations and very complex relations such as "as part of sayings, proverbs etc.". The resulting test set contained 1146 combinations for the noun entries and 886 combinations for the verb entries (see Table 1). As *Woordcombinaties* describes the combinations of high-frequency words, most of the words in our test set are polysemous.

TABLE 1 Overview of the lemmas in our test set, along with the number of combinations that are recorded within *Woordcombinaties*

VERBS	Translation	No. of combinations	NOUNS	Translation	No. of combinations
aanbieden	to offer	125	aanbieding	offer	63
aanbieden, zich	to present oneself	3	aanblik	sight	24
aanbrengen	to apply	58	aandeel	share	59
aandoen	to put on	71	aandeelhouder	shareholder	90
aandringen	to urge	87	aangifte	declaration	60
aanduiden	to indicate	68	aanhanger	trailer	55
aanmoedigen	to encourage	55	Amerikaan	American	45
lukken	to succeed	70	land	country	185
lusten	to fancy	27	luchthaven	airport	53
manipuleren	to manipulate	51	lunch	lunch	50
mankeren	to lack	31	man	man	171
marcheren	to march	24	zonde	sin	34
zwemmen	to swim	57	zone	zone	49
zweren[1]	to swear	29	zoon	son	42
zweren[2]	to fester	3	zorg	care	54
zwerven	to wander	42	zuiden	south	24
zweten	to sweat	21	zus	sister	24
zweven	to float	36	zweet	sweat	31
zwijgen	to be silent	28	zwembad	swimming pool	33

3.2 The corpus

To extract frequency information for the combinations in our test set, we used the CHN and various subsets of this corpus as our test bed. The CHN is a monitor corpus of 4.46 billion tokens (as of February 2026) containing material from the Netherlands, Flanders, Surinam and the Dutch Caribbean. It consists mainly of newspapers, but also contains material from the web as well as fiction. The corpus is lemmatised and tagged, and parts of the corpus have been syntactically parsed. Parsing the whole corpus is planned for 2026. The different test corpora used in our experiment are listed in Table 2.

TABLE 2 Corpora used for the study

(Sub)corpus	Tokens (excl. punctuation)	Annotation
CHN-i	4,487,983,812	tagged
TermWerk	594,851,621	parsed (spaCy)
CWParsed	178,285,778	parsed (Alpino)

CHN-i is the institute-internal version of the CHN and the largest dataset of contemporary Dutch available at the Dutch Language Institute. It is not only much larger than the other two, but also more varied in terms of content, containing books, blogs, magazines and spoken data in addition to newspaper material. However, the two smaller corpora benefit from an additional annotation layer as they have been parsed. The TermWerk subcorpus was created in 2025 and has been parsed with spaCy (Honnibal et al., 2020) using the `nl_core_news_lg` model, whereas the CWParsed subcorpus was created in 2019 specifically for the project *Woordcombinaties* and has been parsed with the Alpino parser (van Noord, 2006). The three test corpora differ in size, composition, and annotation, enabling us to take these factors into account in our analysis.

3.3 Type of co-occurrence

Following the literature (cf. Evert, 2009), we considered surface co-occurrence (words co-occurring within a certain distance of each other), textual co-occurrence (words co-occuring in the same sentence), and syntactic co-occurrence (words co-occurring in a certain grammatical relation to each other), as well as combinations of those. After initial exploration and based on findings in the literature (e.g. Evert et al., 2017), surface and textual co-occurrence were combined and co-occurrence was computed for all three test corpora within sentence context with a span of three words to the left and three to the right. Using textual co-occurrence alone setting the span to within-sentence gave too much noise, especially with longer sentences. Syntactic co-occurrence was only calculated for the two parsed corpora using the grammatical relations mentioned in the appendix.

Based on these parameters, corpus queries were automatically generated for the 2032 combinations in our test set. The resulting queries were then run on the test corpora via the API to see how many hits were found. For example, for the surface/textual co-occurrence for the combination *zwemmen* ('to swim') with *leerling* ('pupil') as subject, the following query is generated[2]:

```
(([lemma="leerling"] | [word="leerling"]) []{0,2}
  [lemma="zwemmen"] | [lemma="zwemmen"] []{0,2}
  ([lemma="leerling"] | [word="leerling"])) within <s/>
```

The query searches for both the collocate as a lemma and a word to cover instances where the lemma form deviates from the collocate as listed in *Woordcombinaties* (e.g. *divers* ('various') in *diverse aanbiedingen* ('various offers')). For the same reason the CQL query does not include part-of-speech information as the correct part-of-speech may differ from the annotation in the corpus.

[2] Note that the information on the grammatical relation being subject is not integrated into the query shown: this information is not present in a corpus that is only tagged and not parsed. It is of course used in the query generation for the syntactic co-occurrence:

```
[lemma="zwemmen"] -nsubj-> [pos="NOUN|PROPN" & (lemma="leerling" |
word="leerling")]
```

As noted above, both headwords and collocates can consist of more than one word in *Woordcombinaties* which complicates the automatic query generation. For instance, in the case of the combination *zweven tussen hemel en aarde* ('to be left in suspense'), we consider *tussen hemel en aarde* as the collocate and allow up to 2 tokens in between the verb and the start of the collocate (or 2 tokens between the end of the collocate and the verb if the expression is inverted). This is still doable, as the collocate forms one unit and does not show any variation, but it gets more difficult with combinations containing optional and/or open slots as in the combination *zwerven van ... tot ...* ('wander from ... to ...') or in the combination *zweren erbij (dat)* ('swear by (that)'). Challenging are also combinations where the collocate contains a verb that can occur in all its inflected forms in the corpus, e.g. *aanbieding* ('offer') in combination with *geldig zijn* ('to be valid'). For those cases more advanced CQL queries with the `containing` operator are used, e.g.:

```
[lemma="geldig"] within <s/> containing [lemma="zijn"]
  within <s/> containing [lemma="aanbieding"]
```

These advanced queries have to be manually verified as they can become quite complex and are prone to errors[3].

3.4 Association measures

Mere co-occurrence is not sufficient for distinguishing collocations. We also need to consider the strength of attraction between the components. To calculate how strongly words are attracted to each other, a large number of association measures have been suggested and used over the past decades. For our study, we compared a set of 9 association measures: t-score (Church et al., 1991), (pointwise) mutual information (MI; Church & Hanks, 1990), MI^3 (Daille, 1994), MI.log-f (Kilgariff et al., 2004), log-likelihood (Dunning, 1993), logDice (Rychlý, 2008), minimum sensitivity (Pedersen & Bruce, 1996), and Delta P (ΔP; Ellis, 2007) in both directions. All measures use the frequency of word A, the frequency of word B and the frequency of words A and B co-occurring, some also use the size of the corpus.

An important point to note is that all these measures were in principle designed to work for two input words (or events), while we have multi-word combinations in the *Woordcombinaties* dataset, as well. Adapting the 9 measures we used to multivariate functions, however, is mathematically complex (cf. also Van de Cruys, 2011), and deserves its own research project. It is furthermore unclear whether two-way association measures can be directly compared to, say, three-way association measures: for instance, whereas a two-way mutual information (not pointwise!) is by definition non-negative, a three-way mutual information is not and can be negative. For the purpose of this study, we therefore considered multi-word headwords and collocates as single events with one frequency (i.e., a collocate such as *in plaats van* ('instead of') was not treated as three distinct words with their own frequencies, but instead as one), so that the existing bivariate association-measure functions could be applied to our data. We realise this may have statistical implications on our output, but we considered it not in the scope of this study to address those.

Another point to note is that the measures were in principle designed to work for surface co-occurrence (with some adaptations suggested for syntactic co-occurrence). In theory, we expect all measures to work with any kind of co-occurrence, because there is nothing in

[3] Separable verbs do not pose a problem as they are lemmatised as the whole verb even when the verb is split in the text, e.g. *hij **doet** zijn jas **aan*** (lit. 'he put his coat on') is tokenised and lemmatised as hij|aandoen|zijn|jas|aan.

the functions themselves that dictates that the co-occurrence must be linear (or indeed needs to concern occurrences of words), and applying the association measures to syntactic (or relational) co-occurrences has been done by others without issue (cf., e.g., Uhrig et al., 2018). Importantly, though, when considering syntactic co-occurrence, we did not consider the frequencies of the individual syntactic relations, as that would make the input trivariate (two words and one relation). The frequency of a word, therefore, was taken as the number of occurrences of a word in a specific syntactic relation, either as head or as dependent.

4 Analysis and results

With all parameters set, we conducted our case study. For surface/textual co-occurrence, we compared the results from CHN-i and CWParsed, while for syntactic co-occurrence we used the two parsed corpora, TermWerk and CWParsed.

4.1 *Surface/textual co-occurrence results*

In terms of coverage, it is clear that searching within the much larger CHN-i corpus misses (i.e. fails to find) much fewer combinations than searching within the CWParsed corpus, respectively 9 in the CHN-i and 77 in the CWParsed corpus. This is to be expected, simply because combinations have a higher chance of occurring in the CHN-i due to its size. Looking more closely at the results, we see that most of the 77 combinations not being found in the CWParsed corpus were combinations with multi-word collocates with automatically generated queries involving the `containing` operator. Manual verification showed that most of these combinations were present in the corpus, but the query did not return any results because of some slight differences in the implementation of CQL between Sketch Engine and the corpus query system used for CHN-i, BlackLab (de Does et al., 2017). If we disregard the combinations involving multi-word collocates, the numbers reduce to 8 noun and 8 verb combinations not being found in the CWParsed corpus and 1 noun and 1 verb combination missing in CHN-I (cf. Table 3).

TABLE 3 Number of combinations that were found and missed using the surface/textual co-occurrence, excluding not-found multi-word collocates

	NOUNS		VERBS	
	found	missed	found	missed
CHN-i	1140	1	883	1
CWParsed	1090	8	865	8

We also investigated whether the results from both corpora yielded similar rankings based on the association measures. We compared the ranked lists per grammatical relation for all words in our test set with Rank-Biased Overlap (RBO; Webber et al., 2010). RBO compares the overlap of two ranked lists at incrementally increasing depths, and quantifies the expected average overlap assuming the lists are of infinite length, describing a similarity between two rankings. It takes one parameter, p, which gives the user control of how heavy higher-ranking elements in the lists count towards the similarity, compared to lower-ranking elements. We set p to 0.9, which means that the overlap in the top-10 of the two ranked lists account for about 85% of the similarity measure. The measure is useful for our purposes, because the highest variation is expected at the bottom of the ranked lists, but RBO and human users focus mainly on the top.

What is seen is that all 9 measures are relatively stable between the two corpora, with an RBO between 0.81 and 0.86. LogDice seemed to vary most between the corpora, while ΔP (direction node-to-collocate) seems to be most stable. This is interesting, because it means that the collocation rankings produced by different association measures are only weakly dependent on the size of the corpus. It must be stressed, however, that their stability between corpora does not necessarily mean that ΔP is the most useful association measure.

4.2 Syntactic co-occurrence results

We also considered syntactic co-occurrence in our case study because, according to the literature (e.g. Evert et al., 2017), this is almost always the best choice provided that parsed corpora are available. In our case, however, it is difficult to fairly compare the two parsed corpora that we have access to, Termwerk and CWParsed, because, as already noted above, they use different interfaces that support slightly different implementations of CQL and can therefore not be queried with the exact same queries. The differences are even more pronounced at the syntactic level. Termwerk is available in BlackLab (de Does et al., 2017) and can be queried using BCQL (BlackLab Corpus Query Language), which supports relations querying[4]. CWParsed is uploaded in Sketch Engine, which does not provide an equivalent for relations querying. We nevertheless present some results.

TABLE 4 Number of combinations that were found and missed using the syntactic co-occurrence

	NOUNS		VERBS	
	found	missed	found	missed
TermWerk	1058	88	785	101
CWParsed	1010	136	606	280

In terms of coverage, it was again observed that searching in the larger corpus, Termwerk, yielded more hits, though the effect of corpus size is much smaller compared to the smaller corpus, CWParsed. This makes sense, because the size discrepancy is also much smaller: Termwerk is about 3 times larger than CWParsed, while CHN-i is about 25 times larger. Cf. also Table 4.

Inspecting the results, we see that the missing numbers for verbs in CWParsed are mainly due to combinations involving a trinary relation[5] in Sketch Engine that have not been properly matched with the data in *Woordcombinaties* (175 of the 280 verb combinations that were not found).

The number of combinations for which no hits were found in the corpora is much higher than for the surface co-occurrence method. This is mainly due to two complications: first, the automatically generated queries for combinations with multi-word collocates were often incorrect and yielded no results; second, more obviously, the parses of the sentences contained many errors, in particular for the Termwerk corpus, leading to a mismatch between the query and the annotation.

[4] See https://blacklab.ivdnt.org/guide/query-language/

[5] As defined in Sketch Engine: https://www.sketchengine.eu/documentation/writing-sketch-grammar/

No comparison was done between the corpora for the association measures, because we deemed it less meaningful to compare lists generated by different querying methods and annotations.

4.3 Discussion of results

Overall, our quantitative results are in line with the findings in the literature described in Section 2. We see that 1) corpus size increases coverage; 2) syntactic co-occurrence leads to loss of coverage; 3) combination of a narrow window span and within-sentence context seems to have the highest coverage; 4) separate treatment is needed to deal with multi-word collocates.

To get an insight into the quality of the output, a small manual evaluation was carried out. Using the automatically generated queries for surface/textual co-occurrence, concordance lines were extracted from CHN-i which were then manually annotated by one lexicographer as being an correct/incorrect instance of a combination. Approximately a quarter of the matching concordance lines per headword were judged incorrect by the lexicographer, e.g. because the headword verb was part of a larger compound (*zwemt mee* ('swims along'), in which *zwemt* is part of the separable verb *meezwemmen*), the headword was used as a noun instead of a verb (*het dagelijks zwemmen* ('the daily swimming')) or for another reason (such as *zwemmend* ('swimming'), used as a present participle). There were also instances where the lexicographer was unsure whether the concordance line can be considered a correct instance of a combination. For instance, is *leerlingen* ('pupils') the subject of the verb *zwemmen* ('to swim') in the sentence *hij leert de leerlingen zwemmen* ('he teaches the pupils to swim')? Those cases were not counted as incorrect.

The same was done for queries generated for syntactic co-occurrence using the TermWerk corpus. About 1 in 9 of the matching concordance lines per headword were judged incorrect by the lexicographer, largely for the same reasons as the incorrect matches for the surface/textual co-occurrence. There were again cases where the lexicographer was unsure whether the concordance line can be considered a correct instance of a combination, which were also not counted as incorrect.

The choice between the two methods, then, seems to be a trade-off between precision and coverage. As we are starting from existing combinations, it is important that the extracted frequency information is as precise as possible. With syntactic co-occurrence we have more control over the output and, in our view, it offers the greatest scope for improvement by refining query results and improving parsing techniques. We therefore intend to use syntactic relations to retrieve the frequency information for the combinations in *Woordcombinaties*.

In order to assess which association measure produces the best ranked lists of combinations, a second manual evaluation was carried out based on data from the TermWerk corpus. Four (computational) lexicographers were presented with ten rankings per syntactic relation (nine association measure + simple frequency) and were asked to rank the ranked lists on their usefulness (limited to a top-three), corresponding to the expectations and wishes of the lexicographers. Based on a small sample of 11 grammatical relations for three verbs and 6 grammatical relations for three nouns, it was chiefly observed that all measures generally yield good rankings. It was also observed that MI and ΔP (direction collocate-to-node) always yielded the same rankings, and that, for verbs, they clearly yielded the best rankings for subjects, indirect objects, and postmodifying prepositional phrase adjuncts without secondary collocates. Furthermore for verbs, the MI.log-f measure seemed to yield the best rankings of direct objects.

Other syntactic relations had no clear trend with regards to preferred association measures. As for nouns, MI and ΔP (direction collocate-to-node) also seemed to yield the best rankings for subjects (where they yielded the same rankings as minimum sensitivity and logDice) and pronoun determiners. Other syntactic relations showed no clear trend.

The fact that all measures generally yield good rankings is not surprising: the measures are not used to extract combinations from a corpus. Instead, they are only used to sort combinations from an already curated list on their salience. This means that any undesirable combination candidates that may show up due to the mathematical nature of the measures are already filtered out. This is most striking with MI, which is famous for extracting low-frequent and undesirable candidates (and therefore not used much anymore), while it shows good results in ranking pre-curated list of combinations.

5 Conclusion

The paper presented a systematic study of ranking Dutch word combinations from a lexicographic perspective using frequency information and association strength. Various factors such as the size and composition of the corpus, linguistic processing and the mathematical properties of various association measures were taken into account.

The results show that syntactic dependencies give the highest precision at the expense of some loss in coverage. As we expect that coverage will increase with the implementation of a new parser for the CHN (foreseen by the end of 2026), we plan to derive frequency information for the combinations in *Woordcombinaties* from the corpus using syntactic dependencies.

So far, our analysis has focused primarily on quantitative measures, with less attention given to qualitative evaluation. In future work, we plan to conduct a larger-scale manual assessment of the results produced by the different association measures. We also intend to refine the automatic query generation process to increase both precision and recall. The refined treatment of multi-word collocates may be based on MWE-Finder (Odijk et al., 2024), a tool for searching for multiword expressions in large Dutch text corpora.

The current work has established the basis for extending the online application of *Woordcombinaties* with a new feature that enables users to rank combinations based on frequency and association score. It has also laid the foundations for the development of a user-friendly collocation functionality within the BlackLab corpus interface (de Does et al., 2017), as well as for compiling a Dutch collocations database on the basis of the CHN (similar to the work described in Köhler et al. (2025) for German). It additionally provides us with a solid basis for investigating the identification and extraction of collocations using machine learning techniques and LLMs in the future.

References

Church, K. & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Computational Linguistics*, 16(1), 22–29.

Church, K., Gale, W. A., Hanks, P. & Hindle, D. (1991). Using statistics in lexical analysis. In U. Zernik (Ed.), *Lexical Acquisition: Using On-line Resources to Build a Lexicon* (pp. 115–164).

Colman, L. & Tiberius, C. (2018). A Good Match: a Dutch Collocation, Idiom and Pattern Dictionary Combined. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (Eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts* (pp. 233–246). Ljubljana University Press, Faculty of Arts.

Daille, Béatrice (1994). *Approche mixte pour l'extraction automatique de terminologie : statistiques lexicales et filtres linguistiques* [Doctoral dissertation, Université Paris 7].

de Does, J., Niestadt, J. & Depuydt, K. (2017). Creating research environments with BlackLab. In J. Odijk and A. van Hessen (Eds.), *CLARIN in the Low Countries* (pp. 151–165). Ubiquity Press.

Dunning, T.E. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 19(1), 61–74.

Ellis, N.C. (2007). Language Acquisition as Rational Contingency Learning. *Applied Linguistics*, 27(1), 1–24. https://doi.org/https://doi.org/10.1093/applin/ami038

Evert, S. (2009). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus Linguistics: An International Handbook*. Volume 2 (pp. 1212-1248). De Gruyter Mouton.

Evert, S., Uhrig, P., Bartsch, S. & Proisl, T. (2017). E-VIEW-alation - A Large-Scale Evaluation Study of Association Measures for Collocation Identification. In I. Kosem, C. Tiberius, M. Jakubíček, J. Kallas, S. Krek & V. Baisa (Eds.), *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2017 Conference* (pp. 531–549).

Gries, S.T. & Durrant, P. (2020). Analyzing Co-occurrence Data. In M. Paquot & S.T. Gries (Eds.), *A Practical Handbook of Corpus Linguistics* (pp. 141–159). Springer.

Hanks, P. (2013). *Lexical Analysis: Norms and Exploitations*. MIT Press.

Honnibal, M., Montani, I., Van Landeghem, S. & Boyd, A. (2020). *spaCy: Industrial-strength Natural Language Processing in Python*. doi: 10.5281/zenodo.1212303

Ježek, E., Magnini, B., Feltracco, A., Bianchini, A. & Popescu, O. (2014). T-PAS: A resource of Typed Predicate Argument Structures for linguistic analysis and semantic processing. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk & S. Piperidis (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (pp. 890–895). European Language Resources Association (ELRA).

Kilgarriff, A., Rychlý, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. In G. Williams & S. Vessier (Eds.), *Proceedings of the 11th EURALEX International Congress* (pp. 105–116). Université de Bretagne-Sud, Faculté des lettres et des sciences humaines.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, 1, 7–36.

Köhler, L., Middell, G. & Geyken, A. (2025). Choosing Suitable Text Corpora for Identifying Collocations – A Case Study of a Large Reference Dictionary of Contemporary German. In: I. Kosem, M. Jakubíček, M. Medveď, K. Zgaga, Š. Arhar Holdt, T. Munda & A. Salgado (Eds.), *Electronic lexicography in the 21st century (eLex 2025): Intelligent lexicography. Proceedings of the eLex 2025 conference*. Bled, 18–20 November 2025. Lexical Computing CZ s.r.o.

Ljubešić, N., Logar, N. & Kosem, I. (2021). Collocation Ranking: Frequency vs Semantics. *Slovenščina 2.0: Empirične, Aplikativne in Interdisciplinarne Raziskave*, 9(2), 41–70. https://doi.org/10.4312/slo2.0.2021.2.41-70

Odijk, J., Kroon, M., Spoel, S., Bonfil B. & Baarda, T. (2024). MWE-Finder: Querying for multiword expressions in large Dutch text corpora. In V. Giouli & V. Barbu Mititelu (Eds.), *Multiword expressions in lexical resources* (pp. 229–267). Language Science Press.

Pedersen, T. & Bruce, R. (1996). *What to infer from a description* (Publication No. 96-CSE-04). Southern Methodist University, Dallas, TX.

Rychlý, P. (2008). A lexicographer-friendly association score. In *Proceedings of Recent Advances in Slavonic Natural Language Processing, RASLAN* (pp. 6–9).

Uhrig, P., Evert, S. & Proisl, Th. (2018). Collocation candidate extraction from dependency-annotated corpora: Exploring the differences between parsers and dependency annotation schemes. In P. Cantos Gómez & M. Almela Sánchez (Eds.), *Lexical Collocation Analysis: Advances and Applications* (pp. 111–140). Springer.

Van de Cruys, T. (2011). Two multivariate generalizations of pointwise mutual information. In *Proceedings of the Workshop on Distributional Semantics and Compositionality* (pp. 16–20). Association for Computational Linguistics.

van Noord, G. (2006). At Last Parsing Is Now Operational. In P. Mertens, C. Fairon, A. Dister, & P. Watrin (Eds.), *TALN06. Verbum Ex Machina. Actes de la 13e conférence sur le traitement automatique des langues naturelles* (pp. 20–42).

Webber, W., Moffat, A. & Zobel, J. (2010). A similarity measure for indefinite rankings. *ACM Transactions on Information Systems*, 28(4), 1–34. http://dx.doi.org/10.1145/1852102.1852106

Webpages

BlackLab: BlackLab Corpus Search Engine. Retrieved 1 June 2026, from https://blacklab.ivdnt.org/

CHN: *Corpus Contemporary Dutch*. Retrieved May 1, 2026, from https://chn.ivdnt.org/

PDEV: *Pattern dictionary of English Verbs*. Retrieved May 15, 2026, from https://pdev.org.uk/

Sketch Engine. Retrieved 1 June 2026, from https://www.sketchengine.eu

SKELL: *Sketch Engine for language learning*. Retrieved May 15, 2026 from https://skell.sketchengine.eu/#home?lang=en

Woordcombinaties. Retrieved May 1, 2026, from https://woordcombinaties.ivdnt.org/

Contact information

Carole Tiberius
Instituut voor de Nederlandse Taal/Leiden University Centre for Linguistics
carole.tiberius@ivdnt.org

Martin Kroon
Instituut voor de Nederlandse Taal
martin.kroon@ivdnt.org

Lut Colman
Instituut voor de Nederlandse Taal
lut.colman@ivdnt.org

Dalí Chirino
Instituut voor de Nederlandse Taal
dali.chirino@ivdnt.org

Appendix

TABLE A Grammatical relations used for noun entries in the test data

Grammatical Relation Nouns	No of Combinations	Combination	Example
subject of	156	aandeel + stijgen	de **aandelen** stegen ‘the shares rose’
object of	272	aandeel + kopen	ze kopen **aandelen** ‘they buy shares’

TABLE A (*CONTINUED*)

Grammatical Relation Nouns	No of Combinations	Combination	Example
premodifier: adjective	358	lunch + gezond	een **gezonde** lunch 'a healthy lunch'
postmodifier: prepositional phrase or conjunction phrase with secondary collocates	221	aandeel + van + bank	de aandelen **van** de **banken** 'the shares of the banks'
postmodifier: prepositional phrase or conjunction phrase without secondary collocates	14	zus + van	een zus **van** mijn vader 'the sister of my father'
pronoun determiner	82	aandeel + veel	**veel** aandelen 'many shares'
noun determiner	43	aandeel + aantal	het **aantal** aandelen 'the number of shares'

TABLE B Grammatical relations used for verb entries in the test data

Grammatical Relation Verbs	No of Combinations	Combination	Example
noun subject	217	lukken + plan	het **plan** lukt 'the plan succeeds'
pronoun subject	26	lukken + het	**het** lukt 'it succeeds'
noun object	176	lusten + brood	hij lust **brood** 'he likes bread'
pronoun object	46	lusten + iets	hij lust **iets** 'he likes something'
noun indirect object	44	aanbieden + klant	wij bieden **klanten** diensten aan 'we offer customers services'
pronoun indirect object	38	aanbieden + hem	wij bieden **hem** hulp aan 'we offer him help'
adverbial adjunct	149	zwemmen + dagelijks	hij zwemt **dagelijks** 'he swims every day'
prepositional object	37	aandringen + op + verandering	aandringen **op verandering** 'press for change'
prepositional phrase adjunct with secondary collocates	122	zwemmen + in + aquarium	**in** het **aquarium** zwemmen 'swim in the aquarium'

TABLE B (*CONTINUED*)

Grammatical Relation Verbs	No of Combinations	Combination	Example
prepositional phrase adjunct without secondary collocates	23	zwerven + van . . . tot . . .	zwervend **van** paviljoen **tot** paviljoen 'wandering from pavilion to pavilion'
subordinate clause introduced by	8	aanmoedingen + (om) te	aanmoedigen **om** mee **te** denken 'encourage to contribute ideas'

Ágoston Tóth

(UG)TPEX

A (Usage-Guided) Typical Example Score

Abstract The aim of this paper is to explore methods to utilize the linguistic knowledge accumulated in language models to evaluate corpus sentences as potential examples in dictionary entries. We examine the TPEX (Typical Example) measure in three variants in the task of assessing the typicality of a sentence based on how probable the appearance of the headword in the given context is, as per the selected language model. Then we present an extension, the UGTPEX (Usage-Guided Typical Example) measure, in six variants, where the contextual typicality is supplemented by data reflecting the similarity of the headword's contextualized embedding to representations seen in known-good (dictionary-derived, sense-specific) examples. The combination of the two components, i.e. the contextual typicality and the embedding-based similarity, results in a score that is sensitive to both frequent usage patterns and the usage categories (sense categories) established by the lexicographer. We will examine the applicability of these measures in case studies involving five headwords using dictionary example sentences and corpus samples. The proposed measures are for ordering concordance lines, in combination with other ranking methods, as part of the lexicographic workflow.

Keywords Example selection; sense alignment; LLM-assisted human lexicography; BERT

1 Introduction

Professionally selected example sentences in dictionary entries serve several purposes. Examples provide evidence for the existence of the headword (attestation), and by illustrating the word in authentic context, they demonstrate contextual features, grammatical structures, collocations and stylistic characteristics. In the case of polysemous words, examples also assist the dictionary users in finding the correct sense of the word, which is particularly useful for abstract concepts where definitions may not be clear, especially for language learners.

Large Language Models (LLMs) have had an influence on many specific tasks in applied linguistics. Generative LLMs produce a sequence of tokens based on learning from large compilations of authentic text. However, they are not themselves sources of authentic natural language text. Imagine that you take a photo of the Moon with your phone, which replaces it with an AI-generated version to boost visual quality – this is a function that we find in some devices. Does it matter if you see the real Moon or a generated one? If it is a visual accessory, then it may not matter, but when you want to learn how exactly the Moon looked at that exact celestial moment, or you want to discover new detail, it certainly matters. Also consider the issue of hallucination: if you need reliable, factual information, then use authentic data rather than generated data – just to achieve the attestation goal in our case.

According to Atkins & Rundell (2008, pp. 458–461), the criteria for good examples are:

- naturalness and typicality: the example must align with the recurring patterns in the language,
- informativeness: the example must supplement the definition, not contradict it, and help the user understand the concept, and
- intelligibility: the example must be understandable; for instance, it should not contain unnecessarily complex structures or difficult words.

Dictionary editors also handle common issues related to example selection, which include recognizing (and filtering out) individual language use, creative solutions, managing "decontextualization" (an example loses meaning when taken out of context; Atkins & Rundell, 2008, p. 457), as well as noticing "over-contextualization" (where the example encodes more context than necessary; ibid.).

Given the complexity of the task, and also in light of the existing literature (see Section 2), the human lexicographer's experience and intuitions are indispensable, and this lexicographical subtask is not a good target for full automation. In this paper, we explore ways of assisting the human lexicographer in ranking corpus sentences according to typicality – based on context in general and in sense-specific ways – so that they can find relevant examples more easily.

Section 2 includes an overview of the directly relevant literature. In Section 3, the TPEX and UGTPEX measures will be presented, then tested quantitatively and qualitatively in Section 4 with dictionary examples for five headwords.

The present investigation is a proof-of-concept pilot study. The general research question addressed in this paper is whether the proposed measures are suitable to help with the lexicographic task of example sentence selection. To address this goal, the metrics must first be established, then additional research questions will be formulated and answered in Sections 4.2.2 through 4.2.4.

2 Related work

Sinclair et al. (1998) introduced a procedure that could be used as a ranking method, and developed a corresponding program called Typical. The program compared the relative frequency of context words found in the context window of target headwords in corpus sentences to their relative frequencies measured in the entire corpus (Sinclair et al., 1998, p. 233). The context contained 3 words on either side of the target word. In their example, they created a concordance for the word *hot* and sorted it using Typical, which was also observed to group certain collocations containing the headword (e.g., *red hot pokers*, *hot flushes*, *hot and cold flushes*, *blisteringly hot*, *hot and humid* etc.). This procedure also helped distinguish between the headword's meanings, insofar as the emerging collocations reflected the similarities in meaning.

While Sinclair et al. (1998) sought to capture the typicality of headwords as described above, Kilgarriff et al. (2008) aimed to ensure the readability of examples by introducing the GDEX (Good Dictionary Examples) measure. Three main criteria were highlighted: typicality (frequent, characteristic usage patterns), informativeness (clarity of meaning and usage) and readability. The goal is to ensure that the example is clear to the learner, to avoid unnecessarily difficult vocabulary, confusing pronominal and deictic references and elements that are incomprehensible without a specific context. The concept of readability also reflects how "self-contained" the sentence is within the entry. The scoring algorithm approaches the above goals using simple tools: for example, it is sensitive to sentence length, the position of the headword within the sentence, the occurrence of certain characters and words, the presence of the PRON part-of-speech tag, word repetition in the sentence, etc. The best-known, configurable implementation of GDEX scoring is in the Sketch Engine toolkit (Kilgarriff et al., 2014; https://www.sketchengine.eu/guide/gdex).

Since the criteria for a good example are partly language-specific (e.g., typical sentence length, pronoun system, taboo vocabulary), GDEX should be configured in a language-specific way. Kosem et al. (2019) report on GDEX parameter sets for English, Slovenian, Dutch, Estonian and Portuguese.

In addition to the classic, rule-based GDEX approach, there have also been studies that use machine learning (including neural deep learning) methods to predict the extent to which a sentence meets the criteria for a good dictionary example. Khan et al. (2021) combine supervised learning, manually labeled example corpora and word embedding-based neural models, and demonstrate that – with sufficient data – machine learning can emulate the performance of manually tuned GDEX rule systems. This suggestion builds on the GDEX approach, except that the patterns are learned from a training corpus and are not set manually.

Extracting large quantities of contextual information from electronic corpora while learning to represent tokens in large artificial neural networks in a context-specific way contributed to the development of language models (including the model used in the present paper) and modern generative AI.

In the field of lexicography, the question arose shortly after the emergence of ChatGPT as to whether these systems could be used for generating complete dictionary entries. There is work that specifically explores the possibilities of automatically generating example sentences, too (Cai et al., 2024). In the present paper, generated examples will not be used for reasons discussed in Section 1.

The existing practice in Natural Language Processing of using language models to evaluate the fluency and acceptability of sentences was extended by Salazar et al. (2020), who proposed a scoring method for evaluating sentence acceptability that exploited the fact that predicting masked tokens is a native capability of neural language models trained on the Masked Language Modeling (MLM) task.

Building on a similar foundation, in a different context, Tóth (forthcoming) proposed the use of MLMs to evaluate the predictability of the occurrence of the headword in a given sentence as part of a lexicographical task: example selection.

The present paper extends Tóth's above-mentioned proposal while also building on a different line of research. The use of embedding-based word similarity, extracted from the same type of neural language models, has also been studied in the literature. Contextualized token embeddings have been shown to be highly specific to their context, which can also be observed visually after dimensionality reduction and clustering. For a lexicographical application, see Tóth & Abdelzaher (2023); there, in one of the examples, the uses of the English word *mouth* were grouped into clusters roughly corresponding to these meanings: 'mouth', 'mouth of a river', 'human speaker' and 'manner of speaking', which were similar to (although not equivalent to) reference sense categories retrieved from a dictionary.

Tóth & Abdelzaher (2024) examined the extent to which clusters of BERT (Devlin et al., 2019) token embeddings corresponded to lexical meanings. The study combined hierarchical clustering, two-dimensional visualization and detailed quantitative (Silhouette score) and qualitative analyses, demonstrating that BERT representations were not only sensitive to syntactic differences, but also accurately reflected part-of-speech variations, semantic preferences and literal usages versus metaphorical extensions.

Teglia et al. (2025) also worked with contextualized token embeddings, but rather than focusing on a specific lexicographical task, they investigated the extent to which encoder-based language models (BERT, DeBERTa, etc.) could distinguish word meanings based on existing semantic inventories (WordNet, Oxford Dictionary of English). The authors calculated sense centroids and compared model representations going layer by layer, demonstrating, for

example, that DeBERTa-v3's middle layers performed best in a given experiment, and that the structure of reference sense inventories significantly influenced performance.

The UGTPEX proposal introduced in this paper makes use of masked token prediction for calculating contextual typicality (Section 3.1), as well as the similarity of the uses of the headword in potential example sentences to known-good dictionary examples already tied to dictionary senses (Section 3.2).

3 The TPEX and UGTPEX measures

The measures described in this section can be used as lexicographical tools to score potential dictionary example sentences, with the aim of assisting the lexicographer in example sentence ranking by leveraging the linguistic knowledge accumulated in the selected neural language model. These measures are suited to workflows where editors work with large concordance lists and need a sorting technique that prioritizes more typical usages while maintaining an editing process based on authentic, corpus-derived sentences.

3.1 TPEX

The TPEX (Typical Example) metric provides the probability of occurrence of the headword in a given textual context. Encoder-based transformers, such as BERT (Bidirectional Encoder Representations from Transformers; Devlin et al., 2019) and similar models (e.g., RoBERTa, ALBERT, BigBird), are typically trained directly to predict masked tokens in token sequences, which is known as the Masked Language Modeling (MLM) task. This model capability is directly utilized in TPEX: in the potential example sentence, the headword is masked, then the neural network is activated to make a prediction for the token in the masked position. Unlike Transformer decoders (such as GPT), BERT uses the entire context, including the left and right context of the masked token, to make the prediction. For every token it knows, the model indicates how likely the token is to be a candidate for that position by calculating a score (a logit). The higher the score, the more likely the token is to appear in that position, meaning that it is more typical in that context.

Let s be an example sentence containing the headword h, where h is an element of the token vocabulary of the language model. We mask the position of h in the sentence, and let c denote the resulting context. The conditional probability calculated from the values returned by the model for any element (t) of the model's token vocabulary:

$$P(t|c) = \frac{\exp(z_t)}{\sum_{i \in V} \exp(z_i)} \quad (1)$$

where

z_t: the raw model score (logit) produced by the language model for token t

V: the token vocabulary of the model.

Unmasking code libraries, including the one used in this experiment (see Section 4.1), may return probabilities rather than logits by default.

The probability for the *headword* to appear in its masked position:

$$P(h|c) = \frac{\exp(z_h)}{\sum_{i \in V} \exp(z_i)} \tag{2}$$

The most probable token in the given position:

$$t_1 = \arg\max_{t \in V} P(t|c) \tag{3}$$

The suggested TPEX score variants are the following:

Absolute typicality:

$$\text{TPEX}_{\text{abs}}(s) = P(h|c) \tag{4}$$

Relative typicality:

$$\text{TPEX}_{\text{rel}}(s) = \frac{P(h|c)}{P(t_1|c)} \tag{5}$$

Rank typicality:

$$\text{TPEX}_{\text{rank}}(s) = \text{rank}(h|c) \tag{6}$$

where the rank indicates the position of the headword in the probability order predicted by the model; rank 1 is for the most probable token. For UGTPEX measures discussed below, $\text{TPEX}_{\text{rank}}$ is normalized to the [0,1] interval. Let us introduce $\text{TPEX}_{\text{rank-norm}}$ for this purpose:

$$\text{TPEX}_{\text{rank}-\text{norm}}(s) = 1 - \text{minmax}(\text{TPEX}_{\text{rank}}(s)) = 1 - \frac{rank(h|c) - 1}{|V| - 1} \tag{7}$$

Examples will be provided in Section 4.2.1.

3.2 UGTPEX

The UGTPEX measure helps in the selection of example sentences in relation to the senses that have already been identified and documented with examples. This is motivated by the fact that example sentences are typically tied to specific senses within the entries, and one of their functions is to guide the user to the appropriate section (sense) within the entry through prototypical exemplification, especially in the case of highly polysemous or abstract concepts.

The UGTPEX score is made up of two components.

First, we run a neural language model (potentially the same model that we use for calculating the TPEX score) to produce a contextualized token representation (a numerical vector representation of a token produced by the neural language model; an embedding) v_s of the headword, and compare it to embeddings $\{v_e : e \in E\}$ produced for the headword in "known-good" example sentences, which serve as reference points for the given sense in the dictionary. We calculate the word embedding-based cosine similarity, normalized to the [0,1] interval for our use case, as follows:

$$\text{norm\_sim}(s) = \frac{\cos(v_s, v_e) + 1}{2} \tag{8}$$

For candidate sentence s, we select the most similar reference sentence; the similarity of sentence s to this reference sentence is $\max\_\text{sim}(s)$.

UGTPEX scores can be calculated for pairs of sentences, where one sentence is the sentence under scrutiny, and the other one is dynamically selected from a set of reference sentences (dictionary examples) as described above.

Second, for UGTPEX to work, we need the contextual typicality value calculated for the headword in the given example sentence as described in Section 3.1. That value falls within the [0,1] interval, where 1 indicates a perfectly typical usage of the headword in the given sentence.

The UGTPEX score can then be calculated using either multiplication (an AND-like condition):

$$\text{UGTPEX}_{mul}(s) = \max\_\text{sim}(s) \cdot \text{TPEX}(s) \tag{9}$$

or by using addition ('strong in at least one dimension'):

$$\text{UGTPEX}_{add}(s) = \max\_\text{sim}(s) + \text{TPEX}(s) \tag{10}$$

and the TPEX function can be computed using any of the methods discussed in the previous section (TPEX_{abs}, TPEX_{rel} or $\text{TPEX}_{\text{rank-norm}}$). The UGTPEX score falls within the [0,1] interval for multiplication, and within [0,2] for addition.

Examples of how these variants work will be provided in Section 4.2.4.

4 Examining dictionary data and corpus sentences using TPEX and UGTPEX

4.1 Methods

Example sentences from the relevant entries of the online Oxford Learner's Dictionaries (OALD; https://www.oxfordlearnersdictionaries.com) were retrieved for the following headwords: *full*, *mouth*, *serious*, *sound* and *risk*, using all the matching example sentences (including "extra examples") across different part of speech categories. This set of sentences was expanded to 1000 for each headword by adding corpus examples from the British National Corpus (BNC) available via the Sketch Engine (Kilgarriff et al., 2014) using sentence concordancing and shuffled output. Part-of-speech filtering was not used for these examples.

Contextualized word representations for the headwords in the example sentences were produced using a pre-trained BERT model ("bert-large-uncased", https://huggingface.co/bert-large-uncased). BERT (Devlin et al., 2019) is a Transformer encoder: it does not produce a token sequence as an output, but it creates embeddings for input tokens. The embeddings produced in the last layer of BERT at the position of the headword were used in this study. An embedding was a vector of 1024 elements with this particular model. The resulting token embeddings are available at https://github.com/BERThelps/withUGTPEX

All pairwise similarities over the set of 1000 sentences (dictionary examples and corpus sentences) were computed using cosine similarity normalized to the [0,1] interval (norm_sim (*s*), as in Section 3.2). The comparison served two purposes. *(A)* Sentences belonging to the same dictionary sense were examined to see if they grouped in coherent regions of the embedding space. To determine whether the embeddings align with the sense structure provided by the dictionary, the similarity between sentences from the same sense and those from different senses was measured. *(B)* They were also used to determine, for each BNC corpus sentence, which dictionary example it was closest to (which sense and which example), as it was necessary for the computation of the UGTPEX scores (see Section 3.2).

The outputs required for calculating the TPEX scores were obtained using the *predict_mask()* method of the Happy Transformer PyTorch library (https://happytransformer.com). The UGTPEX scores were calculated using a Python script for six UGTPEX varieties as described in Section 3.2: $UGTPEX_{mul\_abs}=TPEX_{abs}\cdot max\_sim$, $UGTPEX_{add\_abs}=TPEX_{abs}+max\_sim$, $UGTPEX_{mul\_rel}=TPEX_{rel}\cdot max\_sim$, $UGTPEX_{add\_rel}=TPEX_{rel}+max\_sim$, $UGTPEX_{mul\_ranknorm}=TPEX_{rank\text{-}norm}\cdot max\_sim$, $UGTPEX_{add\_ranknorm}=TPEX_{rank\_norm}+max\_sim$.

The statistical analysis included the following:

- For the comparison of pairwise sentence similarities (based on the relevant token embedding similarities for the given uses of the headword) the mean, the standard deviation, the minimum and maximum similarities were calculated. Same-sense and different-sense similarities were compared; Welch's *t*-test was used to compare group means. A sense-by-sense mean similarity matrix was compiled to visualize similarity patterns.
- To see if TPEX correlated with semantic clarity (which was *not* expected), Pearson and Spearman correlations were calculated between the contextual typicality of headwords and their embedding-based semantic similarity to dictionary examples. A low or statistically non-significant correlation could signal that high predictability does not imply clear sense alignment.
- To compare UGTPEX parameter combinations (multiplication or addition, the three TPEX varieties), BNC sentences were ranked for each sense using each UGTPEX scoring method. The top-20 sentences were selected for each sense and each UGTPEX variation, and cluster quality was assessed using the mean pairwise similarity and the standard deviation of these similarities. Higher means and lower standard deviations were interpreted as indicating a tighter and more coherent selection.

4.2 Results

4.2.1 TPEX variants

The three TPEX variants described in Section 3.1 sort sentences in slightly different ways.

$TPEX_{abs}$ reflects the absolute probability of the headword in a given context. It assigns very high values to high-frequency expressions with strong collocations (e.g., "full of life," "full of energy," "increased risk of . . . "). The question it answers is: How likely is it that headword *h* appears here?

One of the BNC sentences with the word *full* reads: "By the end of the nineteen fifties, life in Orkney had come full circle, and like everyone else, we had acquired a taste for the material things in life." In this sentence, *full* has maximum probability to appear (1.0). The phrase

"come full circle" is idiomatic, and – according to the model – the word *full* is maximally probable to appear there.

In the OALD dictionary example "The sky was full of brightly colored fireworks" the word *full* also receives a high $TPEX_{abs}$ value (0.9994), as a member of the phrase "full of". Note that the probability of occurrence for *full* mostly depends on the next word in this case, and BERT is capable of taking this into account (as it is "bidirectional"), whereas Transformer decoders, such as GPT, are trained to work with the left context only.

A high $\mathbf{TPEX_{rel}}$ score can highlight contexts where the headword is not very common, but it has no real alternative, such as in more idiomatic constructions. Take, for example, a sentence with a moderate $TPEX_{abs}$ but a maximal $TPEX_{rel}$ score, such as "'Permanent' staff (yourself, the Cook/Kitchen Supervisor and Assistant Cook) are full time". In this example, *full* appears with $TPEX_{abs} = 0.3364$ and $TPEX_{rel} = 1$. The sentence itself is not typical, it is rather convoluted, but the word *full* fits perfectly –according to the model – in this specific context and points to a useful pattern.

The combination of low $TPEX_{abs}$ and high $TPEX_{rel}$ values may indicate highly idiomatic structures. Relevant examples in the data included "hand to mouth", "by word of mouth", "safe and sound", as well as expressions such as "sound of the summer" and "mouth of the Amazon".

$\mathbf{TPEX_{rank\text{-}norm}}$ (normalized rank) smooths out differences in contextual typicality, so that even among hundreds of sentences, the absolute differences become smaller. As we combine TPEX with an embedding-based similarity measurement for UGTPEX, this TPEX variation may be useful, too: we can allow sentences close to the "known-good" reference examples to rank near the top, and we can still factor in enough contextual typicality to ensure that those sentences that represent the typical flow of text (with regard to the use of the headword) are listed near the top.

The TPEX scores calculated for the five headwords used in this investigation are at https://github.com/BERThelps/withUGTPEX.

It is advantageous to be able to select example sentences that align with sense categories previously defined in the dictionary. In what follows, we investigate how similarity measurement based on contextualized token embeddings can be employed in this role.

4.2.2 Same-sense versus different-sense embedding similarity

Do BERT embeddings of headwords exhibit significantly higher similarity within the same dictionary sense compared to different senses?

Embeddings of headwords in sentences for the same sense were consistently more similar. For the word *sound*, for example, the average normalized cosine similarity of sentence pairs representing the same usages was 0.861 (SD=0.040), compared with 0.775 (SD=0.054) for sentence pairs belonging to different senses. Welch's *t*-test: $t=28.17$; $p<0.001$ (i.e., the observed difference in mean similarity was 28.17 estimated standard errors away from zero). This pattern was consistent across the five target words:

- *sound*: same-sense mean similarity was higher than different-sense similarity by 0.0865; Welch's $t=28.174$, $p=6.469 \cdot 10^{-90}$
- *mouth*: same-sense mean similarity was higher than different-sense similarity by 0.0391; Welch's $t=11.960$, $p=8.547 \cdot 10^{-30}$

- *serious*: same-sense mean similarity was higher than different-sense similarity by 0.0337; Welch's t=11.389, p=1.339·10^{-25}
- *full*: same-sense mean similarity was higher than different-sense similarity by 0.0405; Welch's t=9.522, p=5.467·10^{-18}
- *risk*: same-sense mean similarity was higher than different-sense similarity by 0.0298; Welch's t=9.302, p=2.048·10^{-18}

4.2.3 The correlation between TPEX and maximum embedding similarity

Is there a strong correlation between a headword's contextual typicality and its embedding-based semantic similarity to dictionary examples? If so, the combination of TPEX with embedding similarity would be redundant.

In this examination, correlations between TPEX scores and max_sim values were computed. For the first dictionary sense of *full* ('with no empty space'), for example, the Pearson correlation of $TPEX_{abs}$ and max_sim was $r \approx 0.42$ ($p \approx 1.8 \cdot 10^{-41}$), which was one of the strongest correlations seen in this phase of research. For the word *risk*, the highest r was ≈ 0.25 ($p \approx 1.6 \cdot 10^{-14}$), with $TPEX_{rank\text{-}norm}$, sense #5. For the entire set of data, after excluding non-significant correlations, Pearson r ranged from -0.26 to 0.48 ($p<0.001$) depending on the headword, sense category and TPEX variety. While medium-strength correlations were found in specific cases, especially for the words *full* and *sound*, the inclusion of embedding-based similarity was not found redundant in general.

4.2.4 The comparison of the six UGTPEX varieties

How did the different configurations of the UGTPEX measure impact the coherence and other features of selected corpus sentences?

The six UGTPEX varieties were evaluated in the following ways:

1. Per-sense cluster coherence: examining the mean of similarities and standard deviation of the pairwise embedding similarity of the top 20 BNC sentences associated with a given OALD sense. The higher the mean pairwise similarity, the more compact and semantically coherent the cluster is.
2. Aggregated performance across sense categories: the average of the previous sense-level averages across all senses. This provides a global ranking among the six UGTPEX combinations for a given headword.

For the word *serious*, the highest ranking UGTPEX configuration was $UGTPEX_{add\_ranknorm}$ (mean across senses=0.8783; SD=0.0175). The lowest-ranked configuration was $UGTPEX_{add\_abs}$ (mean across senses=0.8201; SD=0.0034). For the remaining four words, the differences between the best and the worst configurations were lower. The strongest configurations for the five words were the following:

- $UGTPEX_{add\_ranknorm}$ for *risk* and *serious*,
- $UGTPEX_{mul\_ranknorm}$ for *mouth* and *sound*,
- $UGTPEX_{mul\_abs}$ for *full*.

When examining the homogeneity of examples for individual meanings, the result depended more on the specific meaning than on the chosen UGTPEX configuration.

In qualitative analysis, we could see, for example, that the first meaning of the word *full* ('with no empty space') yielded the following high-ranking corpus sentence assignments from the BNC corpus using $\text{UGTPEX}_{\text{mul\_abs}}$: "There were cardboard boxes stuffed full of clothes.", "In one corner were three big cardboard boxes full of filthy old jam jars." In a different case, the first sense of *risk* generally attracted medical and statistical examples of 'increased risk' selected from the corpus sentences using the $\text{UGTPEX}_{\text{mul\_ranknorm}}$ variety, with a mean pairwise similarity ranging from 0.9346 to 0.94. Here are two of the top 20 examples: "Smoking low-tar cigarettes does not reduce the risk of heart disease.", "Living close to overhead power lines may increase the risk of cancer in humans." If we examine the same meaning using, say, $\text{UGTPEX}_{\text{mul\_abs}}$, the mean of pairwise similarity is slightly lower, indicating that more heterogeneous sentences are included, such as general examples of the type 'any business venture contains an element of risk', not just medical ones – a behavior that can be useful when selecting examples.

The above observations support the idea that the UGTPEX metrics can be used to select sentences in a usage-specific manner, that is, by prioritizing sentences similar to previously chosen dictionary examples. The lessons learned from comparing them, particularly in qualitative case studies similar to those described above, indicate that it is worthwhile to implement multiple variants of these measures and offer them as options for ranking corpus sentences. With regard to the multiplication vs. addition dilemma: a decrease in either factor (TPEX or embedding similarity) significantly reduces the score in the case of multiplication, so only those sentences that simultaneously exhibit context-based typicality and are very similar to at least one dictionary example are ranked higher. When using addition, sentences that excel in one area may also be included. In the case of the headword *full*, for example, the additive variants allowed more "borderline" sentences to appear (for example, administrative-legal contexts, such as "full investigation", "full European Community VAT reporting"), while the multiplicative variants leaned more strongly toward the classic "full of . . . " examples.

It should be noted that it was common to find sentences with high (UG)TPEX scores that contained lexically undesirable elements or were not optimal for other reasons as illustrative dictionary examples. Therefore, combining these measures with GDEX (Kilgarriff et al., 2008) in an integrated editorial workflow seems advantageous. For instance, when editors need some more examples to illustrate a sense, they could sort concordance lines by GDEX first, then by UGTPEX, and then select good ones based on human expertise and intuition, while also considering the definitions and previously selected examples.

5 Conclusion

In these proof-of-concept experiments, UGTPEX has been introduced as a bridge between general headword typicality and sense-specific example selection. A sentence can be a typical environment for a headword without being strongly aligned with a specific dictionary sense, and vice versa, a sentence can be semantically close to a sense-specific, professionally selected dictionary example sentence without achieving high TPEX values. TPEX alone does not guarantee semantic alignment, but UGTPEX systematically reorders sentences by incorporating semantic similarity to known examples.

The UGTPEX measures extract linguistic knowledge accumulated in BERT (or a similar language model) in two ways. For the TPEX component (Section 3.1), we exploit the model's native pre-trained task, masked language modeling. For the UG component (Section 3.2),

we use the model's ability to compute highly contextualized token representations, which are sensitive to both the structural and semantic contexts of the token. This allows us to measure the similarity of headword usage across sentences, and thus we can use dictionary example sentences as reference points for ranking corpus sentences in a semantically relevant way in a potentially very long list of candidate sentences.

The UGTPEX measure has been introduced in six varieties. Their performance was tested in a series of preliminary quantitative tests and some qualitative case studies. We have seen that UGTPEX provides relevant data and assists sense-aligned example selection.

The suggested method is limited, in practice, by the tokenizer algorithm (and the resulting token vocabulary) used by the selected language model. For the proposed measures to work, a token vocabulary that contains the given headword (and those inflected forms that we want to include in the examples) is required. In language modeling, words may be represented as token sequences. Infrequent words, complex or simplex, especially for languages underrepresented in the vocabulary-construction step, tend to be handled as sequences. Open compounds and other multiword expressions are also handled in this way, as long as standard sub-word tokenization is used. In our case, when a headword is represented as a token sequence, the major alternatives are: *(A)* creating a language model with a custom vocabulary for the given purpose, or *(B)* modifying the UGTPEX measures to cover token sequences. In the latter case, options include multiplying the probabilities under the assumption that a multi-token lexical unit is probable if each part is separately probable based on the context, or using a sum (each token receives a match score, and the score for the entire unit is the sum of these). These are conceptually different approaches, and more research is needed to develop a method that does not favour or disadvantage multi-token words relative to single-token units. The alternative of training new models with all the necessary words (headwords, word forms) included in their token vocabularies requires model pre-training or fine-tuning, which may be computationally expensive. However, Transformer encoders may use far fewer trainable parameters and require less training data and time than their text-generating counterparts (e.g., GPT), which are often meticulously optimized for artificial intelligence applications and multi-lingual environments. Currently, pre-trained language models for a large number of frequent word forms are readily available for English and other languages, but the analysis of the token vocabulary is indispensable.

The analyses shown here are exploratory: only 5 headwords were examined, and sentences that contained the exact headwords were used. The use of frequent, familiar and highly polysemous words was a conscious design choice, however, to test the usefulness of the UG-component of the UGTPEX measure (Section 3.2), rather than being a consequence of the vocabulary limitation described above. Frequent, highly polysemous headwords produce long lists of concordance lines, for which useful ordering of potential illustrative examples is particularly important.

The ultimate feedback on the usefulness of such measures would come from lexicographers who could test them in actual lexicographical workflows, and from dictionary users who make use of the example sentences in the entries in various ways. Interested colleagues are invited to implement and test these measures in their work.

References

Atkins, S. B. T. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press.

Cai, B., Ng, C., Tan, D. & Hotama, S. (2024). Low-Cost Generation and Evaluation of Dictionary Example Sentences. In K. Duh, H. Gomez & S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 3538–3549). Association for Computational Linguistics.

Devlin, J., Chang, M., Lee, K. & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of NAACL* (pp. 4171–4186). Association for Computational Linguistics.

Khan, M. Y., Qayoom, A., Nizami, M. S., Siddiqui, M. S., Wasi, S. & Raazi, S. M. K. (2021). Automated Prediction of Good Dictionary EXamples (GDEX): A Comprehensive Experiment with Distant Supervision, Machine Learning, and Word Embedding-Based Deep Learning Techniques. *Complexity*, 2021, Article 2553199. https://doi.org/10.1155/2021/2553199

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine. *Lexicography*, 1(1), 7–36. http://doi.org/10.1007/s40607-014-0009-9

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In E. Bernal & J. DeCesaris (Eds.), *Proceedings of the XIII EURALEX International Congress* (pp. 425–432). Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra.

Kosem, I., Koppel, K., Zingano Kuhn, T., Michelfeit, J. & Tiberius, C. (2019). Identification and automatic extraction of good dictionary examples: the case(s) of GDEX. *International Journal of Lexicography*, 32(2), 119–137. https://doi.org/10.1093/ijl/ecy014

Salazar, J., Liang, D., Nguyen, T. Q. & Kirchhoff, K. (2020). Masked Language Model Scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 2699–2712). Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.acl-main.240

Sinclair, J., Mason, O., Ball, J. & Barnbrook, G. (1998). Language-independent statistical software for corpus exploration. *Computers and the Humanities*, 31(3), 229–255.

Teglia, S., Tedeschi, S. & Navigli, R. (2025). How much do encoder models know about word senses? In W. Che, J. Nabende, E. Shutova & M. T. Pilehvar (Eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 2266–2277). Association for Computational Linguistics.

Tóth, Á. (forthcoming). TPEX: Neurális nyelvi modellek alkalmazása példamondatok kiválasztásában ['TPEX: The application of neural language models in selecting example sentences'].

Tóth, Á. & Abdelzaher, E. (2023). Probing visualizations of neural word embeddings for lexicographic use. In M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček & S. Krek (Eds.), *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 conference. Brno, 27–29 June 2023* (pp. 545–566). Lexical Computing.

Tóth, Á. & Abdelzaher, E. (2024). BERT may help in lexicographic sense delineation. *International Journal of Digital Humanities*, 6(3), 291–311. https://doi.org/10.1007/s42803-024-00093-2

Contact information

Ágoston Tóth
University of Debrecen, Institute of English and American Studies, Department of English Linguistics
toth.agoston@arts.unideb.hu

Acknowledgements

The present study has been supported by the Institute of English and American Studies, University of Debrecen.

Ágoston Tóth & Esra Abdelzaher

GPT AS A SENSE-TAGGER FOR SELECTING DICTIONARY EXAMPLES

Abstract Aligning a target word in a sentence with a dictionary sense is a challenging task, given the dynamicity of meaning (in context) and the fixed discrete senses in dictionaries, among other reasons. However, it is still a practical necessity for a lexicographer to select example sentences that best illustrate dictionary senses and, hence, find a way to address this challenge. Further challenges appeared after the established use of large corpora as sources of lexicographic evidence and the need to update dictionary entries with additional illustrative examples. The present study tested the usability of an LLM as a facilitator in sense-aligning and example-selecting tasks. The role of the model was limited to assigning a score that reflects the match between (corpus-based and dictionary-cited) examples and existing dictionary senses under several experimental conditions. Quantitative and qualitative analyses showed the usefulness of the model in selecting candidate senses that best represent a dictionary sense, detecting misalignment cases in dictionaries, excluding less representative examples despite their high GDEX scores and identifying highly overlapping senses (especially in the category of adjectives) that are likely to be puzzling to the users.

Keywords Dictionary examples; GPT; sense alignment; sense delineation

1 Introduction

Example sentences or citations have been an integral part of lexicographic resources since Sir Johnson's *The Plan of an English Dictionary*. In addition to proving the existence of a word in a language, they serve multiple pedagogical purposes by illustrating syntactic and phraseological patterns, distinguishing literal and metaphorical meanings and highlighting the distribution of a word across general and specialized discourses (Fontenelle, 2008).

After the shift to corpora as the "evidence base" in lexicography (Rundell et al., 2020, p. 18) and the use of Natural Language Processing (NLP) tools, it has been easier for lexicographers to improve their selection of examples. Features such as word sketches, originally introduced by Kilgarriff and Tugwell (2002), and Good Dictionary Examples (GDEX), proposed by Kilgarriff et al. (2008), enabled the lexicographer to select representative examples of the collocational and colligational associations of a headword from a candidate list of sentences. However, this shift highlighted the need to reconsider the efficiency of dictionary senses and examples in practical contexts. Lexicographers encountered multiple inconsistencies between words as described in dictionaries and the linguistic reality of word use in corpora. Prominent (computational) lexicographers such as Kilgarriff, Hanks and Lew were challenged by the discrepancy between actual language use in corpora and meaning representation as separate sense categories in dictionaries. They not only found one-to-many correspondences between a sentence and multiple dictionary senses, but they also encountered corpus sentences that did not match any of the senses listed in a dictionary. It is not uncommon for a word in a context to activate several meaning components that may map into more than one dictionary sense (e.g., Hanks, 2000; Grefenstette & Hanks, 2023; Lew, 2013).

The same dilemma has been encountered by dictionary users. Whereas the pedagogical value of example sentences for enhancing the encoding performance of dictionary users has been stressed in multiple user-based experiments, the effectiveness of dictionary senses and/or examples in improving learners' decoding performance is still questionable. While some

dictionary users reportedly do not benefit from example sentences in meaning comprehension (Ptasznik, 2023), other users typically find more than one dictionary sense perfectly matching a single word use (Abdelzaher, 2022).

In light of the increasing calls for the addition of further corpus examples to dictionary entries (e.g., Jackson, 2018) and the implementation of these recommendations in several leading online dictionaries, addressing the challenges of lexicographic sense-sentence alignment has become more pressing. The online versions of COBUILD, Oxford Advanced Learner's Dictionary (OALD), Merriam-Webster (MW), Longman Dictionary of Contemporary English (LDOCE) and Cambridge Dictionary (CAMB) have already increased the number of example sentences. However, little or no editorial effort has been exerted to categorize the additional examples into their discrete senses or evaluate their syntactic and semantic value. The absence of sense alignment is a major drawback that limits the usability of the extra examples sections by learners or translators (Josselin-Leray, 2023).

This paper uses LLMs to examine the challenges of lexicographic sense-sentence alignment and facilitate the selection of example sentences. The model is prompted to benefit from the lexicographically edited examples in dictionary entries in order to sort uncategorized examples, sourced from corpora and lexicographic resources, into a lexicographic sense inventory. The study addresses the following questions:

1. Can LLMs detect potential cases of sense-sentence (mis)alignment in lexicographically edited entries that may confuse dictionary users during sense identification?
2. Can LLMs identify possible overlaps between sense categories that may preclude the alignment of a sentence with a single sense?

2 Related work

The literature on the use of LLMs as a sense-tagger in lexicography is extremely diverse across at least six variables: the target language, the list of senses, selection versus generation of examples, the use of conversational versus non-conversational LLMs, the formulation and prioritization of the sense-tagging task, and evaluation methods.

In this study, we enter the lexicographic workflow at a stage when the system of senses for the headword is established, the definitions have been written, and example candidates are available to select from. This choice was deliberately made to avoid the drawbacks reported in previous studies, which prompted LLMs to generate lists of definitions and sets of aligned example sentences. For instance, Conde et al. (2024) tested the ability of several conversational LLMs to produce definitions and example sentences for a set of Spanish words. The qualitative and quantitative analyses showed that not only did the list of senses poorly cover the meanings of the target words, but also that the generated examples were not always aligned with the senses they were intended to illustrate.

Jakubíček & Rundell (2023) reported further problems while generating lexicographic entries for English words, such as false polysemy (i.e., the same sense is paraphrased multiple times), fabricated senses and unnatural examples. Similarly, Marković & Stanković (2025) stressed the recurrent failure of the LLMs in addressing polysemy, particularly when dealing with multiple word expressions (MWEs). The automatic generation of entries for Serbian idioms resulted in the same problems: hallucinated senses, false polysemy, poor alignment, and low-quality example sentences, in addition to misidentifying idiomatic and non-idiomatic uses. Moreover,

Nam et al. (2025) showed that the ability of LLMs to delineate literal and metaphoric uses of Korean neologisms does not guarantee the generation of well-aligned examples for the successfully delineated senses.

Whereas using LLMs in the production phase only (i.e., writing the entry based on the model's suggested senses and sentences) resulted in the abovementioned drawbacks, using the same models during the analysis and synthesis processes obtained better results across languages. Widemann (2025), for instance, used Python code and structured XML files with a human-in-the-loop procedure to prompt LLMs to process corpus examples of Danish headwords. The study did not rely on a pre-existing list of senses. Instead, the model was prompted to extract representative example sentences, cluster them according to their senses and suggest a definition for the sense represented in each cluster. The qualitative evaluation of the output demonstrated significant minimization of the hallucinations and increased transparency, as the source of the examples is verifiable.

Tóth and Abdelzaher (2024) used a non-conversational LLM to identify semantic clusters of corpus and dictionary examples based on their embeddings. Despite the significant gaps between the senses in the automatically suggested clusters and dictionary senses, qualitative and quantitative analyses revealed that the model successfully captured different syntactic, lexical and semantic patterns of the target words. Some dictionary examples evidently failed to illustrate the abstractness of their sense definitions or overspecified a certain use. In this connection, Teglia et al. (2025) also used non-conversational models to organize sense-tagged example sentences into existing sense categories. They analyzed headword embeddings in encoder-based large language models (including BERT), comparing the use of headwords in dictionary example sentences (treating them as ground truth) and other sentences taken from datasets and corpora. They crystalized the role of the sense inventory as a determining factor in the good or poor performance of the model.

3 Methodology

3.1 Data Collection

To test our methodology, we initially collected senses and example sentences for twelve headwords in different parts of speech and frequency bands. Our sample includes content and function words because, as already noted in the literature, the number of sense-tagged examples for function words is always lagging behind even in the most prominent dictionaries (Josselin-Leray, 2023). Whereas the sense inventory includes only sense definitions cited from the online version of Oxford's Advanced Learners' Dictionary (OALD), example sentences were compiled from different resources. First, we include the examples provided for each sense in OALD. Second, we make use of the additional examples recently presented in OALD, COBUILD, Cambridge Dictionary (CAMB) and Merriam Webster (MW). The additional examples in OALD are sense-tagged, unlike the rest of the dictionaries, which barely apply some POS-tagging to the extra examples. CAMB occasionally performs some sense-tagging, but it frequently includes untagged examples for the headwords. We test whether our suggested method may help lexicographers expedite the process of annotating such additional examples. Finally, we extract approximately 300 example sentences with the highest GDEX scores from the BNC for each headword using the default configuration available in Sketch Engine (accessed in February, 2026). GDEX judges the goodness of an example based on the length of the sentence, presence of pronouns and proper nouns and use of C1 words, among others, to maximize the benefits of the example for lexicographic purposes (Kilgarriff et al., 2008). The datasets used in our experiment currently include twelve headwords, 101 senses and approximately 5000 example sentences. Table 1 describes the words in our dataset.

TABLE 1 Description of the target words, senses and examples in the dataset

Headword	Number of senses	Examples division
and (conj)	8	BNC: 300; CAMB: 13; COBUILD: 46; MW: 1
in (prep)	17	BNC: 300; CAMB: 75; COBUILD: 8; MW: 30
absolute (adj)	6	BNC: 300; CAMB: 26; COBUILD: 48; MW: 34; OALD: 5
serious (adj)	6	BNC: 300; CAMB: 41; COBUILD: 30; MW: 30; OALD: 19
see (v)	18	BNC: 285; CAMB: 20; COBUILD: 10; MW: 12; OALD: 32
accommodate (v)	5	BNC: 300; CAMB: 99; COBUILD: 6; MW: 30
bank (n)	7	BNC: 300; CAMB: 10; COBUILD: 56; MW: 31; OALD: 13
introduction (n)	7	BNC: 245; CAMB: 26; COBUILD: 34; MW: 30; OALD: 16
life (n)	14	BNC: 300; CAMB: 5; COBUILD: 45; MW: 12; OALD: 144
viral (adj)	2	BNC: 300; CAMB: 96; COBUILD: 28; MW: 31
say (v)	7	BNC: 300; CAMB: 15; COBUILD: 26; MW: 12; OALD: 17
cloud (n)	4	BNC: 300; CAMB: 5; COBUILD: 10; MW: 30; OALD: 10

3.2 The sense alignment method

In our sense alignment experiment, we tested whether a generative Large Language Model (LLM) was capable of classifying the use of the headword into dictionary sense categories. In our prompts, we presented the sense categories using OALD definitions and corresponding example sentences. The sentences to be classified by the model were meant to help two types of analysis. We used sentences taken from the OALD "extra examples" sections for quantitative analysis: in this dictionary, the extra examples are assigned to sense categories, which allows us to compare the GPT sense classification decisions against professionally edited "ground truth". We also used sentences from other data sources, as described above, without OALD sense alignment, for qualitative analysis.

The LLM prompt included the headword, the definitions of its senses with all the corresponding examples (but the extra examples), as well as clear instructions for the model to align the use of the headword in given sentences to the existing sense categories using scores on a 0-5 scale for each sense. We also requested a rationale to be generated to support the scoring decisions, which we were able to evaluate qualitatively.

We processed the test sentences in batches of 50 to stay within the context size limits of the selected model. For example, we processed the 359 sentences containing the verb *see* in 8 GPT requests in each trial run.

During the pilot phase, we used ChatGPT (free tier with default settings as of February 2026). The results of sense-tagging for the experimental cases were promising. However, the limitation on the length and number of the prompts made it extremely time-consuming to include all senses, extra dictionary examples and corpus dictionaries for even a single headword. In the main experiments, the results of which are reported in this paper, we avoided using the ChatGPT interface, as it offers much less control over reasoning, temperature, consistent LLM model choice over the entire experiment, extra features (e.g., RAG/online search), etc. We used OpenAI's GPT 5.4 nano model via the OpenAI Application Programming Interface (API). This model is recommended for high-volume classification tasks by OpenAI as it is cost-effective and fast. This recommendation matches our use case.

The chosen model also supports reasoning, that is, the generation of additional context (usually a more detailed task specification) before the visible text generation begins. We made use of this operating mode: we compared trials with the reasoning effort set to "high" and with reasoning disabled. When reasoning was disabled, we could set the token generation temperature to 0 to decrease indeterminism and increase reproducibility; the seed value was not constant, however, to allow for variation across test runs. We ran 5 trials with reasoning set to none, and 5 trials with reasoning set to high. We checked the number of input, output and reasoning tokens to ensure compliance with context size limitations. OpenAI does not reveal the result of the (pre-)reasoning phase, but we obtained and used the decision rationales that we asked for in the prompts, with both reasoning enabled and disabled.

Find our GPT prompt template and the GPT sense alignment scores at https://github.com/LLMfor/sensetagging.

The total GPT cost was approximately $30 (+VAT) using the GPT 5.4 nano model via the API in May, 2026. At the time of writing, the standard GPT 5.4 model was approximately ten times more expensive, and the "pro" version was about one hundred times more expensive. In our view, the model choice reflected the financial reality of a typical budget-conscious applied linguistics project.

3.3 Quantitative evaluation: The Sense Purity Score

Many of the sentences that we examined using GPT were OALD "extra examples". To make use of the sense-alignment information available for these sentences, we introduce a measure, the Sense Purity Score (SPS), to evaluate the accuracy of the GPT-assigned scores in the following way:

$$SPS = 100 \cdot \frac{s_c}{\sum_i s_i}$$

where: s_i is the GPT-generated score for sense index i, and c is the index of the correct sense. This measure is optimized for detecting classification into the wrong categories. SPS returns 100% if and only if the correct sense is the only active choice.

In addition to the quantitative analysis, we performed qualitative analysis of the sense alignment scores and the rationale behind them to identify the most beneficial scenarios for a lexicographer making use of GPT models in sense-sentence alignment.

4 Results

We evaluated the sense alignment performance of the selected GPT model quantitatively using sense purity scoring (see section 3.3), where we relied on dictionary examples taken from the "extra examples" sections of OALD entries. Eight of the twelve headwords that we collected data for had relevant, sense-aligned extra examples. Mean and standard deviation values calculated for SPS scores from 10 trials of GPT sense alignment runs (5 with reasoning, 5 without it) are shown in Table 2.

An interesting finding was that reasoning could not boost performance in this task. With regard to statistical means, the model without reasoning, running at temperature=0, outperformed the reasoning model for five out of the eight headwords. In terms of spread, we saw a similar situation. This is despite the fact that the high reasoning effort setting used in our experiments produced considerable text generation effort: the text produced in the reasoning phase

TABLE 2 Mean and standard deviation matrix of sense purity scores (SPS) over 2·5 trials

	no reasoning (5 trials)		reasoning effort=high (5 trials)	
Headword_POS	SPS mean	SPS SD	SPS mean	SPS SD
absolute_ADJ	72	4.73	67	6.04
bank_N	99	1.15	99	1.40
cloud_N	100	0	89	5.55
introduction_N	85	3.12	78	2.93
life_N	58	3.47	62	1.78
say_V	83	1.27	70	10.55
see_V	65	4.71	62	7.75
serious_ADJ	52	2.97	55	3.48

was actually longer than our original prompts. It translated into noticeably higher inferencing time and cost, too.

An important finding of the quantitative evaluation of the GPT sense alignment decisions was the wide range of purity scores: the lowest SPS value was 55% for *serious*_ADJ, which turned out to be the most difficult headword to align in our experiment, and we got 100% purity for *cloud*_N. It shows the importance of qualitative, headword-by-headword analysis, and it also points toward the fact that fully automatic alignment may not be a realistic goal, which is most probably due to the nature of the task rather than model limitations.

Results revealed five scenarios where LLMs can assist sense-tagging and example selection. First, the model mapped a sentence to a single sense and assigned the highest possible score (5) to it. This scenario was frequently associated with good candidate sentences for representing the selected sense. This should speed up the mapping of the extra examples to their relevant senses. Moreover, it also drew attention to cases where the lexicographic sense-tagging was well worth reconsidering. Table 3 shows a sample of the sentences that were mapped to a single sense with a 5 matching score.

Despite the exhaustive effort needed to edit the corpus-driven examples in OALD and align them with their relevant senses, the LLM model pointed to areas that should be revised by lexicographers, as shown in Table 3. As, for example, the dictionary assigns the literal sense of *life* to inanimate objects, the model could successfully detect the metaphoric reading of *life* when attributed to *puppets* and *figures*. Whereas such examples were not frequent, other cases which raised controversy regarding the representativeness of a certain sentence to a specific sense were more frequent. The LLM, for instance, associated *winning the gold medal* with the presence of a kind of performance and accordingly assigned *watch . . . performance* as the best match. The presence of certain collocates seems to be a contributing factor in sense-tagging with high scores. The model assigned all the sentences that contained *absolute* as a target word and *value*, *terms*, *measures*, *zero* or *relative* in a 3-word window to the sense *existing or measured independently and not in relation to something else*, with the highest matching score. The sense *definite and without any doubt* was also selected as the best match to the sentences including *absolute* and *truth*, *certainty*, *answer* and *knowledge*.

The second scenario occurs when the model chooses more than one sense and assigns a high value (4 or 5) to all of them. It frequently suggested a significant overlap among the senses. The differences between the two senses of *absolute*, for instance, seem to be hardly distinguishable from each other. It is very challenging to choose either *total and complete* (sense 1) or *not limited*

TABLE 3 Sample of the cases in scenario 1 and possible lexicographic decisions

Example sentence	LLM-assigned sense	Lexicographer-assigned sense	Possible lexicographic decision
I'd love to *see* her win the gold medal.	to watch a game, television programme, performance, etc.	to become aware of somebody/something by using your eyes	Reconsider sense alignment
Instead teachers can use a *bank* of tests for their own assessments.	an amount of something that is collected; a place where something is stored ready for use	Unannotated	Consider adding it to the selected sense
She creates wonderful puppets, brimming with *life*.	the quality of being active and exciting	the ability to breathe . . . that people, animals and plants have . . . and that objects do not have	Reconsider sense alignment
The canal can *accommodate* vessels up to 1000 tons.	accommodate somebody/something to provide enough space for somebody/something	Unannotated	Consider adding it to the selected sense
These two-dimensional figures are somehow given *life*.	the quality of being active and exciting	the ability to breathe . . . that people, animals and plants have . . . and that objects do not have	Reconsider sense alignment
You can *see* at a glance if all is well.	to become aware of somebody/something by using your eyes	to find out something by looking, asking or waiting	Reconsider sense alignment

in any way (sense 5) in most cases. The same challenge applies when the LLM is asked to choose the best sense for *serious* from a list that separates *not silly* (sense 2) from *not joking* (sense 3) and *thinking . . . in a careful way* (sense 2) from *needing to be thought about carefully* (sense 4). The presence of the same collocates in the sense-aligned sentences included in the prompt and the unannotated sentences in the tagging task was not always fully exploited by the model. For instance, *serious issue* was aligned with the first sense of *serious* by OALD lexicographers and was already in the prompt. However, when the model encountered the same collocate in another sentence, it assigned the highest score to a different sense (sense 4: score=5) while assigning the second-best score to the first sense. Table 4 shows a sample of the sentences which were aligned with more than one sense with a high matching score.

In our third scenario, the model chooses one sense as a perfect match to a sentence and assign low or average values to other senses. The rationale given by the model sheds light on the possibility of activating multiple senses, primarily and secondarily, by the same sentence. Table 5 shows a sample of the target words mapped to multiple senses with a wide gap in the alignment scores.

Whereas scenarios 1, 2 and 3 occur when the model successfully identifies one or more senses as perfectly matching a sentence, scenarios 4 and 5 are relevant to the cases in which the model could not find a single sense as perfectly or highly matching a sentence. Scenario 4 represents cases in which the model assigns average or low values to one or more senses and fails to assign

TABLE 4 Sample of cases in scenario 2 and the overlapping senses

Example sentence	LLM-assigned senses	Lexicographer-assigned sense
He demands *absolute* obedience from his men.	total and complete (score 4); not limited in any way (score 5)	Total and complete
He devoted his *life* to the education of deaf children.	the period between somebody's birth and their death . . . (score 4); a person's experiences . . . the activities that form a particular part of a person's life (score 4)	The period between somebody's birth and their death
Is it ever justifiable to take a *life*?	the ability to breathe, grow, produce young, etc. that people, animals and plants have before they die . . . (score 4); the state of being alive as a human (score 4)	The state of being alive as a human
It can be addictive, for power tends to corrupt and *absolute* power corrupts absolutely.	total and complete (score 4); not limited in any way (score 5)	Unannotated
This is a *serious* issue which must be dealt with on a rational basis.	bad or dangerous (score 4); needing to be thought about carefully; not only for pleasure (score 5)	Unannotated

TABLE 5 Sample of cases in scenario 3 and the rationale for multiple sense-alignment

Example sentence	The best sense (score)	Other chosen senses (score)	Rationale
Chris Patten could reasonably *say* to house-builders: Go to MK.	to speak or tell somebody something, using words (4)	to give particular information or instructions (3); to express an opinion on something (1)	. . . is a direct directive utterance. The verb "say" introduces the spoken instruction, matching Sense 1. Because the content is explicitly an instruction/command, Sense 3 (give particular information or instructions) is also fairly strong, but the headword usage is still the act of speaking, so Sense 1 remains highest.
She came from a very *serious*, Puritan family.	thinking about things in a careful and sensible way (4)	sincere about something (2); Must be treated as important (1)	"Puritan family" licenses the meaning of "serious" as strict, morally solemn, and not frivolous—closest to sense 2 . . . It can also imply a general non-joking attitude (sense 3), but there is no explicit reference to jokes; it's more about temperament/values.
Sometimes they were illustrations for *serious* study or important social statements.	needing to be thought about carefully (5)	Must be treated as important (3); thinking about things in a careful and sensible way (2)	"Serious study" contrasts with lighter/pleasure-oriented uses and points to careful, non-trivial intellectual engagement . . . reinforced by the coordination with "important social statements." There's some overlap with sense 5 . . . Sense 2 (careful and sensible thinking) is related but less directly signaled . . .

TABLE 5 (*CONTINUED*)

Example sentence	The best sense (score)	Other chosen senses (score)	Rationale
Well it's *absolute* rubbish,' I said.	used, especially in spoken English, to give emphasis to what you are saying (5)	total and complete (4); not limited in any way (3)	"Absolute rubbish" is a colloquial intensifier meaning "utterly rubbish," i.e., emphasis on the speaker's judgment. This is a clear match for Sense 2. It also conveys totality (Sense 1) because the rubbishness is portrayed as complete.

4 or 5 to any sense. In this connection, scenario 5 represents a more extreme case in which the model assigns zero values for all senses. Further exploration of such instances and the generated rationale behind the scoring showed a number of lexicographically relevant points. It unveiled cases of POS misidentification in dictionaries, even after manual lexicographic sense alignment in CAMB. It also correctly identified idiomatic uses as not corresponding to any standard sense. Moreover, it pointed to cases where the example sentence may not be beneficial for a user in sense identification, although it may receive high GDEX scores. The rationale for assigning 0 or low scores occasionally pointed to the possibility of a missing sense or the need to avoid the fine-granularity of some senses to accommodate the diversity in word use. It also showed the model's sensitivity to the metaphoric uses that cannot be directly mapped to literal senses, even if they extend from them. Table 6 shows a sample of cases in scenarios 4 and 5.

TABLE 6 Sample of cases in which the LLM could not find a matching sense for a test sentence

Example sentence (GDEX)	LLM-assigned scores	Rationale
And with that she disappears in a *cloud* of tulle. (0.8)	a large mass of something in the air, for example dust or smoke, or a number of insects flying all together (1), 0 for the rest of the senses	Figurative use
Bank curlers help Scots to a famous victory! (0.8)	0 for all senses	Proper noun
Cloud chocolate chip cookies are made to be light and airy (like a cloud, as the name implies), and should be just as fluffy as one, too. (0.5)	a large mass of something in the air, for example dust or smoke, or a number of insects flying all together (1), 0 for the rest of the senses	Figurative use
I was on *cloud* nine. (0.7)	0 to all	Idiomatic use
She must be living in *cloud* cuckoo land. (0.9)	something that makes you feel sad or anxious (2), 0 for the rest of the senses	Idiomatic use
Ten young dancers passed through a series of actions presumably inspired by a short film *introduction* of spinning and blurry but pretty colours and shapes. (0.5)	0 to all	. . . nonstandard way to mean an opening/ depiction that introduces elements in a film.
There was just enough *bank*, eh. (0.9)	0 for all senses	Lack of context
While he was at the library, Steve decided to seek out some information on *accommodation* in the area. (0.5)	to provide somebody with a room or place to sleep, live or sit (1), 0 for the rest of the senses	Misidentified POS

5 Discussion

The consistent scoring of 5 or 0 to one or multiple senses across several reruns and the rationale behind these systematic choices showed the ability of the model to distinguish metaphoric and literal, as well as idiomatic and non-idiomatic uses of a word. This distinction is essential, though not always enough, to successful sense alignment. Previous studies have already reported the ability of LLMs to detect metaphoric and idiomatic uses in different languages (e.g. Beliga & Petrovic, 2025; Nam et al., 2025). Moreover, the model relied on collocation patterns to be consistently aligned with certain senses. This approach has been called for by prominent lexicographers such as Kilgarriff (2005) when he attempted to Put Corpora Into Dictionaries (PCID) by assigning collocates to dictionary senses, and later developed this approach into the Word Sketch tool, dominantly used in modern lexicography. However, it should be noted that some collocates may overlap across two or more dictionary senses. *Absolute truth*, for instance, is shared between the first sense of *absolute* (i.e., 'total and complete') and the third sense ('definite and without doubt') in OALD.

In addition, sense categories may differ from one dictionary to another. Abstracting senses from corpus sentences is a lexicographic decision that may vary considerably, i.e., what a lexicographer may split into multiple senses another lexicographer may lump in a single sense. Tóth and Abdelzaher (2024), for instance, revealed how the distributional representations of headwords in example sentences representing a single sense in OALD may form separate clusters that actually correspond to multiple senses delineated by lexicographers compiling a different dictionary, such as Collins COBUILD.

Still, the abilities the GPT model manifested in the current study can fundamentally nominate a list of sentences for a lexicographer to align with existing senses. In addition, it also spotted cases of misalignment in OALD's extra example section. This misalignment is a major drawback that should be added to the list of challenges lexicographers need to deal with before adding more example sentences to current entries (see Josselin-Leray, 2023, for a detailed report).

The model also showed sensitivity to the linguistic and pragmatic context. Some sentences were aligned to their senses based on their pragmatic readings as directives, evaluative or stance-expressing. Previous studies suggested that part of the pragmatic competence has been transferred to LLMs from the training data, and it enabled the models to identify and interpret metaphors, humor and implicature at least to some extent (e.g., di San Pietro et al., 2023). Moreover, the model pointed out some linguistic contexts of the target words that are too general or ambiguous to be accurately aligned with a sense, despite their high GDEX scores. As Kosem et al. (2019) explained, GDEX is one of the most used automatic measures that attempt to capture authenticity, typicality, informativeness and intelligibility. They suggested improving the coverage of typicality in the GDEX measure by considering corpus domains, text types and issuing year. Based on the results of our experiment, it can be argued that LLMs can serve as sense-sensitive selectors of dictionary examples with high GDEX scores.

The low values of SPS and the dual or multiple choices of senses as best matching for a single sentence highlighted highly overlapping senses in all entries. Whereas lexicographic studies typically consider a single sense as the correct answer in sense identification tasks (e.g., Dziemianko, 2019), the choice of two or more senses in alignment tasks have long been accepted in NLP. Sánchez (2009), for instance, opposed the rejection of dictionary sense inventories as inadequate for Word Sense Disambiguation (WSD) because of their overlapping fine-granular senses. Instead, the study advocated the choice of two or multiple senses when the context of use does not allow clear-cut discrimination.

6 Conclusion

This study explored the possible role of LLMs in facilitating sense-sentence alignment and the selection of example sentences. The results showed that LLMs can be integrated into the lexicographic workflow to detect possible cases of misalignment, especially in the additional example sections, which need massive editorial effort to be categorized and updated. The models can be effectively used to recommend a list of candidate sentences that can best match a specific sense and exclude sentences that may be more puzzling than guiding, despite their high GDEX scores. The study also highlighted how the overlapping senses in dictionaries and the meaning dynamicity in corpus sentences make it almost impossible to align every sentence with a single sense. However, it should be noted that there will be cases in which the model is more confusing than helpful, as explained in scenario 4 in particular.

It should be noted that the results of the current study are based on a limited number of headwords in each POS. Expanding the sample may reveal different directions not only in English lexicography but also across languages. Moreover, we depended on a single GPT model in our experiment and a platform that hides the internal reasoning step (which is standard practice in commercial systems). The use of different models may lead to different quantitative and qualitative results, while open-weight models may reveal internal reasoning. We framed the task, including the prompts and the scoring formula, to detect possible sense overlaps; a different task, such as asking for a single best choice, may likewise affect the accuracy of sense alignment.

References

Abdelzaher, E. (2022). A classroom-based study on the effectiveness of lexicographic resources. *Lexicography*, 9(2), 139–174. https://doi.org/10.1558/lexi.22164

Beliga, S., & Filipović Petrović, I. (2025). AI-and Corpus-Based Strategies for Identifying Phraseme Constructions: A Pilot Study on Croatian Repetitive Constructions. In Kosem, I., Jakubíček, M., Medveď, M., Zgaga, K., Arhar Holdt, Š., Munda, T. & Salgado, A. (Eds.), *Electronic lexicography in the 21st century (eLex 2025): Intelligent lexicography* (pp. 95–115). Lexical Computing CZ s.r.o.

Conde, J., González, M., Melero, N., Ferrando, R., Martínez, G., Merino-Gómez, E., Hernández, J. A. & Reviriego, P (2024). Open Conversational LLMs do not know most Spanish words. *arXiv preprint* arXiv:2403.15491.

di San Pietro, C., Frau, F., Mangiaterra, V., & Bambini, V. (2023). The pragmatic profile of ChatGPT: Assessing the communicative skills of a conversational agent. *Sistemi intelligenti*, 35(2), 379–400. https://doi.org/10.1422/108136

Dziemianko, A. (2019). Homogeneous or heterogeneous? Insights into signposts in Learners' dictionaries. *International Journal of Lexicography*, 32(4), 432–457. https://doi.org/10.1093/ijl/ecz011

Fontenelle, T. (Ed.) (2008). *Practical lexicography: a reader*. OUP Oxford.

Grefenstette, G., & Hanks, P. (2023). Competing Views of Word Meaning: Word Embeddings and Word Senses. *International Journal of Lexicography*, 36(2), 211–219. https://doi.org/10.1093/ijl/ecad005

Hanks, P. (2000). Do word meanings exist? *Computers and the Humanities*, 34, 205–215. https://doi.org/10.1023/A:1002471322828

Jackson, H. (2018). English Lexicography in the Internet Era. In Fuertes-Olivera, P.A. (Ed.), *The Routledge Handbook of Lexicography* (pp. 540–553). Routledge.

Jakubíček, M., & Rundell, M. (2023). The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography?. In Medveď, M. & Měchura, M. & Tiberius, C. & Kosem, I. & Kallas, J. & Jakubíček, M. & Krek, S. (Eds.), *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography*. (pp. 518–533). Lexical Computing CZ s.r.o.

Josselin-Leray, A. (2023). Contextual information in the dictionary: A critical approach of the additional example feature in online monolingual English dictionaries with a focus on translators' needs. *Lexis. Journal in English Lexicology*, 21. https://doi.org/10.4000/lexis.7198

Kilgarriff, A., & Tugwell, D. (2002). Sketching words. In M. Corrard (Ed.) *Lexicography and natural language processing: a festschrift in honour of BTS Atkins* (pp.125-137). EURALEX.

Kilgarriff, A. (2005). Putting the corpus into the dictionary. *In Proceedings the Second Meaning Workshop*. Fondazione Bruno Kessler/FBK.

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M., & Rychlý, P. (2008, July). GDEX: Automatically finding good dictionary examples in a corpus. *In Proceedings of the XIII EURALEX international congress* (pp. 425-432). Barcelona: Universitat Pompeu Fabra.

Kosem, I., Koppel, K., Zingano Kuhn, T., Michelfeit, J., & Tiberius, C. (2019). Identification and automatic extraction of good dictionary examples: the case (s) of GDEX. *International Journal of Lexicography*, 32(2), 119–137. https://doi.org/10.1093/ijl/ecy014

Lew, R. (2013). Identifying, ordering and defining senses. In Jackson, H. (Ed.). *The Bloomsbury companion to lexicography* (pp. 284–302). Bloomsbury Publishing.

Marković, A., & Stanković, R. (2025). So close but still far: Case study on application of LLMs in idioms identification, definition, and generation of illustrative examples. In Kosem, I., Jakubíček, M., Medveď, M., Zgaga, K., Arhar Holdt, Š., Munda, T. & Salgado, A. (Eds.), *Electronic lexicography in the 21st century (eLex 2025): Intelligent lexicography* (pp. 79-94). Lexical Computing CZ s.r.o.

Nam, K., Lee, S., & Jung, H. Y. (2025). Detection and Description of Neologisms in Korean Lexicography: Methodological Issues in Corpus Balance, Word Unit Bias and LLM Assistance. *Lexikos*, 35, 414–438. https://doi.org/10.5788/35-1-2045

Ptasznik, B. (2023). More examples may benefit dictionary users. *International Journal of Lexicography*, 36(1), 29–55. https://doi.org/10.1093/ijl/ecac015

Rundell, M., Jakubíček, M., & Kovář, V. (2020). Technology and English Dictionaries. In S. Ogilvie (Ed.), *The Cambridge Companion to English Dictionaries* (pp. 18–30). Cambridge University Press.

Sánchez, M. A. (2009). The treatment of word sense inventories in the'LACELL WSD project'. *IJES: international journal of English studies*, (1), 21–37.

Teglia, S., Tedeschi, S., & Navigli, R. (2025). How much do encoder models know about word senses? In W. Che, J. Nabende, E. Shutova, & M. T. Pilehvar (Eds.), *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers) (pp. 2266–2277). Association for Computational Linguistics.

Tóth, Á., & Abdelzaher, E. (2024). BERT may help in lexicographic sense delineation. *International Journal of Digital Humanities*, 6(3), 291–311. https://doi.org/10.1007/s42803-024-00093-2

Widemann, (2025). A Pipeline for Automated Dictionary Creation with Optional Human Intervention. In Kosem, I., Jakubíček, M., Medveď, M., Zgaga, K., Arhar Holdt, Š., Munda, T. & Salgado, A. (Eds.), *Electronic lexicography in the 21st century (eLex 2025): Intelligent lexicography* (pp. 30–40). Lexical Computing CZ s.r.o.

Contact information

Ágoston Tóth
University of Debrecen, Institute of English and American Studies, Department of English Linguistics
toth.agoston@arts.unideb.hu

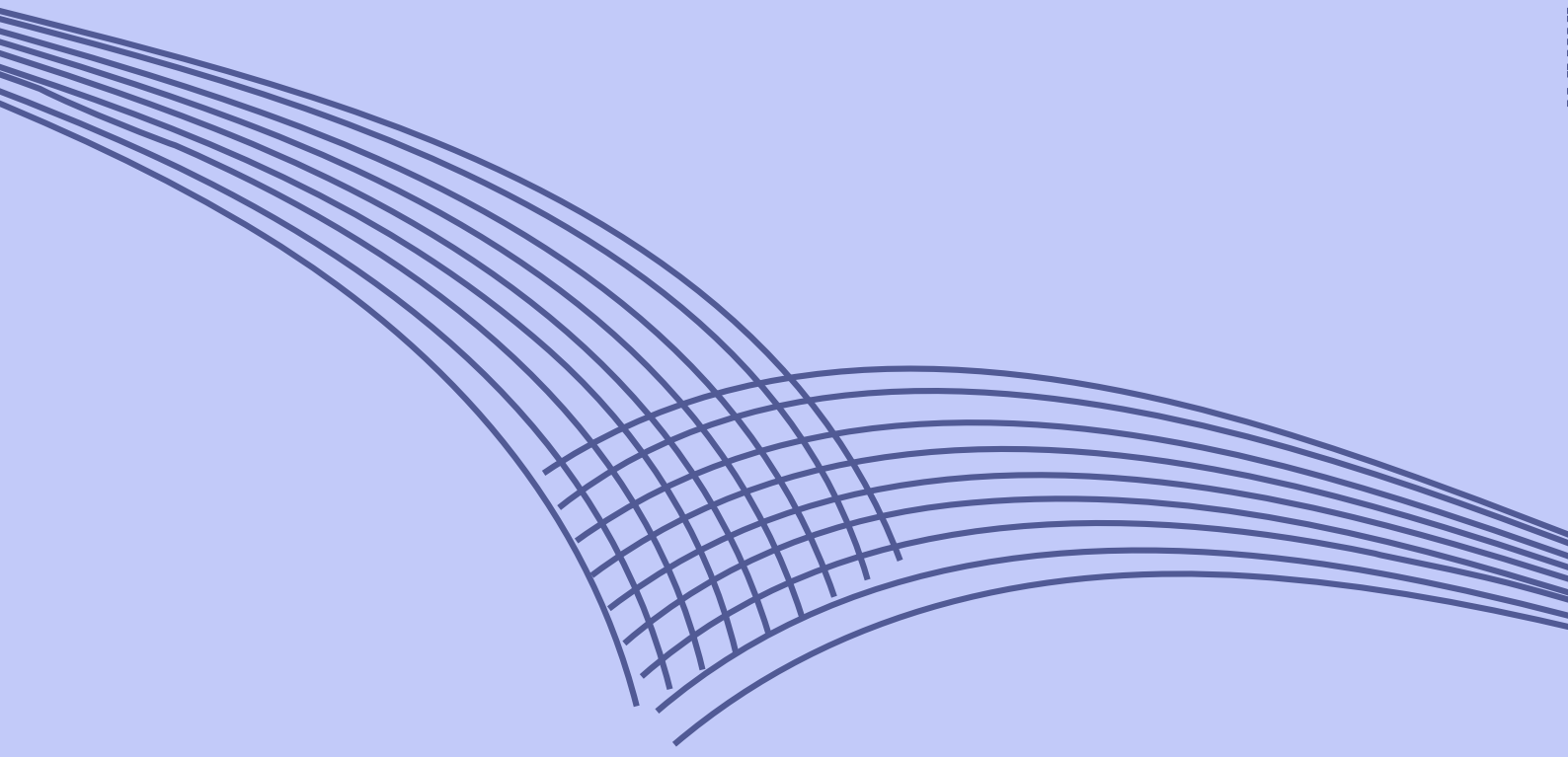
Esra Abdelzaher
University of Debrecen, Institute of English and American Studies, Department of English Linguistics
esra.abdelzaher@arts.unideb.hu

Acknowledgements

The present study has been supported by the Institute of English and American Studies of the Faculty of Humanities, University of Debrecen.

MORPHOLOGY, WORD FORMATION & GRAMMATICAL LEXICOGRAPHY

EURALEX 2026



Ben Black Bear, Jr., Richard Two Dogs & Jan Ullrich

BUILDING THE LARGEST DICTIONARY FOR AN INDIGENOUS LANGUAGE OF NORTH AMERICA

Abstract This paper describes the development of the *New Lakota Dictionary*, a corpus-based dictionary of Lakota that, with 48,000 entries, is, to our knowledge, currently the largest dictionary of an Indigenous language of North America. Drawing on more than three decades of lexicographic work, we discuss the methodological and practical challenges of compiling a comprehensive dictionary for a severely endangered, under-resourced, and highly polysynthetic language. Particular attention is given to corpus construction, lexical documentation, collaboration with native-speaker consultants, and editorial decisions concerning lexical categories, argument structure, orthography, and the treatment of derived forms. We also describe the integration of corpus data into both print and electronic editions, lemmatization, valency marking, inflection charts, corpus search, audio recordings, and monolingual Lakota definitions. Beyond documenting the development of a major lexical resource, the paper illustrates how sustained collaboration between native speakers, linguists, and software developers can support both rigorous lexicographic description and long-term language revitalization.

Keywords Endangered language; bilingual and monolingual dictionary; corpus; polysynthetic language; print and electronic dictionary

1 Introduction

Over the past decades, several major dictionaries of Indigenous languages of North America have been published, reflecting a variety of lexicographic traditions and methodological approaches. The New Lakota Dictionary (NLD) contributes to this body of work by documenting 48,000 lexical entries (excluding the English reversal entries) and by integrating a large corpus of historical and contemporary texts, extensive collaboration with native speakers, audio recordings, and computational tools. Compiling a comprehensive corpus-based dictionary for a severely endangered language presents challenges fundamentally different from those encountered in the lexicography of well-documented languages. Limited textual resources, declining numbers of native speakers, and restricted funding often make it difficult to systematically document the lexicon. This paper describes how these were addressed during the development of the *New Lakota Dictionary*.

An earlier Lakota dictionary containing 20,000 entries was compiled by the Jesuit missionary Eugene Buechel during the first half of the twentieth century (Buechel, 1970). For decades, it served as the principal reference work for the language, but like many dictionaries of its time, it reflects the limited documentary resources and linguistic knowledge then available. Much of its lexical material and many English glosses were adopted from the earlier Dakota dictionary compiled by Riggs (1852), based on the assumption that closely related Dakota cognates generally shared the same form and meaning in Lakota. Subsequent research has shown that this assumption was often inaccurate. The dictionary also employs an orthography that is not fully phonemic and contains lexical and semantic analyses that have since been revised in light of extensive corpus evidence and collaboration with native speakers. These limitations, together with the availability of a much larger body of textual and spoken data, demonstrated the need for a new comprehensive dictionary based on contemporary lexicographic principles. The NLD was not conceived as a revision of Buechel's dictionary, but as an entirely new lexicographic description of Lakota based on corpus evidence and extensive work with native speakers.

The project was initiated in 1992 by two Lakota speakers from Pine Ridge, SD, Rudolf Fire Thunder and John Around Him, who invited Jan Ullrich to serve as editor. Since 2003, the project has been carried out under the auspices of the Lakota Language Consortium (LLC), a nonprofit organization dedicated to language preservation and revitalization, whose board of directors represents all Lakota communities. The founding of the LLC was supported by a 2002 resolution of the Oglala Sioux Tribe council. The NLD is, to our knowledge, the only Lakota dictionary owned by a Lakota organization rather than a private or academic publisher.

Throughout the project, editorial decisions have been developed in close collaboration with native-speaker consultants. Two highly competent native speakers, Ben Black Bear, Jr. and, later, Richard Two Dogs served as the principal native-speaking consultants responsible for the linguistic verification and editorial review of entries, while many other fluent speakers contributed their linguistic knowledge during different stages of the project. Throughout its history, the project has relied on sustained collaboration between native Lakota speakers and a linguist.

Lakota is a critically endangered Siouan language spoken on five large reservations in South Dakota and North Dakota, as well as in several smaller communities in Montana and Canada. Typologically, it is a predominantly head-marking, left-branching language with a high degree of polysynthesis. Its verbal system is characterized by split intransitivity (Split-S), a fundamental distinction between active and stative verbs, and extensive derivational morphology. These typological characteristics have important consequences for lexicographic description, particularly with respect to lexical categorization, argument structure, valency, and the treatment of derived forms, all of which require editorial solutions that differ from those typically employed in dictionaries of well-documented Indo-European languages.

Although reliable statistics on the number of speakers are unavailable, estimates place the number at approximately 10,000–15,000 in the early 1990s. Today, the number is closer to 2,000, with the decline accelerating during the COVID-19 pandemic because most fluent speakers belonged to the oldest generation. The past three decades, therefore, represented a critical period for documenting the language while extensive knowledge of its vocabulary, grammar, and traditional usage was still available from first-language speakers.

This paper describes the methodology used in compiling the NLD and discusses several lexicographic challenges posed by an endangered polysynthetic language. Particular attention is given to corpus construction, lexical organization, orthographic standardization, lexicographic software, and the integration of computational tools.

The challenges described here are not unique to Lakota. A growing body of work addresses the lexicography of endangered and under-resourced languages specifically, including its distinct constraints on corpus size and balance, the integration of community consultation into editorial workflow, and the trade-offs of adapting general-purpose lexicographic software to languages with typologically unusual morphology (Corris et al., 2002; Mosel, 2011). More broadly, corpus-based lexicography for smaller languages raises methodological questions distinct from those addressed in the corpus lexicography literature for well-resourced languages, where corpora of hundreds of millions or billions of words are standard (Hanks, 2012; de Schryver, 2010). The methodological decisions discussed in this paper — regarding corpus construction, entry architecture, and computational tool development — are offered as a case study within this broader effort.

A key factor in the project's success was the availability of a large corpus of Lakota texts written by or recorded from native speakers, including materials dating back to 1871. Beginning in the

1960s, additional efforts by tribal colleges and individual researchers focused on recording oral narratives, resulting in an extensive corpus documenting the language as spoken by highly fluent speakers over the past five decades. This combination of historical written sources and modern audio recordings provided the empirical foundation on which the dictionary was compiled.

2 Scale comparison

The chart below compares the NLD with a selection of major dictionaries of Indigenous languages of North and Central America, Hawai'i, and New Zealand. Although the list is not intended to be comprehensive, it includes many of the larger published dictionaries with more than approximately 10,000 lexical entries and provides context for the scale of the NLD.[1] A comparison of entry counts across these dictionaries should be interpreted with caution, as lexicographic traditions differ considerably in how an "entry" is defined and counted. Some of the dictionaries listed may count only headwords, while others count individual senses, subentries, or derived and inflected forms separately; some are print publications with a fixed, closed entry list, while others are continuously updated electronic databases. The entry count for the NLD reported here (48,000) refers to Lakota headwords with distinct lexical status, excluding English reversal entries, but including archaisms, neologisms, and geographical names as separate entries. Because comparable methodological information is not available for all dictionaries in the list, and was in several cases estimated rather than confirmed by the original compilers, the entry counts below should be read as broadly indicative of project scale rather than as a precise, methodologically uniform metric.

Language	Dictionary	Approximate entries
Lakota (Lakȟóta)	Black Bear, Two Dogs, Ullrich (2026)	48,000
Dakota (Dakhóta)	Dakhóta Iápi Okhódakičhiye (2025)	31,000
Hawaiian	Pukui and Elbert (1986)	30,000
Hopi	Hopi Dictionary Project (1998)	29,394
Plains Cree	The itwêwina Plains Cree Dictionary	25,000
Arapaho	Cowell (2026)	25,000
Maori	He Pātaka Kupu (monolingual online dictionary)	24,000
Kaskaskia-Illinois	Masthay (2002)	ca 23,000
Navajo	Young and Morgan and Midgette (1992)	ca 21,000
Nahuatl	Wood (2026)	20,000 +
Ojibwe	Ojibwe People's Dictionary	20,000
Lakota (Lakȟóta)	Buechel (1970, 2002)	20,000
Choctaw	Byington (1915)	19,888
Cheyenne	Fisher et al. (2013)	18,000
Dakota (Dakhóta)	Riggs (*1992* [1852])	18,000
Yup'ik Eskimo	Jacobson (2012)	16,000
Koyukon	Jetté, and Jones (2000) (unpublished database)	15,177
Maori	Williams (1971)	15,000
Cherokee (Tsalagi)	Feeling (1975)	15,000
Tlingit (Língit)	TlingitLanguage.com (2019)	15,000
Saanich	Montler (2018)	11,518
Crow	Crow Language Consortium (2022)	11,000
Western Apache	Bray (1998)	ca 10,000
Klallam	Montler (2012)	9,653

[1] We want to thank Willem de Reuse for his help in compiling parts of this list and providing some of the estimates for those dictionaries in which the number of entries is not disclosed.

Language	Dictionary	Approximate entries
Halkomelen	Galloway (2009)	8,014
Alabama	Sylestine et al. (1993)	ca 8,000
Creek	Martin and Mauldin (2000)	7,000
Oneida	Michelson Doxtator (2002)	ca 7,000
Tuscarora	Rudes (1999)	ca 7,000
Northern Paiute-Bannock	Liljeblad, Fowler and Powell (2011)	ca 6,500
Kaw (Kaáⁿze)	Cumberland and Rankin (2012)	5,000
Skiri Pawnee	Parks and Pratt (2008)	4,500
Hocank	Helmbrecht and Lehmann (2010)	4,000

The comparison demonstrates that the NLD ranks substantially above the largest dictionaries compiled for Indigenous languages of North America. At the same time, dictionary size alone does not determine lexicographic quality or methodological sophistication. The purpose of this comparison is simply to provide context for the project's scale. Among the dictionaries listed, the Hawaiian Dictionary (Pukui & Elbert, 1986) and the Hopi Dictionary (Hopi Dictionary Project, 1998) are widely regarded as landmark works of Indigenous-language lexicography and have received high praise from reviewers (e.g., Ashford, 1987; Bright, 1999).

The Dakota dictionary project, compiled by the Dakhóta Iápi Okhódakičhiye partnered with the Lakota Language Consortium, which expedited their dictionary expansion by incorporating a large number of cognates and Dakota variants from our lexical data. They verified this data with Dakota speakers and historical sources, and have been working on building their text corpus to also include examples in their dictionary's entries (Bendickson, 2026). Not included in the chart is a Blackfoot relational database of 63,493 lexical forms because this number comprises inflected words, stems, and morphemes and does not represent the number of lemmas (Weber et al., 2023).

The purpose of the chart is not to compare the scholarly merits of these dictionaries but to place the NLD in the broader context of major lexicographic projects for Indigenous languages. Beyond its size, the principal contribution of the NLD lies in its corpus-based methodology, community-driven effort, extensive collaboration with native speakers, and integration of computational tools, as discussed in the following sections.

3 Lexicographic coverage and types of entries

The lexical database of the New Lakota Dictionary grew gradually. In the first phase, between 1992 and 2003, about 15,000 Lakota entries were created. By the time the first edition of the dictionary was published in 2008, it had 21,000 entries, surpassing the size of the earlier missionary dictionary while incorporating many previously undocumented words, most from recordings of contemporary speakers and some from older written sources. The second edition, published in 2011, contained 25,000 entries, and the third edition (2022) had 41,000 entries. As of 2026, the electronic version has 48,000 words documented from recordings and texts, representing more than 3,500 native speakers. Our work on expanding the lexical database continues.

The NLD takes an all-inclusive approach to documenting the lexicon. Entries include not only words documented from or recognized by contemporary speakers, but also archaisms and words that have gone out of use entirely (provided their pronunciation, semantics, and argument structure can be reliably reconstructed from textual evidence). This is to enable students

of the Lakota language, history, and culture to read even the oldest texts. It also contains highly specialized vocabulary, such as many entries on plant names, traditional tools and several hundred geographical names together with cultural and encyclopedic information. In this sense, the NLD documents not only the Lakota language but also Lakota culture.

Among the entries are many neologisms as well. We established a committee to create neologisms and collaborated with various immersion schools in Lakota communities to meet their needs for new words. Following contact with Europeans, Lakota speakers spontaneously coined numerous words for newly introduced objects and concepts while adopting relatively few borrowings. This process largely ceased by the 1950s, leaving lexical gaps that later needed to be filled for immersion education and contemporary communication. Our team has created numerous useful neologisms, many of which have been adopted by immersion schools and by second-language learners who access them through the regularly updated NLD electronic application.

The dictionary also includes some spelling and pronunciation variants, though these do not significantly increase the total number of entries. Variants from the two closely related Dakota dialects (Western Dakota and Eastern Dakota) are also included, but as they are listed inside Lakota entries, they do not increase the total entry count.

Each entry is identified and documented through corpus analysis, subjected to grammatical analysis by the editorial team, and then verified in consultation with native speakers. Native-speaker consultation remains particularly important for low-frequency words, whose corpus attestations often provide insufficient evidence for inflection, morphophonology, semantic distinctions, or usage.

4 Lakota corpus and other sources of words

From the outset, the NLD was conceived as a corpus-based lexicographic project. Work therefore began with digitizing historical texts and transcribing audio recordings. Because the available material was insufficient in both size and genre diversity, the corpus was systematically expanded through recordings of contemporary speakers. Between 1992 and 2026, over 500 contemporary speakers were interviewed for the project, resulting in many hours of additional audio-recorded material that needed to be transcribed.

For approximately the first fifteen years, the project operated as a grassroots initiative with virtually not financial budget, so the linguistic work was carried out on a voluntary basis. Limited funding initially slowed digitization and transcription, requiring the editor to manually identify lexical items in texts and recordings before computational corpus methods became feasible.

Although modest compared with corpora of major world languages, the Lakota corpus is exceptionally large for an endangered Indigenous language. We estimate that the Lakota corpus contains around five to seven million words (lexical tokens); the reason why this is an estimate is that a large proportion of the available data has not been digitized. This estimate places the Lakota corpus in a similar range to corpora used in other well-regarded corpus-driven dictionaries of smaller languages, such as the approximately 7.5-million-word corpus underlying the Zulu dictionary discussed by de Schryver (2010) and Hanks (2012), and above the roughly one-million-word threshold that Hanks (2012) suggests is needed to reliably distinguish systematic from accidental patterns. The Lakota corpus continues to expand through ongoing digitization and documentation. It is used to identify new lexical items, collocates, multi-word expressions, and patterns of Lakota morphosyntax. Thanks to our efforts to expand

the corpus, we began creating and publishing monolingual Lakota literature, which is a necessary component of a successful language revitalization (see especially Ullrich, 2020; Two Dogs & Ullrich, 2025).

Despite our efforts to enlarge the corpus by recording more narratives, it remains too small, unbalanced in genre, and with limited lexical diversity. In particular, it lacks various expressions commonly used in daily conversation, as well as specialized vocabulary from various semantic domains. Therefore, we sought words using methods other than corpus search.

Additional lexical material came from three supplementary sources. Firstly, our editorial team's workflow is typically conducted in Lakota, and we also participate in spontaneous conversational interaction with other Lakota speakers. These immersive events have helped us identify many conversational words and collocations. Secondly, many lexical items were reported to us from informed educators and community members. The third source of words is elicitation. Because translational elicitation is well known to yield unreliable lexical and grammatical data (Chelliah & de Reuse, 2011; Mosel, 2012), we employ it only sparingly and under carefully controlled conditions, typically prioritizing active or domain elicitation over translational elicitation (Mosel, 2004). We ensure that the reported and elicited words are meticulously cross-checked and undergo the same rigorous attestation process as corpus-derived words. This combination of corpus evidence, careful supplementary documentation, and independent verification by native speakers enables the NLD to maximize lexical coverage while maintaining high standards of linguistic reliability.

5 Orthographic standardization

One of the first lexicographic decisions in compiling the New Lakota Dictionary was the selection of an orthography. Because several competing writing systems were in use, establishing a consistent orthographic standard was essential for dictionary compilation and language pedagogy. The first orthography was created by Christian missionaries for the closely related Dakota language in the 1830s to facilitate Bible translation. The missionary orthography did not mark aspiration or stress, both of which are phonemic in Lakota. Although native speakers often recover these distinctions from context, their omission substantially impairs their level and speed of reading comprehension, literacy development, pronunciation, interpretation of historical texts and development of phonological awareness among second language learners.

Building on the orthography developed by Boas and Deloria (1941), the University of Colorado Lakota program (Rood & Taylor, 1976), working extensively with native speakers from Pine Ridge and Rosebud, introduced several practical modifications, including replacing various superscript diacritics with the háček (*č, ǧ, ȟ, š, ž*) and simplifying the notation of aspiration and glottalization. During the 1970s, this orthography was adopted by Ben Black Bear, Jr., who was the chair of the Lakota language department at the Sinte Gleška University. Although a new department chair eventually abandoned this orthography, it was eventually approved as the writing system for the NLD project, with three additional modifications agreed upon by Lakota elders. The decision was made collectively by a group of respected Lakota elders, including Johnson Holy Rock, Leonard Little Finger, Matthew Kayo Bad Heart Bull, Catherine Looking Elk, Elizabeth Makes Him First, Ben Black Bear, Jr., Richard Broken Nose, etc.

During the first decade after the dictionary's publication, the orthography was probably its most controversial aspect because many native speakers were understandably attached to the traditional missionary spelling. However, through extensive teacher training, educational workshops, and the growing use of the NLD and related teaching materials, the orthographic

reform gradually gained increasing acceptance among Lakota language educators and adult learners. Many respected elders became strong advocates of the revised orthography, including Philomine Lakota, Richard Two Dogs, Levi Left Hand, Sandra Black Bear, Ben Black Bear, Jr., Dewey Bad Warrior, Paulette High Elk, Chief Dave Bald Eagle, Jay Taken Alive, and Delores Taken Alive.

As with many orthographic reforms around the world, the orthography continues to draw its share of critics, and debates about spelling are occasionally intertwined with broader discussions of language revitalization. Nevertheless, it is now used by many immersion schools and most bilingual programs. The NLD and its orthography have been extensively used by Oglala Lakota College in Pine Ridge and Sitting Bull College in Standing Rock. Although some institutions continue to use various versions of the missionary orthographic conventions, younger instructors increasingly adopt the NLD and its orthography in their teaching. The increasing adoption of the orthography reflects both its practical value for developing phonemic awareness and native-like pronunciation, and the availability of a comprehensive dictionary together with associated pedagogical resources, such as the comprehensive pedagogical grammar by Ullrich and Black Bear, Jr. (2016).

It must be added here that we use the term standardization only to refer to establishing a single spelling convention for representing individual phonemes of the language, rather than promoting one variety of the language over another. Lakota is phonologically homogeneous, and consequently, orthographic standardization does not affect the use of local variants. The NLD documents and gives equal status to the relatively few geographic variants.

The adoption of a consistent orthography has been essential not only for compiling the NLD but also for ensuring consistency across educational materials and digital language resources.

6 Refined lexical categories and grammatical analysis

The compilation of the first edition relied largely on existing grammatical descriptions of Lakota. However, continued corpus expansion and lexicographic analysis after 2008 revealed that several traditional grammatical categories required revision. For instance, our corpus research showed that the words that traditional Dakota and Lakota dictionaries and grammars (Buechel, 1939; Boas & Deloria, 1941; Rood & Taylor, 1996) treat as adverbs actually belong to several distinct lexical categories, specifically (i) adverbs, (ii) converbs, (iii) coverbs, and (iv) attributive derived modifiers (see Ullrich, 2018, 2022b). For example, the word *šayá* (derived from the stative verb *šá* 'to be red') is traditionally categorized as an adverb and glossed as "redly" (Riggs, 1992, p. 441; Buechel, 1970, p. 461), but Ullrich (2018) showed that these words exhibit a variable modificational scope, most commonly modifying morphologically coded arguments, and as such cannot be considered adverbs. We categorize them as derived modifiers (DMOD) and our gloss for *šayá* is "red, being red".

Words that the Lakota research literature traditionally categorized as postpositions (Buechel, 1970; Pustet, 2000) fall into the following two categories: (i) postpositions proper, (ii) coverbs with postpositional semantics (Ullrich, 2026a). Many of these findings emerged directly from the lexicographic process rather than from a separate grammatical research project.

Another example concerns two types of dative verbs in Lakota, first described by Buechel (1939). Rood and Taylor (1976, 1996) categorized these two groups as benefactive verbs and dative verbs, following Boas and Deloria (1941), who asserted that these two types of verbs signal activities performed with, and without consent of the affectee, respectively. According to this analysis, a verb stem such as *ičú* 'to take sth', receives the affixes *ki-* and *kiči-* to derive

into the dative verb *ikíču* 'to take sth from sb' and a benefactive verb *ikíčiču* 'to take smth for sb', respectively. However, Ullrich (2013, 2022b) provided evidence that the semantic import of these affixes is context-dependent and that both types of derivative verbs can variably code benefaction, malefaction, recipient, and other semantics typically associated with the dative case. Consequently, instead of the terms **dative** and **benefactive**, we adopted the terms **first dative** (DAT1) and **second dative** (DAT2) as the subcategories of verbs with the two respective affixes. Creating English glosses for such verbs can be challenging because of the context-driven nature of their semantics. For example, both *ikíču* and *ikíčiču* can be glossed as "to take sth for/from sb" or more neutrally as "to take sth pertaining to sb or affecting sb". The revised analysis required corresponding revisions to all dictionary entries for dative verbs before the second edition.

As part of our ongoing research on Lakota morphosyntax, we also revised the syntactic behavior of stative verbs, which is one of the largest open classes of the Lakota lexicon. The traditional account of post-nominal stative verbs in Lakota analyzes their syntactic function as modification, consequently treating them like adjectives in familiar languages. However, Ullrich (2018, 2020, 2026b) considers this the result of an Indo-European bias and, in his analysis, a post-nominal stative verb functions syntactically as a co-subordinate of the noun, despite semantically conveying an attributive property of the noun. This explains various constraints on Lakota stative verbs, such as their inability to form chains.

Our lexicographic research also extended beyond lexical categories. Before the third edition, we began investigating Lakota prosody and its interplay with morphosyntax, which helped us develop standards for punctuation and use of hyphens to mark lexical and syntactic compounding (Ullrich, 2022a). This positively improved the representation of example sentences in the dictionary.

More broadly, the sustained interaction between corpus research, lexicographic analysis, and native-speaker consultation has led to substantial revisions of earlier descriptions of Lakota lexical categories and grammar. These findings have improved not only the linguistic accuracy of dictionary entries but also the translation of examples, pedagogical materials, and subsequent grammatical research. The lexical categories and their syntactic functions are described in detail in the extensive grammar section inserted between the Lakota and English sides of the dictionary (Ullrich, 2022b).

7 Representing argument structure, valency, and derived verbs

Every Lakota verb has an inherent fixed valency ranging from 0 to 4, and, with a small handful of exceptions, there are no ambivalent verbs in the language. Accordingly, every verb entry in the dictionary includes a valency number, and the English glosses are designed to reflect the verb's argument structure. This is illustrated with four simplified entries below. (The abbreviation *AV* stands for "active verb", the digits indicate valency, and the square brackets provide a morphological breakdown of the entry word).

entry word	lexical category	morphology	gloss
lowáŋ	AV:1		to sing
ahíyayA	AV:2	[a- hiyáyA]	to sing sth
kahíyayA	AV:3 (dat1)	[ki- ahíyayA]	to sing sth for sb
akíčičaȟpA	AV:4 (dat2)	[kíči- akáhpA]	to cover sth for sb with sth

Note that older Lakota dictionaries typically glossed such entries merely with "to sing" or "to cover". Such glosses obscured the verb's argument structure and prevented learners from learning how to use them correctly. The NLD, on the other hand, helps learners by introducing valency numbers and consistently using glosses that illustrate the argument structure.

Numerous derivational affixes increase or decrease valency and alter the argument structure and sometimes even the semantics of the derived verbs, which has important lexicographic consequences. Derivational processes can yield 17 distinct derivative forms, including possessive, reflexive, reciprocal, causative, and several other derivational categories. A comprehensive list of derivatives is given at the end of each verb entry, as illustrated in Figure 1.

Deciding how to represent derivative verbs on the English side of the dictionary was an important and complicated decision. In most cases, the derivatives fall under the same sense as the base verb, which had to be reflected in the architecture of the entry, as illustrated in Figure 2. However, some English entries contain so many Lakota equivalents that it would be impractical to also include all derivatives. In such cases, the derivatives are listed only in the Lakota entry. For example, the English reversal entry "cut" has 44 Lakota equivalents, so including all of the derivatives would result in an enormously long entry, which would also be very difficult to navigate. Nevertheless, representing derivatives in the Lakota entry and, when possible, also in the English reversal entries, substantially improves users' ability to identify verbs with the desired argument structure.

The NLD treats all derived verbs as independent lexical entries because they frequently differ from the base verb in both argument structure and meaning. For instance, the verb *nážiŋ* 'to stand' can be derived with the dative 2 affix, resulting in *nakíčižiŋ*, which literally means "to stand for/on behalf of sb" but the idiomatic meaning is "to defend sb".

waŋyáŋkA *AV:2* **1** to see smth/sb. *USAGE:* see also the contracted form, which can express either 'to see smth/sb' or 'to look at/watch smth/sb' depending on the verb which follows it. *VARIANT: REDUCED SP.* **waáŋka**. • **Wičháša kiŋ hé waŋbláke šni.** *I haven't seen that man.* **Waŋmáyaŋke šni séče.** *It seems that he didn't see me.* **Owátohaŋyaŋ waŋčhíyaŋke šni.** *I haven't seen you for a while.* **Íthehaŋtaŋhaŋ waŋwíčhauŋyaŋkapi.** *We saw them from afar.* **Waŋmáyalaka yačhíŋ kéyapi.** *They said you wanted to see me.* **Akhé waŋčhíyaŋkiŋ kte.** *I will see you again (used as a farewell).* **Waŋúŋyaŋkapi.** *We saw him/her/it.* **2** meet sb. *USAGE:* mostly in phrases like "nice to meet you," compare with **átayA**. • **Iyúškiŋyaŋ waŋčhíyaŋke.** *Nice to meet you.* **Hé tóhaŋni waŋbláke šni.** *I have never met/seen him.* **Tóhaŋni waŋčhíyaŋke šni.** *I have never met you.* **3** to look at smth/sb, to take a look at. *USAGE:* only in commands, comp. with **ayúta**. • **Lé taŋyáŋ waŋyáŋka yo.** *Take a good look at this.* **Kȟolá, waŋmáyaŋka po.** *Friends, look at me.* **Waŋúŋyaŋka po.** *Look at us.* **Waŋwíčhayaŋka po.** *Look at them.* ► *1s:* **waŋbláke**. *2s:* **waŋláke**. *1p:* **waŋúŋyaŋkapi**. *CONT:* **waŋyáŋg**; *POS:* **waŋglákA, waŋkíglakA**; *REFL:* **waŋíglakA, waŋíč'iglakA**; *DAT1:* **waŋkíyaŋkA**; *DAT2:* **waŋkíčiyaŋkA**; *RECIP:* **waŋkíčhiyaŋkA**; *RECIP-POS:* **waŋkíčhiglakA**; *CAUS1:* **waŋyáŋgkhiyA**; *ABS:* **wawáŋyaŋkA**; *DIRECTIONAL:* **éwaŋyaŋkA, ahíwaŋyaŋkA, aglíwaŋyaŋkA, akhíwaŋyaŋkA**; *LOC1:* **awáŋyaŋkA**; *LOC2:* **iwáŋyaŋkA**; *LOC3.RECIP:* **íčhiwaŋyaŋkA**; *N:* **wówaŋyaŋke**; *COMP. WITH:* **ayúta, tuŋwÁŋ**; *SV:* **waŋyáŋgphiča, waŋgyáŋg-phičašniyaŋ, oówaŋyaŋg wašté, oówaŋyaŋg šíčA**; *INCOR:* **thi-wáŋyaŋkA**.

FIGURE 1 Lakota entry

see *V* **1a waŋyáŋkA** *AV:2* to see smth/sb. ► *CONT:* **waŋyáŋg** seeing smth/sb; *REFL:* **waŋíč'iglakA, waŋíglakA** see oneself; *POS:* **waŋglákA** see one's own; *RECIP:* **waŋkíčhiyaŋkA** see each other; *DAT2:* **waŋkíčiyaŋkA** see smth for sb, see what belongs to sb; *DAT1:* **waŋkíyaŋkA** to see or look at smth belonging to sb; *CAUS:* **waŋyáŋgkhiyA** cause sb to see smth/sb; *ABS:* **wawáŋyaŋkA** see things or people, observe; be a spectator; *SV:* **waŋyáŋgphiča** it is worth seeing, it can be seen; *CONV (NEG):* **waŋyáŋkešniyaŋ** without seeing smth/sb; *DIRECTIONAL:* **éwaŋyaŋkA, ahíwaŋyaŋkA, aglíwaŋyaŋkA, akhíwaŋyaŋkA**. **1b waŋglákA** *AV:2 (POS)* see one's own. ► *VARIANT:* **waŋkíglakA**; *CONT:* **waŋglág** seeing one's own; *RECIP-POS:* **waŋkíčhiglakA** see each other (of relatives); *CAUS:* **waŋglágkhiyA** let sb see his/her own; *ABS:* **wawáŋglakA** see one's own things; *CONV (NEG):* **owáŋglakešniyaŋ**. **2 tuŋwÁŋ** *AV:1* to open and use one's eyes (be able to see, to not be blind). ► *REDUP:* **tuŋwáŋwaŋ**; *CAUS:* **tuŋwékhiyA**; *DIRECTIONAL:* **ahítuŋwAŋ, étuŋwAŋ**; *LOC:* **atúŋwAŋ, otúŋwAŋ**. **3 thóéyaš** *INTERJ* let me see, give me a second. **4 tókša akhé** *PH* 'see you later/eventually'. *USAGE:* an informal farewell. • **Tókša akhé waŋčhíyaŋkiŋ kte.** *I will see you eventually.* **5 wablésyA** *AV:1 (CAUS ABS)* to see well, have a clear sight. **6 akhíȟpeyA** *AV:2 (CAUS)* to see sb off home or somewhere, drop sb off (for this meaning see also entry for 'take'). **7 ékičituŋwAŋ** *AV:3 (DAT2)* to see to smth for sb. **8 tókhel éš** *INTERJ* let me see (used when one can't recall smth). **9 ehánk'uŋ** *ADV* I see! Aha! Indeed! **10 ištíwaŋyaŋkA** *AV:2* to lay one's eyes on smth/sb.

FIGURE 2 English entry

8 Conjugation in the print and electronic versions

Lakota verbs fall into several classes, each with a different set of personal affixes that can be prefixed or infixed in positions that are often difficult to predict. Knowing the verb forms of the 1st person singular and the 1st person plural is usually sufficient for deducing the remaining

inflection forms, although sometimes the 2nd person singular is also needed, which is why all verb entries include these two to three verb forms. After finding these forms, users can navigate to a chart at the back of the dictionary to identify the matching conjugation paradigm. In verbs with valency 2 or higher, the various combinations of subject and object affixes create 32 distinct verb forms, each expressing a different subject-object relationship. There are several distinct paradigms of these affix combinations, all of which are explained in the grammar section of the dictionary, but identifying the appropriate one is not always trivial due to various morphophonemic changes. Navigating to the paradigm chart and identifying the appropriate paradigm can be time-consuming, particularly for less experienced learners.

The accompanying electronic version of the dictionary simplifies this process; each verb entry includes a link to a full inflection chart showing all forms (7 for intransitive verbs and 32 for verbs with valency 2 or higher). An example of the inflection chart is in Figure 3. Note that affixes are color-coded: Subject affixes are shown in red, object affixes in blue, and the ablauting vowel is highlighted in green (spelled with a capital in the head word). This feature of the application makes complex verbal morphology substantially easier to navigate. Furthermore, the lemmatizer ensures that looking up any inflected form leads the user to the relevant entry (it reliably parses the documented regular and irregular verb forms).

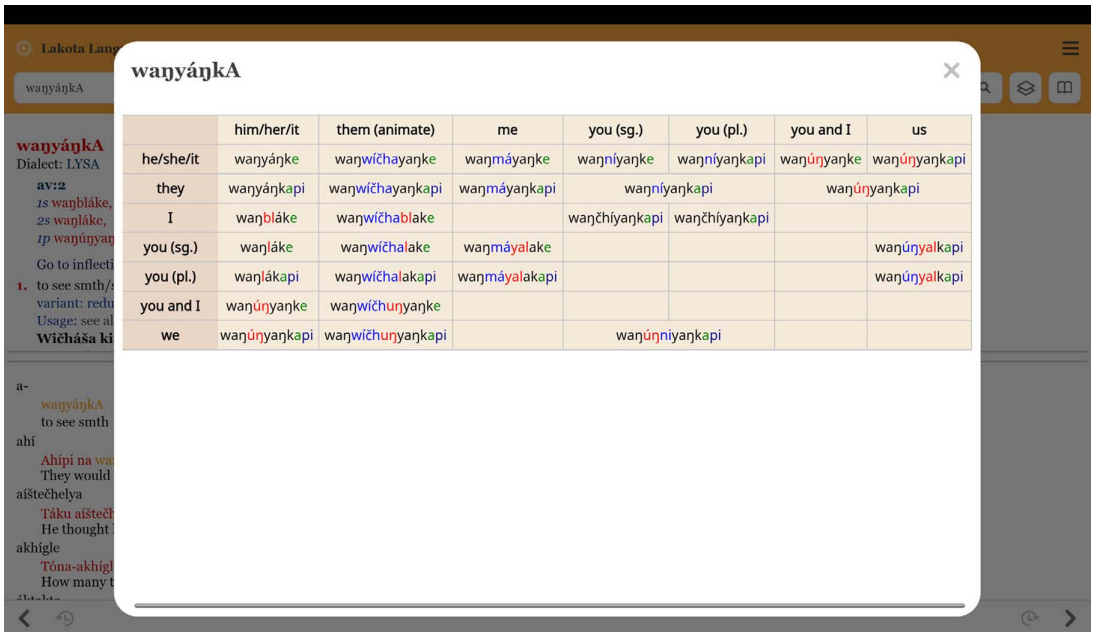


FIGURE 3 Complete conjugation paradigm accessed from an entry in the electronic version

9 Toward a monolingual Lakota dictionary

To further support language acquisition, we began adding Lakota-language definitions alongside the English glosses in the lexical database. As of now, only about 7,000 entries contain Lakota definitions; however, since these are among the most frequently used words, dictionary users are likely to encounter them. The electronic version allows users to display either English glosses or Lakota definitions, or both simultaneously, enabling a transition from bilingual to monolingual consultation. User feedback has been overwhelmingly positive, particularly because Lakota definitions encourage learners to acquire meanings directly through the target language rather than exclusively through English translation.

Including both English glosses and Lakota definitions would be impractical in the printed dictionary, but the electronic version provides a practical platform for progressively expanding monolingual lexicographic resources. A fully monolingual Lakota dictionary has therefore become a realistic long-term objective.

10 Examples, corpus access, and speaker attribution

The print version of the dictionary contains around 50,000 example sentences and collocations, each chosen to illustrate the meaning or usage of a specific entry or sense. The electronic version provides additional full-text access to a corpus of approximately 100,000 sentences (continuously expanded). The lemmatizer in the electronic version allows users to find examples regardless of the inflected form entered. Corpus search enables comparing the use of a lexical item across multiple authentic contexts.

All examples originate in authentic texts representing a wide range of genres, including narratives of personal experience, letters, traditional stories, trickster tales, legends, accounts of historical events, descriptions of cultural and religious practices, and everyday conversations. Consequently, the example sentences document not only lexical meaning and grammatical usage but also traditional knowledge, historical experience, and cultural heritage embedded **within** authentic language use.

Each example sentence, whether in an entry or among corpus search results, is followed by a citation tag indicating the native speaker who produced it (see Figures 4 and 5). Clicking the tag opens a pop-up window with the speaker's full name and additional information, such as their home community, reservation, and the dates they were interviewed or wrote their text. This allows users to identify the provenance of each example and preserves the connection between the lexical data and the individual speakers from whose texts the examples were drawn.

11 Audio component

Around the time of the second print edition in 2011, we sought to create studio-quality audio recordings of native speakers pronouncing all entry words. Funding, however, was the main obstacle to accomplishing this goal. Over several consecutive years, we applied for National Science Foundation (NSF) grants without success, receiving reviewer comments that studio recordings of isolated lexical items were not considered linguistic research because they were produced outside natural communicative settings. This illustrated the divergent criteria between traditional academic funding bodies and the specific requirements of community-based documentation projects for language revitalization and education. We eventually found funding for the project from other sources and recorded around 24,000 entry words from both male and female native speakers. Interestingly, once the recordings became publicly available through the app, numerous linguists requested access for research purposes, demonstrating that resources developed primarily for language revitalization can also serve as valuable research data. In addition to the audio for entry words, users can listen to authentic audio recordings of about 30,000 sentences in the corpus and several hundred sentences in entries. The recordings serve both as a pronunciation resource and as a permanent archive of native-speaker speech.

Ben Black Bear, Jr., and Iris Eagle Chasing provided recordings for the majority of the lexical entries, although Bernadine Little Thunder, Kenny Little Thunder, and Manny Iron Hawk also recorded a smaller number of words. Given the cost of the project, it is unlikely that audio recordings will be produced for the lexical entries added since then unless additional funding becomes available. Future development may incorporate speech synthesis (TTS), whose

quality has improved substantially in recent years. This would make it possible to offer the audio version not only for entry words but also for all inflected forms.

12 Electronic app

After the second edition of the dictionary was published in 2011, the native board of directors of the organization expressed their wish to publish the dictionary as an electronic multimedia application as well. After a considerable but fruitless effort to find IT developers for this task in the community, the decision was made to partner with a company that has enough experience and track record in developing dictionary applications, especially given the complexity and number of the desired features, as the board members wanted the dictionary application to match the quality of apps used for highly developed dictionaries of modern languages, such as English.

The board also emphasized that time is of the essence and that Lakota education could not wait several years for an application to be developed. Consequently, we partnered with a company called Lingea, which has a long track record of developing dictionary applications for over 40 languages, and whose data and language tools are used by companies such as Adobe, Microsoft, or Oxford University Press. Lingea was able to cater to our needs regarding database format, entry architecture, and lemmatization strategy, and their application offers a modern, user-friendly interface. Lingea also implemented inflection charts for all verb entries, using the data we created and following the format we designed. To improve accessibility, the application also accepts searches entered in the simplified missionary orthography and automatically redirects users to the corresponding entries written in the phonemic orthography. This allows users accustomed to the earlier writing system, as well as learners who have not yet mastered the phonemic orthography, to locate entries without having to manually convert spellings.

The NLD is currently available on all major platforms, including Android, iPhone/iOS, and Windows. Figures 4-6 show screenshots of a sample Lakota entry, a Lakota entry with corpus search, and an English entry, respectively.

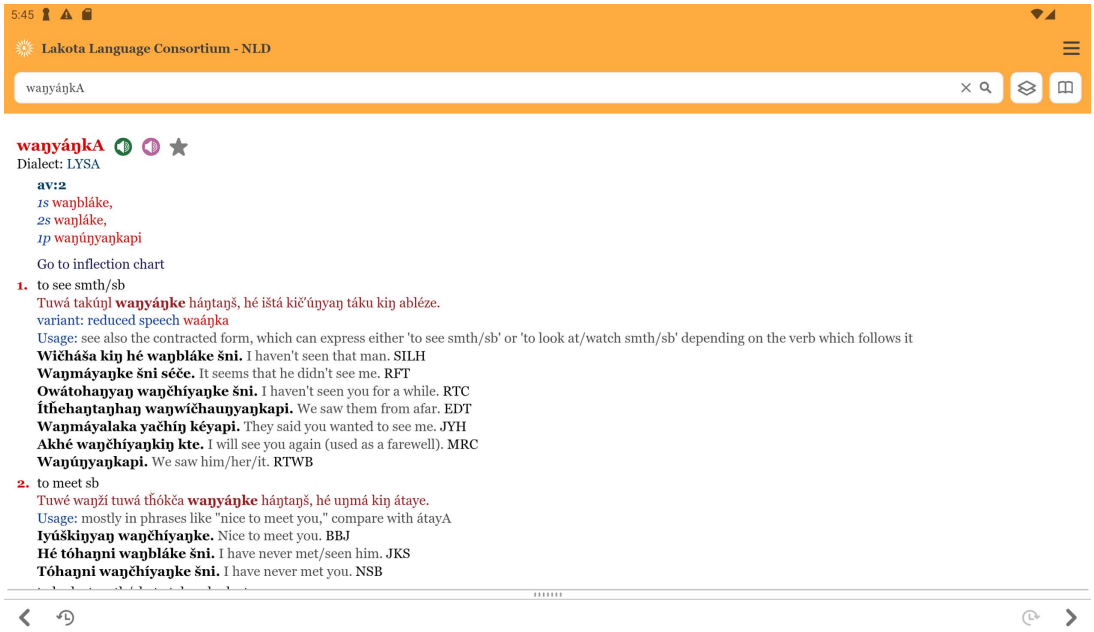


FIGURE 4 Lakota entry in the application

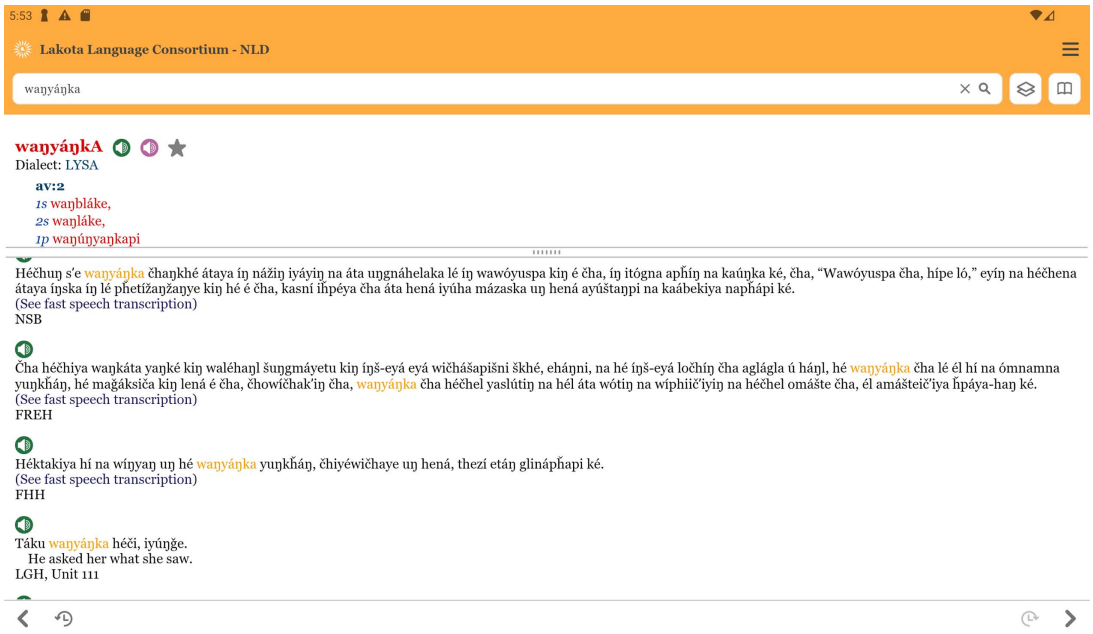


FIGURE 5 Lakota entry with corpus search results

13 Lexicographic software

Although the software tools described below are internal and not visible to dictionary users, they are essential for editing and maintaining the lexical database, integrating corpus evidence, and generating many of the features available in the electronic dictionary.

When the project began in 1992, we chose the Linguistic Toolbox (formerly Shoebox), developed by SIL Global. As a freely available application, it was a natural choice for a project with no initial funding. Throughout the project, we evaluated other lexicographic tools, especially FLEx and TshwaneLex (see de Schryver & Joffe, 2004 and de Schryver, 2003, on the design and use of dictionary writing systems such as TshwaneLex). Although each offered particular strengths, none adequately supported the combination of large-scale lexicographic editing, corpus integration, editing speed, and the entry architecture required for our project. Consequently, we continued to maintain the lexical database in Toolbox while developing complementary software to provide the missing functionality.

Around 1996, IT specialist Konstantine Chmielnicky joined the project as a volunteer. His work eventually led to the development of Virtual Toolbox (VTB), an internal software utility that complements Toolbox by automating corpus-related tasks and advanced language processing. VTB generates comprehensive lists of inflected forms for lemmatization and corpus indexing. From approximately 48,000 lemmas, it derives around 900,000 lexical forms used in the inflection charts and recognizes hundreds of millions of other word forms generated by combinations of affixes and ablaut, enabling efficient lemmatization in both the lexical database and electronic applications.

We also use VTB for corpus annotation and analysis, including phonemicization of older texts, searches for collocations, concordances, multi-word expressions, and morphosyntactic constructions, automated assignment of corpus citations to dictionary entries, and generation of corpus statistics, including frequencies, geographic and temporal spans of documented words, etc. We continue to improve its morphoparsing capabilities, thereby enhancing

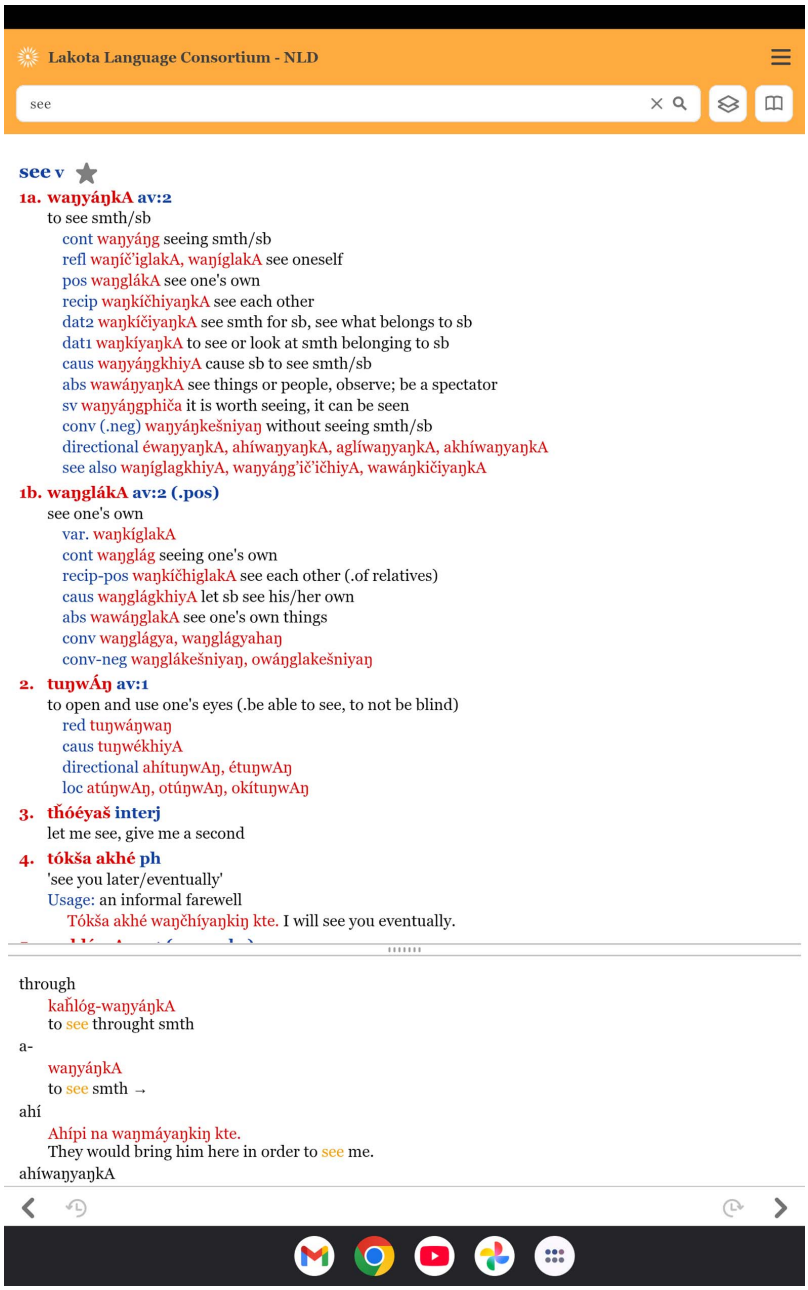


FIGURE 6 English entry

homograph disambiguation and the accuracy of corpus analysis. Work is also underway on a morphosyntactic parser and syntactic treebank, although these remain under development.

14 Conclusions

With its 48,000 entries based on a text corpus of several million words and spanning more than 155 years and 3,500 native speakers, the New Lakota Dictionary is not simply a large dictionary. It represents an integrated model of endangered-language lexicography in which corpus documentation, linguistic research, community collaboration, and digital technology mutually reinforce one another.

The project required revisiting established grammatical categories for Lakota — including the analysis of adverbs, converbs, derived modifiers, postpositions, coverbs, and dative verbs, and the reanalysis of post-nominal stative verbs as copredicates rather than adjective-like modifiers — demonstrating that large-scale lexicographic work can itself generate new grammatical findings, not merely apply existing ones.

This project resulted in both a traditional bilingual print dictionary and a comprehensive multimedia digital resource that integrates bilingual and monolingual dictionaries, full-text search in a large text corpus, studio-recorded pronunciation of words in isolation, and recordings of sentences from authentic discourse, a lemmatizer, and comprehensive inflection charts for all verbs. It is predominantly corpus-based but supplements missing lexical data with carefully vetted input from other sources. Its high accuracy stems from careful analysis of corpus data, extensive consultation with native speakers, grammatical analysis, and effective use of computational tools.

By documenting authentic connected speech and making it accessible to educators and learners, the dictionary has become an important resource for Lakota language revitalization. This is reflected in its widespread acceptance among Lakota language educators, learners, and institutions, where it serves as a cornerstone of Lakota language education and revitalization, both of which require reliable, comprehensive reference materials. The dictionary has also become a standard reference for linguistic research on Lakota, a role acknowledged by leading Siouanists, including Robert Van Valin, who described it in an email sent to our team as *"the best dictionary of any language that I have ever seen . . . and the primary reference source for the language"* (quoted with permission).

The project demonstrates that, with sustained collaboration, corpus-based methodology, rigorous linguistic analysis, and appropriate computational tools, lexicography for endangered, underdocumented, and under-resourced languages can yield a dictionary that meets contemporary lexicographic standards in terms of content quantity, quality, layout, and user-friendliness. Many of the methodological principles of working with this polysynthetic language are relevant to other morphologically complex languages, and as such, this is a case study within a recognized methodological challenge.

We do not consider our work finished. We plan to create additional corpus data, digitize and analyze more of the existing corpus, and continue documenting the language from native speakers for as long as possible. More work is needed to improve our computational tool's ability to serve future editions of the print version and the electronic app.

The following quote is from an introductory letter published in the first edition of the dictionary and co-authored by the highly respected Lakota elders and scholars Johnson Holy Rock, Ben Black Bear, and Delores Taken Alive:

> We are happy that this New Lakota Dictionary has come to our people. When we pray to the Creator, we speak in the Lakota language, and that is why our language is sacred. . . . If we use it well, we will unite with all of our relatives, and we will honor the memory of our ancestors. With this dictionary, we have come back in full circle to where we can once again stand as a people.

The New Lakota Dictionary is the product of many Lakota speakers who shared their language through texts, recordings, conversations, and decades of collaboration with the editorial team. Their voices remain at the heart of the dictionary, ensuring that the linguistic and cultural knowledge they shared will remain accessible to future generations.

References

Ashford, M. K. (1987). The Evolution of the Hawaiian Dictionary and Notes on the Early Compilers, with Particular Attention to Manuscript Resources of the Bishop Museum Library. *Bishop Museum Occasional Papers* 27.

Black Bear, B. Jr., Two Dogs, R. & Ullrich, J. (2020). *New Lakota Dictionary (electronic app version)*. Lakota Language Consortium.

Bendickson, J. (2026). *Dakhóta Iápi Wičhóie Wówapi*. A presentation at the 46th Annual Siouan and Caddoan Languages Conference.

Boas, F. & Deloria, E.C. (1941). *Dakota Grammar*. Memoirs of the National Academy of Sciences, Volume XXIII, second memoir. US Govt. Printing Office.

Bray, D. (Ed.) (1998). *Western Apache-English Dictionary: A Community-Generated Bilingual Dictionary*. Bilingual Press/Editorial Bilingüe.

Bright, W. (1999). Note. *Language in Society*, 28, 481–483. doi:10.1017/S0047404599213073.

Buechel, E. (1939). *A Grammar of Lakota. The Language of the Teton Sioux Indians*. Rosebud Educational Society.

Buechel, E. (1970). *A Dictionary of the Teton Dakota Sioux language. Lakota-English: English - Lakota*, ed. by Paul Manhart, Red Cloud Indian School Inc.

Buechel, E. (2002) *Lakota Dictionary: Lakota-English/English-Lakota. New comprehensive ed*. University of Nebraska Press.

Byington, C. (1915). *A Dictionary of the Choctaw Language. Edited by John R. Swanton and Henry S. Halbert*. Bureau of American Ethnology Bulletin 46. Government Printing Office.

Chelliah, S. L. & de Reuse, W. J. (2011). *Handbook of Descriptive Linguistic Fieldwork*. Springer.

Corris, M., Manning, C., Poetsch, S. & Simpson, J. (2002). *Dictionaries and endangered languages*. In D. Bradley & M. Bradley (Eds.), *Language Endangerment and Language Maintenance* (pp. 329–347). RoutledgeCurzon.

Cowell, A. (2026). *Arapaho Lexical Database and Dictionary*. The Arapaho Project, University of Colorado Boulder. Retrieved July 23, 2026, from www.colorado.edu/project/arapaho/language/dictionary

Crow Language Consortium (2022). *Crow Dictionary*.

Cumberland, L. A. & Rankin, R. L. (Eds.) (2012). Kaá^{n}ze *Íe Wayáje: An Annotated Dictionary of Kaw (Kanza): Kaw-English/English-Kaw*. Kanza Language Project of the Kaw Nation.

Dakhóta Iápi Okhódakičhiye (2025). *Dakota-English and English Dakota Dictionary* (an electronic app).

de Schryver, G.-M. (Ed.) (2010). *Oxford Bilingual School Dictionary: Zulu and English*. Oxford University Press Southern Africa.

de Schryver, G.-M. (2003). Lexicographers' dreams in the electronic-dictionary age. *International Journal of Lexicography*, 16(2), 143–199.

de Schryver, G.-M., & Joffe, D. (2004). On how electronic dictionaries are really used. In *Proceedings of the Eleventh EURALEX International Congress*.

Feeling, D. (1975). *Cherokee–English Dictionary*. Cherokee Nation

Fisher, L., Leman, W., Pine, L., Sr. & Sanchez, M. (2013). *Cheyenne dictionary*. Lulu Press.

Galloway, B.D. (2009). *Dictionary of Upriver Halkomelem*. University of California Press.

Hanks, P. (2012). The Corpus Revolution in Lexicography. *International Journal of Lexicography*, 25(4), 398–436.

Helmbrecht, J. & Lehmann, C. (2010). *Hocak Teaching Materials: Elements of Grammar/Learner's Dictionary*. State University of New York Press.

Holy Rock, J., Black Bear, B. Jr., & Taken Alive, D. (2008). Introductory letter. In J. Ullrich (Ed.), *New Lakota Dictionary*. Lakota Language Consortium.

itwêwina Plains Cree Dictionary. Retrieved from https://itwewina.altlab.app/

Jacobson, S. A. (2012). *Yup'ik Eskimo dictionary*. Alaska Native Language Center, University of Alaska Fairbanks.

Jetté, J. & Jones, E. (2000). *Koyukon Athabaskan Dictionary*. Alaska Native Language Center, University of Alaska Fairbanks.

Michelson, K. & Doxtator, M. (2002). *Oneida-English/English-Oneida Dictionary*. University of Toronto Press.

Liljeblad, S., Fowler, C. S. & Powell, G. (2011) *Northern Paiute-Bannock Dictionary*. University of Utah Press.

Martin, J. B. & Mauldin, M. M. (2000). *A Dictionary of Creek/Muskogee*. University of Nebraska Press.

Masthay, C. (Ed.) (2002). *Kaskaskia Illinois-to-French Dictionary*. Carl Masthay.

Montler, T. (2012). *Klallam Dictionary*. University of Washington Press.

Montler, T. (2018). *SENĆOŦEN: A Dictionary of the Saanich Language*. University of Washington Press.

Mosel, U. (2004). Dictionary making in endangered speech communities. In P. K. Austin (Ed.), *Language Documentation and Description*, Vol. 2 (pp. 39–54). School of Oriental and African Studies (SOAS).

Mosel, U. (2011). Lexicography in endangered language communities. In P. Austin & J. Sallabank (Eds.), *The Cambridge Handbook of Endangered Languages* (pp. 337–353). Cambridge University Press.

Mosel, U. (2012). Morphosyntactic analysis in the field: A guide to the guides. In N. Thieberger (Ed.), *The Oxford Handbook of Linguistic Fieldwork* (pp. 72–89). Oxford University Press.

Parks, D.R. & Pratt, L.N. (2008) *A Dictionary of Skiri Pawnee*. University of Nebraska

Pukui, M.K. & S.H. Elbert, (1986). *Hawaiian dictionary: Hawaiian-English/English-Hawaiian. Revised and enlarged edition*. University of Hawaii Press.

Pustet, R. (2000). Lakota Postpositions. *International Journal of American Linguistics*, 66(2), 157–80.

Riggs, S.R. (1852). *Grammar and dictionary of the Dakota language*. Smithsonian Contributions to Knowledge.

Riggs, S.R. (1992 [1890, 1952]). *A Dakota-English dictionary*. Minnesota Historical Society Press.

Young, R.W., Morgan, Wm. Sr. & Midgette, S. (1992) *Analytical Dictionary of Navajo*. University of New Mexico Press.

Rood, D. S. & Taylor, A. R. (1996). Sketch of Lakhota, a Siouan Language. In Goddard, I. (Ed.) *Language. Handbook of North American Indians*, vol. 17. (pp. 440–482). Washington: Smithsonian Institution.

Rood, D. S., & Taylor, A. R. (1976). *Beginning Lakhota*. University of Colorado Lakota Project.

Rudes, Blair A. (1999). *Tuscarora-English/English-Tuscarora Dictionary*. University of Toronto Press.

Sylestine, C., Hardy, H. K. & Montler, T. (1993). *Alabama-English Dictionary*.

The Hopi Dictionary Project (1998). *Hopi dictionary/Hopìikwa lavàytutuveni: A Hopi-English dictionary of the Third Mesa dialect, with an English-Hopi finder list and a sketch of Hopi grammar*. The University of Arizona Press.

TlingitLanguage.com (2019). *Tlingit Dictionary*. Compiled by Lance Twitchell, James Crippen, Keri Eggleston, and Jeff Leer. Retrieved October 12, 2019, from https://tlingitlanguage.com/wp-content/uploads/2019/10/dictionary.pdf

Two Dogs, R. & Ullrich, J. (Eds.) (2025). *Remembering My Ancestors: Lakota Narratives*. Lakota Language Consortium.

Ullrich, J. & Black Bear, B. Jr. (2016). *Lakota Grammar Handbook: A Pedagogically oriented self-study reference and practice book*. Lakota Language Consortium.

Ullrich, J. (2008). *New Lakota Dictionary*, First edition. Lakota Language Consortium.

Ullrich, J. (2011). *New Lakota Dictionary*, Second edition. Lakota Language Consortium.

Ullrich, J. (2013). *Revisiting Lakota Dative Affixes: Analysis of Benefaction and Malefaction in Lakota*. A paper presented at the 33rd Siouan Caddoan Language Conference. Sitting Bull College.

Ullrich, J. (2018). *Modification, Secondary Predication and Multi-Verb Constructions in Lakota*. [Doctoral dissertation, Heinrich Heine University].

Ullrich, J. (2020). Complex Predicates with Nouns and Stative Verbs in Lakota, A Role and Reference Grammar Analysis. *International Journal of American Linguistics*, 86(3), 407–446.

Ullrich, J. (2022a). *Lakota Stories Collected by Ella Deloria*. Lakota Language Consortium.

Ullrich, J. (2022b). Lakota lexical categories, their morphology and syntactic functions. In *New Lakota Dictionary*, Third edition. Lakota Language Consortium.

Ullrich, J. (2022c). *New Lakota Dictionary*, Third edition. Lakota Language Consortium.

Ullrich, J. (2026a). Lakota Postpositions and Coverbs with Postpositional Semantics: A Role and Reference Grammar Analysis. In D. Bentley, L. Guerrero, M. Shimojo, & J. Fleischhauer (Eds.), *Role and Reference Grammar: New Developments and Perspectives*. De Gruyter Mouton.

Ullrich, J. (2026b). *Assigning Attributive Property Without Ad-nominal Modification in Lakota*. (to appear)

Wall, L. & Morgan, W. (1958). *Navajo-English Dictionary*.

Weber, N. et al. (2023). Blackfoot Words: A lexical database of Blackfoot legacy sources. *Language Resources and Evaluation*, 57, 1207–1262. https://doi.org/10.1007/s10579-022-09631-2

Williams H. W. (1971). *A dictionary of the Maori language*. Legislation Direct.

Wood, S. (Ed.) (2026). *Online Nahuatl Dictionary*. Wired Humanities Projects, University of Oregon. Retrieved July 23, 2026, from nahuatl.wired-humanities.org/

Contact information

Ben Black Bear, Jr.
Lakota Language Consortium
ben.blackbear.jr@lakhota.org

Richard Two Dogs, muŋyáŋ Máni
Lakota Language Consortium
r.twodogs@lakhota.org

Jan Ullrich
Lakota Language Consortium
jfu@lakhota.org

Ivana Brač

SEMANTIC CLASS ASSIGNMENT IN A CROATIAN VERB LEXICON

Challenges of Applying VerbNet's Classes

Abstract This paper examines the challenges of assigning VerbNet semantic classes to Croatian verb senses, based on data from the Croatian verb lexicon Verbion. The lexicon provides a multi-level description of the 500 most frequent verbs in Croatian, integrating lexical, semantic, syntactic, and morphological information. Verb senses are classified into semantic classes according to WordNet and VerbNet, using English equivalents as an intermediary. Since VerbNet is based on Levin's syntactic-semantic classification of English verbs, applying these classes to Croatian presents certain challenges. The analysis identifies five types of cross-linguistic mismatch. First, differences in lexicalization patterns reflect language-specific conventions governing what semantic components are encoded in the verb itself as opposed to what is expressed syntactically. Second, Croatian reflexive morphology frequently introduces semantic distinctions that have no direct counterpart in VerbNet classes. Third, alternation-based classification cannot always be applied, as such alternations are either restricted or encoded differently in Croatian. Fourth, some VerbNet classes do not adequately capture the syntactic and semantic properties of Croatian verbs. Finally, certain high-frequency verbs are absent from the VerbNet inventory altogether. These findings suggest that VerbNet requires systematic adaptation for Croatian. The study contributes to the development of a Croatian-specific extension while preserving VerbNet's cross-linguistic comparative potential.

Keywords Verb lexicon; semantic classes; VerbNet; WordNet; Croatian

1 Introduction

Verbs have been thoroughly examined across various linguistic resources, as they are traditionally considered the core element of a sentence. When determining the criteria for processing verbs, different principles may be applied, ranging from morphological ones (classification by conjugation classes) to purely semantic ones (e.g., WordNet, Fellbaum, 1998), or a combination of syntactic and semantic criteria (e.g., VerbNet, Kipper-Schuler, 2005). Assigning a semantic class to verb senses is common practice across linguistic resources employing different methodologies and theoretical frameworks. For example, in WordNet (Fellbaum, 1998), which emphasizes verb semantics, verbs are divided into stative verbs and action verbs, which are further classified into 14 semantic fields. In VerbNet (Kipper, Dang, & Palmer, 2000; Kipper-Schuler, 2005), which deals with the semantics and syntax of verbs, there are more than 270 semantic classes and subclasses. VerbNet's classification is based on Levin's (1993) classification, which relies on syntactic alternations, with certain modifications (Dorr, 1997; Korhonen & Briscoe, 2004; Kipper et al., 2008). As Levin (1993) primarily analysed verbs whose arguments are noun phrases or prepositional phrases, while adjectival, adverbial, predicative, clausal and other arguments were not considered, these limitations were addressed in extensions of VerbNet (Kipper et al., 2008, p. 24). With the expansion of VerbNet, 26 new classes were initially introduced (Dorr, 1997), followed by 55 (Korhonen & Briscoe, 2004), and an additional 35 (cf. Kipper et al., 2008), increasing the number of classes from the original 191 to 274. Beyond English, Levin's classification has served as the foundation for the Czech valency lexicon VALLEX (e.g., Lopatková et al., 2022), the Croatian CROVALLEX (Mikelić Preradović, 2008, 2020), and e-Glava (Birtić, Brač & Runjaić, 2017), reflecting its appeal as a cross-linguistically applicable framework despite the language-specific assumptions it carries (cf. Majewska et al., 2018). FrameNet's (Ruppenhofer et al., 2016) hierarchically organized semantic frames can be regarded as highly elaborated semantic classes that capture fine-grained distinctions

between verb senses. The advantage of classification based on semantic classes lies in the ability to observe similarities and differences between semantically related verbs, as well as their potential for syntactic alternations. However, applying such classifications across languages is not straightforward, as they are often based on language-specific interactions between syntax and semantics.

This paper addresses the challenges of assigning semantic classes in the Croatian verb lexicon Verbion. The aim is to identify and systematize the main types of cross-linguistic mismatch that emerge when VerbNet classes are applied to Croatian verb senses. This will serve as a basis for a future Croatian-specific adaptation of the VerbNet classes, including a revised class inventory and the introduction of Croatian-specific subclasses where necessary.

The paper is organized as follows. In Section 2, the structure of the database is outlined, along with the selection process for the verb list. It also provides general information about the verbs presented at the first level of description, the contents of the semantic level of description, and the structure of the valency frame, in comparison with the consulted resources. Section 3 describes the methodology for assigning semantic classes. In Section 4, the main types of mismatch that arise when applying VerbNet's semantic classes to Croatian verbs are identified and analysed. Section 5 draws conclusions, summarizes the main findings, and discusses the implications for the cross-linguistic adaptation of VerbNet, as well as directions for future research.

2 Croatian verb lexicon

Several Croatian valency resources have been developed since the early 2000s, including CROVALLEX (Mikelić Preradović, 2008, 2020), e-Glava (Birtić, Brač & Runjaić, 2017), and CroaTPAS (Marini & Ježek, 2019), each offering valuable insights into verb semantics and argument structure. These resources differ with respect to their theoretical framework, level of descriptive detail, and scope, ranging from comprehensive valency descriptions to more focused analyses of selected verb classes. The verb lexicon Verbion builds on these resources while aiming to integrate their different descriptive strands into a unified framework. The following subsections outline the structure of the database and its main descriptive levels.

Verbion is a Croatian verb lexicon that will become publicly accessible by the end of 2027. In the initial phase of the project, a detailed multi-level description will be provided for the 500 most frequent verbs, with plans to expand the lexicon to 2,000 verbs. The verb list was compiled by comparing the 1,000 most frequent verbs extracted from corpora containing more than 100 million tokens.[1] The verbs were then preliminarily classified into semantic classes, following the methodology outlined in Brač & Bošnjak Botica (2015), to facilitate further processing. The metalanguage of the database is English, and the lexicon also includes English translations of key information, with the aim of making the database suitable for learning Croatian as a second language and for facilitating potential integration with other linguistic resources.

In the database, aspectual pairs are treated as separate lemmas, in line with CROVALLEX (Mikelić Preradović, 2008, 2020), Wałenty (Przepiórkowski et al., 2014), and e-Glava (Birtić, Brač & Runjaić, 2017; Birtić, 2020). Furthermore, reflexive verbs are also treated as separate lemmas, rather than as one of a verb's senses or a sublemma. This approach differs from CROVALLEX and e-Glava but aligns with VALLEX 2.0 and later versions, which offer a more

[1] CLASSLA (Ljubešić & Kuzman, 2024), MaCoCu (Bañón et al., 2023), hrWac (Ljubešić & Klubička, 2014), ELEXIS (Jakubíček et al., 2022), OpenSubtitles (Tiedemann, 2016), EUR-Lex (Baisa et al., 2016), HNK (Tadić, 2002).

advanced description of reflexive verbs. This decision is motivated by both syntactic and semantic considerations. The reflexive marker *se* cannot be considered a complement, and reflexive verbs differ semantically and syntactically from their (in)transitive counterparts (cf. Kettnerová, Lopatková & Vernerová, 2022). In VALLEX, reflexive verbs are distinguished into reflexiva tantum and derived reflexive verbs, with the latter further categorized into seven classes (Kettnerová, Lopatková & Vernerová, 2022, p. 40). In the present lexicon, reflexive verbs are classified into three groups: reflexiva tantum (e.g. *bojati se* 'to be afraid'), derived reflexives (e.g. *buditi se* 'to wake up'), and reciprocal reflexives (e.g. *ljubiti se* 'kiss each other'), without further subcategorization. Nonetheless, this categorization still represents an enrichment of Croatian linguistic resources, as it has not yet been included in any existing database.

2.1 Lemma level

In the database, the first level of description provides general information about a verb, including the verb lemma, its aspect (perfective, imperfective, or biaspectual), its aspectual pair (if applicable), a reflexivity label, and metadata about a curator, status, and date. Verbal idioms are related to a verb lemma rather than being treated as one of the verb senses or as a sublemma, which differs from the approach adopted in VALLEX (Lopatková et al., 2022) and CROVALLEX (Mikelić Preradović, 2008). A similar solution can be found in e-Glava (Birtić, Brač & Runjaić, 2017), where each idiom is accompanied by its definition and an example from the corpus. In Verbion, the same categories are retained, with the addition of an English translation for both the idiom and its example. The first level also contains a link to a morphological block containing different tenses and modes, and to verb senses, described in the following subsection.

2.2 Sense level

The second level is dedicated to individual verb senses. It includes a definition, its translation into English, synonyms, English equivalents, semantic classes, a semantic frame, and a link to valency frames. Each verb sense is associated with a semantic class from VerbNet (Kipper et al., 2008) and WordNet (Fellbaum, 1998), which is explained in Section 4. Once a representative sample has been processed, a classification based specifically on Croatian verbs is proposed. Definitions are established drawing on existing online dictionaries (*Hrvatski jezični portal*; *Školski rječnik hrvatskoga jezika*) and validated through corpus analysis, primarily using hrWac (Ljubešić & Klubička, 2014) and CLASSLA (Ljubešić & Kuzman, 2024). This corpus-based approach often leads to the identification of additional senses not recorded in dictionaries (more details in Section 3). Semantic frames are incorporated to integrate insights from FrameNet (cf. Kettnerová, Lopatková & Bejček, 2012). Each verb sense is manually assigned a frame, with links provided to the corresponding frames in Berkeley FrameNet (Ruppenhofer et al., 2016) and the Croatian FrameNet, which is currently under development.

2.3 Valency frame level

A valency frame is related to a verb sense, and a single sense may contain one or more valency frames. The valency frame consists of semantic roles, syntactic phrase types, their morphological realizations, and a corpus example, primarily drawn from hrWac (Ljubešić & Klubička, 2014). All examples are translated into English, and each participant is analysed at the syntactic, morphological, and semantic levels. As a Slavic language, Croatian has a rich morphological system with seven cases, so information about the morphological realization of constituents is

systematically included, following the practice in VALLEX (Lopatková et al., 2022), Walenty (Przepiórkowski et al., 2014), CROVALLEX (Mikelić Preradović, 2020), e-Glava (Birtić, Brač & Runjaić, 2017), and CroaTPAS (Marini & Ježek, 2019). Valency frames are retrieved using a random sample of 2,000 sentences from hrWac (Ljubešić & Klubička, 2014) in Sketch Engine (Kilgarriff et al., 2014), while more targeted searches are conducted using Word Sketches and regular expressions. For syntactic annotation, Verbion adopts syntactic phrase types, as they are widely used across various linguistic theories and are also adopted in VerbNet and Walenty. The verb under analysis is labelled as V, while copular (Vcop), semicopular (Vsem_cop), reflexive (VREFL_tantum, VREFL_derived, VREFL_recipr), and light verbs (LV) are further specified.

Other constituents are labelled as follows: NP (noun phrase) and AP (adjective phrase), with one of five morphological cases that can appear without a preposition (nominative, genitive, dative, accusative, instrumental); PP (prepositional phrase), with a preposition-case combination selected from a drop-down menu; AdvP (adverbial phrase), CP (complementizer phrase), where the introducing conjunction is specified or marked as *0* in cases of juxtaposition; and ConjP (conjunction phrase), used for reciprocal reflexive verbs. Since many of the 500 most frequent verbs belong to the class of verbs of communication, the label QUOT was introduced for direct speech or quotation, following the same approach as in e-Glava (*navod*), Walenty (*or*), and FrameNet (*QUO*).

Regarding semantic annotation, a unified approach was adopted to the greatest extent possible given the lack of consensus on semantic role inventories. VerbNet was chosen for its comprehensive coverage of semantic roles and its alignment with role assignment in theoretical linguistics, particularly within the framework of Role and Reference Grammar (Van Valin & LaPolla, 1997; Van Valin, 2000). However, some modifications were necessary. Among others, new semantic roles were introduced for participants representing entities like organizations, cities, etc., through metonymy (WHOLE FOR PART or PART FOR WHOLE), such as Agent_METO, Co-Agent_METO, Recipient_METO, Goal_METO. Furthermore, the analysis of specific verb classes, such as communication verbs, indicated the need for more fine-grained roles than those available in VerbNet. For this reason, roles such as Medium (for the channel of communication) and Mediator (for a participant speaking on behalf of another) were introduced. The role Medium is also attested in FrameNet as a frame element in communication-related frames. The inventory of semantic roles will be further refined as needed. Verbion differs from other resources in that a single participant in an event may be assigned multiple semantic roles, in contrast to the assumption that each argument is associated with a single thematic role (Chomsky, 1981). For example, in the sentence *"Policija traga za počiniteljem", kazali su iz policije*. '"The police are searching for the perpetrator," the police said.', the prepositional phrase *iz policije* 'from the police' = 'the police' has the semantic roles of Source and metonymic Agent. This allows for a more nuanced representation of complex event participants.

In this respect, Verbion builds on existing approaches by combining multiple levels of description. It integrates lexical, semantic, syntactic, and morphological information, while systematically linking verb senses to semantic classes and frames. Particular attention is given to the treatment of reflexive verbs at the level of lemma organization and to the annotation of semantic roles in relation to their syntactic and morphological realization. At the same time, the database is designed to align with existing resources such as VerbNet, WordNet, and FrameNet, while accommodating language-specific features of Croatian.

3 Methodology

At the time of analysis, the database contained 263 processed verb entries comprising 678 verb senses, and 2820 Croatian corpus examples with English translations. As described in Section

2.2, senses are determined using available online dictionaries (*Hrvatski jezični portal*; *Školski rječnik hrvatskoga jezika*) and by examining corpus examples.

An additional method involves analysing a sample of 10,000 randomly selected corpus sentences from hrWac (Ljubešić & Klubička, 2014), corresponding to the maximum export size from Sketch Engine, using large language models to suggest potential sense distinctions. Specifically, GPT-4-based models were used in the earlier stages of processing, while more recent entries have been supported by GPT-5.5. The model is prompted with structured instructions to identify candidate senses based on patterns of argument structure and semantic context across the 10,000-sentence sample. The model's output is treated strictly as a heuristic starting point; it is used to flag potential sense distinctions and draw attention to syntactically or semantically divergent uses that might otherwise require extensive manual scanning. All proposed senses are subsequently verified by at least one lexicographer, and no sense is included without manual validation. The use of LLMs in this workflow therefore serves to accelerate the initial exploration phase while leaving all lexicographic decisions to human annotators.

For each Croatian verb sense, an appropriate English equivalent is first established and used to manually assign the sense to the closest corresponding VerbNet class. The selection of the English equivalent is based on dictionary definitions and corpus context, with the aim of identifying the closest functional and semantic correspondence. All verb senses are subsequently classified according to VerbNet and WordNet, specifically the Croatian version of WordNet – CroWN (Filko et al., 2015), which is currently accessible via the BulNet interface (Rizov & Dimitrova, 2016; https://dcl.bas.bg/bulnet/), as no dedicated public interface for CroWN is available. A semantic class is assigned only if it corresponds to the sense of the Croatian verb; if no suitable equivalent is found or the available classes do not adequately capture the sense, no class is assigned, and this is indicated in the notes. In addition, the distribution of assigned and unassigned classes is quantified, while the mismatch types discussed in Section 4 are treated primarily as qualitative analytical categories.

The annotation process follows a multi-stage validation procedure in which each verb is initially analysed by a single lexicographer, after which the analysis is reviewed by one or more additional lexicographers. Disagreements are resolved through discussion among them. Given the nature of this workflow, in which review serves primarily as a consistency check and a means of resolving linguistic uncertainty rather than as independent parallel annotation, formal inter-annotator agreement is not measured. This approach ensures consistency and quality control while maintaining an efficient workflow.

The present analysis is preliminary, reflecting the current stage of the database. As the lexicon expands towards the planned coverage of the 500 most frequent verbs, the classification of mismatch types will be re-examined and the quantitative findings presented here will be updated accordingly.

4 Types of mismatches in applying VerbNet classes to Croatian verbs

This analysis highlights several recurrent issues and outlines adaptation strategies. Since CroWN (Filko et al., 2015) is based on Croatian lexical material and semantic relations, assigning classes within this framework is generally more straightforward. Difficulties mainly arise when a particular sense is not included in the CroWN inventory (e.g., *isticati* 'highlight' in the sense 'to make something visually stand out by using a specific means', *predlagati* 'nominate' in

the sense 'to propose that a person or entity be appointed to a specific position or granted a particular status or award', *pozdraviti se* 'quit, give up' in the sense 'to give up on a set goal or task', etc.) or when a reflexive verb is not recorded at all (e.g., *predstavljati se* 'present oneself, showcase', *spasiti se* 'save oneself', *zvati se* 'to be called, to be named'). As WordNet comprises fewer and more general classes and is based solely on semantic criteria, this paper focuses primarily on the difficulties involved in applying the VerbNet classification. Of the 678 analysed verb senses, 597 (88.1%) could be linked to an existing VerbNet class, although many required adaptations due to cross-linguistic differences between English and Croatian. The remaining 81 senses (11.9%) could not be satisfactorily assigned to any available VerbNet class. The analysis is primarily qualitative and illustrative rather than based on exhaustive quantitative annotation of individual mismatch types. Future work will include a more systematic quantitative annotation of mismatch types across the complete dataset.

VerbNet classification poses greater challenges, since it is rooted in Levin's approach (1993), which groups English verbs into semantic classes based on syntactic alternations, assuming that syntactic behaviour reflects semantics. Cross-linguistic asymmetry is already evident at the level of class inventory: some VerbNet classes are highly productive in English but marginal in Croatian. For example, classes such as *butter-9.9* and *pocket-9.10* include numerous English denominal verbs that lexicalize a substance or container within the verb itself, whereas only a few Croatian verbs belong to them, and these are generally infrequent (e.g., *asfaltirati* 'asphalt', *pozlatiti* 'gild' vs. *arhivirati* 'archive', *konzervirati* 'preserve'). A similar asymmetry is found with the class *instr_communication-37.4.1* (e.g., *e-mail*, *modem*, *phone*, *radio*, *wire*), which has only one member in Croatian (*telefonirati* 'phone'). Rather than lexicalizing the means of communication within the verb itself, Croatian typically expresses it syntactically through an instrumental noun phrase (*zvati telefonom* 'call by phone'), various prepositional phrases (*zvati na telefon/ preko telefona/putem telefona* 'call on the phone/over the phone/via phone/by phone'), or through adverbial modification (*zvati telefonski* 'call by phone'), combined with more general verbs. This reveals a systematic difference in how the two languages distribute semantic components between the lexicon and argument structure – a difference that recurs across several of the mismatch types discussed below.

This asymmetry has a practical consequence for the lexicon. Naming semantic classes in Croatian sometimes proves difficult due to lexical non-correspondence between English and Croatian verb inventories. Two strategies can be adopted in such cases: either descriptive labels are used, or a different representative verb is selected. The choice between these strategies depends on the availability of a suitable Croatian verb that can function as a representative of the class: when no sufficiently general or frequent equivalent exists, a descriptive label is preferred, whereas in cases where a more prototypical verb is available, it can be selected as the class label. For example, the classes *butter-9.9* and *pocket-9.10* require descriptive labels (e.g., *nanositi tvari označene tvorbenom osnovom* 'apply substances specified by the derivational base noun' or *staviti_u ili na spremnik označen tvorbenom osnovom* 'place into/onto a container specified by the derivational base noun'). In other cases, a different representative verb can be selected as the class label. For example, in the class *floss-41.2.1*, which includes the verbs *brush*, *floss*, *shave*, and *wash*, the verb *floss* has no simple-verb equivalent in Croatian, making *četkati* 'brush' a more appropriate class label. These practical decisions are themselves evidence of the deeper structural mismatches examined in the following subsections.

4.1 Lexicalization patterns

The most straightforward type of mismatch concerns differences in lexicalization patterns: how semantic components are distributed between lexical items and syntactic structure.

As illustrated above, English uses a single verb whereas Croatian expresses the meaning through a more general verb and additional elements (e.g., *butter* vs. *mazati/namazati maslacem* 'to spread butter'; *floss* vs. *čistiti (zube) zubnim koncem* 'to clean (one's teeth) with dental floss'). However, the reverse pattern also occurs where Croatian lexicalizes meanings that are expressed in English through phrasal verbs (*smisliti* 'come up with', *promisliti* 'think through', *patiti* 'long for, yearn for') or predicative constructions (e.g., *odgovarati* 'be responsible/liable', *iznenaditi se* 'to be surprised, to be astonished', *nalaziti se* 'to be located', *napisati se* 'to be done with writing') which are not systematically represented in VerbNet. In both directions, the mismatch reflects language-specific conventions governing what semantic components are encoded within the verb itself.

4.2 Reflexive marker se: Lexical shift and argument structure

Reflexive verb pairs represent an additional type of mismatch, as Croatian reflexive morphology often introduces semantic distinctions that do not align neatly with VerbNet classes. Two subtypes can be distinguished, depending on whether the reflexive marker *se* triggers a lexical-semantic shift or encodes a change in argument structure. In the first subtype, *se* corresponds to a shift in lexical meaning, which may result in the assignment of the reflexive and non-reflexive forms to different semantic classes (e.g., *oprostiti* 'forgive' vs. *oprostiti se* 'say goodbye', *obratiti* 'convert' vs. *obratiti se* 'contact', *ukazivati* 'point out, indicate' vs. *ukazivati se* 'appear, manifest'). In these cases, the meaning is not predictable from the base verb, and the two forms require independent class assignment. In the second subtype, *se* does not introduce a new lexical meaning but encodes a shift in argument structure (e.g., *brijati* 'shave' vs. *brijati se* 'shave oneself'). Here, both forms belong to the same semantic class, with the reflexive form realizing a derived reflexive that encodes coreference between the Agent and the Patient (cf. Kettnerová, Lopatková & Vernerová, 2022). Since VerbNet classes are defined for English verbs, which do not have a comparable reflexive morphology, neither type is systematically captured in the existing classification. The proposed distinction provides a principled criterion for deciding between separate class assignment and same-class annotation with a derived reflexive label.

4.3 Alternation-based mismatches

A third type of mismatch arises from the fact that VerbNet classes are defined by participation in specific syntactic alternations that are not always available or productive in Croatian. In VerbNet, certain verbs are distinguished according to whether they participate in the dative alternation. For example, within verbs of communication, the class *tell-37.2* licenses both the prepositional Recipient (e.g., *tell a story to someone*) and the double-object construction (*tell someone a story*), whereas the class *say-37.7* does not productively license the double-object construction and primarily realizes the Recipient in a prepositional phrase. In Croatian, however, a key criterion for the classification of verbs of communication is the morphosyntactic realization of the Recipient, which may appear in the accusative or dative. Constructions with two accusative objects are limited to a small set of verbs (in our data, the verbs *moliti/zamoliti* 'ask for').

Similarly, for the class of change-of-state verbs, the causative/inchoative alternation is central, but it is not realized as a syntactic alternation in Croatian. While English uses the same verb in both causative and inchoative constructions (e.g., *break* in the examples *The window broke* vs. *Tony broke the window*), Croatian typically distinguishes these meanings through reflexive morphology (*razbiti* in the example *Tony je razbio prozor* vs. *razbiti se* in the example *Prozor se razbio*). As a result, the alternation is not a property of a single verb, but is distributed across related

forms. Other alternations, such as the conative alternation (e.g., *cut at*, *hit at*), which allow the expression of an incomplete or attempted action by demoting the direct object to an oblique position, do not have direct productive counterparts in Croatian. Instead, similar meanings are typically expressed lexically or contextually rather than through systematic syntactic alternation.

These examples show that the alternation-based criteria central to VerbNet classification are not always applicable in Croatian, where many of the same semantic distinctions are encoded through reflexive morphology, case marking, and different lexemes rather than through productive syntactic alternations.

4.4 Mismatch between VerbNet classes and Croatian verb properties

In a number of cases, the assignment of Croatian verb senses to VerbNet classes reveals mismatches that go beyond structural differences and instead concern the mapping between Croatian verbs and their English equivalents. These mismatches fall into three groups: cases where an English equivalent is recorded in VerbNet but its classification does not adequately capture the syntactic or semantic properties of the corresponding Croatian verb; cases where no suitable class or sense is available in the VerbNet inventory; and cases involving copular, semi-copular, and light verb constructions, which are not systematically represented in VerbNet.

4.4.1 Mismatches in the classification of English equivalents

These mismatches arise when an English equivalent can be identified, but its classification in VerbNet reflects a different set of syntactic and semantic properties from those exhibited by the corresponding Croatian verb. For example, the verb *congratulate* is classified under the class *judgment-33.1-1*, together with verbs such as *judge*, *applaud*, *approve*, *belittle*, *bless*, which typically encode evaluative meanings and allow predicative complements. However, although *congratulate* denotes a communicative act comparable to the Croatian *čestitati*, its classification reflects an evaluative perspective that does not align with the behaviour of the Croatian verb. In its primary sense ‘to express joy for someone’s success, achievement, or happy event’, *čestitati* does not denote evaluation, but rather a communicative act directed at another participant. Syntactically, it patterns with verbs of communication, as it regularly licenses a Recipient in the dative (*čestitati komu* ‘congratulate (to) someone’) and does not exhibit predicative complement structures characteristic of judgement verbs.

A particularly clear mismatch is observed with the verb *obratiti se*. One of its closest English equivalents, *address* (as the equivalent *contact* is not recorded), is classified in VerbNet under *illustrate-25.3*, together with verbs such as *adorn*, *decorate*, *ornament*, which typically take a direct object and denote modification. However, the Croatian verb *obratiti se* does not exhibit such syntactic or semantic properties. Instead, it regularly licenses a Recipient in the dative (*obratiti se komu* ‘address (to) someone’) and denotes the initiation of communicative contact.

4.4.2 Gaps in VerbNet class coverage

In some cases, the mismatch arises because VerbNet does not provide a suitable class or sense corresponding to the Croatian verb. These cases reflect gaps in the coverage or granularity of the VerbNet inventory. For example, the verb *apply* is classified under *put-9.1-1* and *use-105.1*, but the Croatian verb *javiti se* in the sense ‘to express interest in a job or task’ is not covered by

either class. A similar issue can be observed with the Croatian verb *moliti*. In VerbNet, its English equivalent *pray* is associated with the class *long-32.2*, in which the subject bears the role of Pivot and the Theme is expressed in a prepositional phrase (e.g., *for something*). However, this configuration does not correspond to the Croatian verb *moliti* in the sense 'to say a prayer' (e.g., *moliti Očenaš* 'to say the Lord's Prayer'), where the verb denotes a ritualized communicative act taking a direct object in the accusative.

4.4.3 Copular, semi-copular, and light verb constructions

Copular, semi-copular, and light verbs are not always adequately represented within VerbNet's semantic class inventory, although VerbNet does contain several classes involving predicative complements. The verb *be* appears in several VerbNet classes, but its core copular use (e.g., *John is a teacher*, *Mary is happy*) is not systematically captured within the standard class inventory. Furthermore, for a number of verbs that have undergone semantic bleaching and developed semi-copular uses, these meanings are not systematically represented. For example, the verb *make* in the sense 'to represent or characterize someone as something', as in *He made me a fool*, is not listed in *characterize-29.2*, although it functions as a verb introducing a predicative complement. A related problem concerns light verb constructions, in which the semantic core is typically carried by the nominal complement, while the verb itself has reduced lexical content. This limitation is not unique to VerbNet as comparable gaps have been noted in PropBank with respect to light verb constructions (cf. Hwang et al., 2010; Bonial et al., 2014), and FrameNet (Petruck & Ellsworth, 2016), suggesting that the systematic treatment of light verbs remains an open problem across major semantic annotation frameworks. In VerbNet, therefore, verbs such as *give*, *make*, *do*, or *conduct* are assigned to classes reflecting their independent lexical semantics, even when they function as light verbs in constructions such as *give a speech*, *make a decision*, *do damage*. As a result, Croatian light verb constructions such as *vršiti ispitivanja* 'conduct surveys', *napraviti štetu* 'do damage', or *dati podršku* 'give support' are left without an appropriate semantic class, since VerbNet does not systematically account for this type of meaning.

4.5 Gaps in the VerbNet inventory

Finally, some verbs do not have corresponding English equivalents recorded in the VerbNet inventory used in this study. In our data, this applies to verbs such as *javiti se* 'contact', *lagati* 'lie', *ispričati se* 'apologize', *izdati* 'betray', *nedostajati* 'lack'. These gaps are particularly significant for high-frequency verbs, as their absence from VerbNet prevents consistent cross-linguistic alignment and highlights the need for language-specific extensions of the inventory.

While some mismatches arise from general typological differences between the two languages, others reflect limitations in the coverage and structure of VerbNet itself. The analysis thus shows that cross-linguistic application of syntactic-semantic classifications is not straightforward and requires both careful interpretation and targeted modification to adequately capture language-specific properties.

5 Conclusion

This paper has examined the challenges involved in assigning VerbNet semantic classes to Croatian verb senses, based on data from the Croatian verb lexicon Verbion. The analysis has shown that these challenges are not limited to isolated cases, and reflect systematic differences between English and Croatian in the encoding of semantic and syntactic information.

Several recurrent types of mismatch have been identified, including differences in lexicalization patterns, reflexive morphology, alternation-based classification, class assignment, and gaps in VerbNet coverage. In particular, semantic components that are lexicalized in one language are often expressed syntactically in another. Croatian reflexive morphology frequently corresponds to lexical-semantic distinctions rather than to purely syntactic alternations. Furthermore, classifications based on syntactic alternations such as the dative, causative/inchoative, or conative alternation cannot always be directly transferred, since these distinctions are often realized through case marking, reflexive morphology, or separate lexical forms. Additional discrepancies arise when English-based VerbNet classes do not adequately capture the semantic or syntactic behaviour of Croatian verbs, or when relevant senses and verbs are absent from the VerbNet inventory altogether.

Taken together, these findings demonstrate that the cross-linguistic application of VerbNet requires systematic adaptation at multiple levels. Rather than directly transferring English-based classes, it is necessary to account for language-specific features. At the same time, some of the identified problems also reflect limitations in the coverage and organization of VerbNet itself, particularly with respect to phrasal verbs, copular, semi-copular, and light-verb constructions.

In this respect, the Croatian verb lexicon Verbion provides a suitable framework for such adaptation, as it integrates lexical, semantic, syntactic, and morphological information within a unified model. The approach proposed in this paper therefore contributes to the development of a Croatian-specific extension of VerbNet, while preserving its comparative and cross-linguistic potential. Despite the identified challenges, the application of VerbNet-style classification enriches Croatian linguistic resources and contributes to broader discussions on the applicability of syntactic-semantic classifications across typologically different languages.

Since the present study is based on the subset of verbs processed so far, the findings should be regarded as preliminary. Future work will include an analysis of the remaining verbs planned within the project. Further research will focus on refining Croatian verb classes, expanding the coverage of the lexicon, and examining the interaction between semantic roles, morphosyntactic realization, and verb classification.

References

Baisa, V., Michelfeit, J., Medveď, M. & Jakubíček, M. (2016). European Union language resources in Sketch Engine. In N. Calzolari et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 2799–2803). European Language Resources Association (ELRA).

Bañón, M., et al. (2023). Croatian web corpus MaCoCu-hr 2.0. Slovenian language resource repository CLARIN.SI. http://hdl.handle.net/11356/1806

Birtić, M. (2020). Usporedba mrežnih valencijskih rječnika u slavenskim jezicima. *Rasprave: Časopis Instituta za hrvatski jezik i jezikoslovlje*, 46(2), 483–510.

Birtić, M., Brač, I. & Runjaić, S. (2017). The main features of the e-Glava online valency dictionary. In I. Kosem et al. (Eds.), *Electronic lexicography in the 21st century. Proceedings of eLex 2017 conference* (pp. 43–62). Lexical Computing CZ s.r.o.

Bonial, C., Bonn, J., Conger, K., Hwang, J. D. & Palmer, M. (2014). PropBank: Semantics of New Predicate Types. In N. Calzolari et al. (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (pp. 3013–3019). European Language Resources Association.

BulNet. Retrieved May 22, 2026, from https://dcl.bas.bg/bulnet/

Brač, I. & Bošnjak Botica, T. (2015). Semantička razdioba glagola u Bazi hrvatskih glagolskih valencija. *Fluminensia*, 27(1), 105–121.

Chomsky, N. (1981). *Lectures on government and binding*. Foris.

Dorr, B. J. (1997). Large-scale dictionary construction for foreign language tutoring and interlingual machine translation. *Machine Translation*, 12(4), 271–322.

Fellbaum, C. (1998). A semantic network of English verbs. In C. Fellbaum (Ed.), *WordNet: An electronic lexical database* (pp. 69–104). The MIT Press.

Filko, M., D. Katunar, O. Antoni, I. Raffaelli, K. Šojat & M. Tadić (2015). CroWN – Croatian WordNet 2.0, HR-CLARIN, http://hdl.handle.net/20.500.14615/19.

FrameNet. Retrieved May 29, 2026, from https://framenet.icsi.berkeley.edu/

Hwang, J. D., Bhatia, A., Bonial, C., Mansouri, A., Vaidya A., Xue, N. & Palmer, M. (2010). PropBank Annotation of Multilingual Light Verb Constructions. In N. Xue & M. Poesio (Eds.), *Proceedings of the Fourth Linguistic Annotation Workshop* (pp. 82–90). Association for Computational Linguistics.

Hrvatski jezični portal. Retrieved May 25, 2026, from https://hjp.znanje.hr/

Jakubíček, M., Suchomel, V., Martelli, F. & Navigli, R. (2022). Semantically annotated corpora. https://elex.is/wp-content/uploads/ELEXIS_D4_6_Semantically_annotated_corpora.pdf

Kettnerová, V., Lopatková, M. & Bejček, E. (2012). Mapping semantic information from FrameNet onto VALLEX. *The Prague Bulletin of Mathematical Linguistics*, 97, 23–41.

Kettnerová, V., Lopatková, M. & Vernerová, A. (2022). Reflexives as part of verb lexemes in the VALLEX lexicon. *The Prague Bulletin of Mathematical Linguistics*, 119, 37–66.

Kilgarriff, A., Baisa, V., Busta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, 1, 7–36. https://doi.org/10.1007/S40607-014-0009-9

Kipper, K., Dang, H. T. & Palmer, M. (2000). Class-based construction of a verb lexicon. In H. A. Kautz & B. W. Porter (Eds.), *Proceedings of the Seventeenth National Conference on Artificial Intelligence and Twelfth Conference on Innovative Applications of Artificial Intelligence (AAAI/IAAI)* (pp. 691–696). AAAI Press.

Kipper, K., Korhonen, A., Ryant, N. & Palmer, M. (2008). A large-scale classification of English verbs. *Language Resources and Evaluation*, 42, 21–40. https://doi.org/10.1007/s10579-007-9048-2

Kipper-Schuler, K. (2005). VerbNet: A broad-coverage, comprehensive verb lexicon [Doctoral dissertation, University of Pennsylvania]. ProQuest.

Korhonen, A. (2002). Semantically motivated subcategorization acquisition. In N. Ide & J. Pustejovsky (Eds.), *Proceedings of the ACL-02 Workshop on Unsupervised Lexical Acquisition* (pp. 51–58). Association for Computational Linguistics.

Korhonen, A. & Briscoe, T. (2004). Extended lexical-semantic classification of English verbs. In *Proceedings of the HLT/NAACL Workshop on Computational Lexical Semantics* (pp. 38–45). Association for Computational Linguistics.

Levin, B. (1993). *English Verb Classes and Alternations*. University of Chicago Press.

Lopatková, M., Kettnerová, V., Jiří, M., Vernerová, A., Bejček, E. & Žabokrtský, Z. (2022). *VALLEX 4.5*, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), http://hdl.handle.net/11234/1-4756.

Ljubešić, N. & Klubička, F. (2014). {bs,hr,sr}WaC – Web corpora of Bosnian, Croatian and Serbian. In F. Bildhauer & R. Schäfer (Eds.), *Proceedings of the 9th Web as Corpus Workshop (WaC-9)* (pp. 29–35). Association for Computational Linguistics.

Ljubešić, N. & Kuzman, T. (2024). CLASSLA: A linguistic annotation pipeline for South Slavic languages enriched with linguistic and genre annotation. In N. Calzolari et al. (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 3271–3282). ELRA–ICCL.

Marini, F. & Ježek, E. (2019). CroaTPAS: A lexicographic resource for Croatian verbs and its potential for Croatian language teaching. In Z. Gavriilidou, M. Mitsiaki & A. Fliatouras (Eds.), *Proceedings of the 19th EURALEX International Congress* (pp. 529–534). Democritus University of Thrace.

Majewska, O., Vulić, I., McCarthy, D., Huang, Y., Murakami, A., Laippala, V. & Korhonen, A. (2018). Investigating the cross-lingual translatability of VerbNet-style classification. *Lang Resources & Evaluation*, 52, 771–799.

Mikelić Preradović, N. (2008). Pristupi izradi strojnog tezaurusa za hrvatski jezik [Doctoral dissertation, Filozofski fakultet Sveučilišta u Zagrebu].

Mikelić Preradović, N. (2020). *CROVALLEX: Valencijski leksikon glagola hrvatskoga jezika*. Filozofski fakultet Sveučilišta u Zagrebu.

Petruck, M. R. L. & Ellsworth, M. (2016). Representing Support Verbs in FrameNet. In V. Kordoni et al. (Eds.), *Proceedings of the 12th Workshop on Multiword Expressions* (pp. 72–77). Association for Computational Linguistics.

Przepiórkowski, A., Hajnicz, E., Patejuk, A., Woliński, M., Skwarski, F. & Świdziński, M. (2014). Walenty: Towards a comprehensive valence dictionary of Polish. In N. Calzolari et al. (Eds.), *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)* (pp. 2785–2792). ELRA.

Rizov, B. & Dimitrova, T. (2016). Hydra for Web: A browser for easy access to wordnets. In Ch. Fellbaum et al. (Eds.), *Proceedings of the 8th Global WordNet Conference (GWC)* (pp. 342–346). Global Wordnet Association.

Ruppenhofer, J., Ellsworth, M., Petruck, M. R. L., Johnson, C. R. & Scheffczyk, J. (2016). *FrameNet II: Extended theory and practice*. International Computer Science Institute. https://my.eng.utah.edu/~cs6961/papers/FrameNet_book.pdf

Školski rječnik hrvatskoga jezika. Retrieved May 20, 2026, from https://rjecnik.hr/

Tiedemann, J. (2016). Finding alternative translations in a large corpus of movie subtitles. In N. Calzolari et al. (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)* (pp. 3518–3522). European Language Resources Association (ELRA).

Unified Verb Index. Retrieved May 25, 2026, from https://uvi.colorado.edu/

Van Valin, R. D. (2000). *An introduction to syntax*. Cambridge University Press.

Van Valin, R. D. & LaPolla, R. J. (1997). *Syntax: Structure, meaning and function*. Cambridge University Press.

Contact information

Ivana Brač
Institute for the Croatian Language
ibrac@ihjj.hr

Acknowledgements

This work has been supported by the Croatian Science Foundation under the project *Semantic-Syntactic Classification of Croatian Verbs* (SEMTACTIC) (IP-2022-10-8074) and by the European Union – NextGenerationEU under the project VERBION. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Katja Brankačkec & Barbora Štěpánková

MULTIFUNCTIONAL FUNCTION WORDS IN THE UPPER SORBIAN–CZECH DICTIONARY MUDRA 2.0

Abstract This paper focuses on the lexicographic treatment of multifunctional function words in the emerging Upper Sorbian–Czech dictionary Mudra 2.0. As Upper Sorbian lexicography remains strongly dependent on German and lacks a monolingual explanatory dictionary, the compilation of bilingual Upper Sorbian–"non–German" dictionaries poses specific methodological challenges. These are particularly evident in the treatment of expressions with predominantly grammatical or pragmatic functions, such as conjunctions, modal words, and intensifiers. These expressions are often characterised by polysemy and multifunctionality resulting from grammaticalization processes, and their meanings are highly context–dependent. The treatment of these expressions in monolingual dictionaries has been inconsistent so far, whereas in bilingual dictionaries it is typically limited to providing translation equivalents. In this paper, we discuss the various possibilities and limitations of lexicographic descriptions of function words in the context of the minority language. Through a case study of the Upper Sorbian expressions *docyła* ('completely') and *cyle* ('quite'), we analyse their semantic and pragmatic properties in comparison to their equivalents in Czech and German. Based on this analysis, we present the lexicographic approach adopted in the dictionary, including sample entries.

Keywords Upper Sorbian; function words; bilingual dictionary; minority language; grammaticalization

1 Introduction

Function words, i.e. expressions with predominant grammatical or pragmatic function, whose meaning is fully apparent only in a context (cf. Filipec & Čermák, 1985, p. 39), are often homonymous (due to their origin in content words) and polysemous/multifunctional (due to the process of grammaticalization,[1] cf. Diewald, 2011). Their treatment in monolingual dictionaries is rather marginal (cf. Hanks, 2009), in bilingual dictionaries, it is generally limited to translation equivalents. This paper focuses on the possibilities and pitfalls of the lexicographic treatment of these expressions. We demonstrate this for the purpose of an emerging bilingual Upper Sorbian-Czech dictionary Mudra 2.0, which additionally reflects an unusual linguistic situation: no monolingual dictionary exists for the source language, and another language (namely German) serves as an intermediary in the building of the translation dictionary. This special situation also defines our main target users: Czech learners of Upper Sorbian with little knowledge of German. Nevertheless, our theoretical conclusions could be useful for the treatment of such expressions in bilingual dictionaries in general.

The paper is organized as follows: Firstly, we describe the current state of the Upper Sorbian language and its linguistic resources (Sec. 2). Then, we introduce the emerging Upper Sorbian-Czech dictionary Mudra 2.0 (Sec. 3). In Section 4, we discuss possible approaches to the lexicographic treatment of function words, focusing on expressions with a predominant pragmatic function (such as pragmatic markers/particles, modal words, intensifiers, etc.). As a case study,

[1] In agreement with Degand & Evers-Vermeul (2015), we use the term *grammaticalization* as an umbrella term encompassing related concepts such as *pragmaticalization*, etc.

we analyse the Upper Sorbian words *docyła* 'completely' and *cyle* 'quite',[2] focusing on the semantic and pragmatic features relevant to their lexicographic description and we compare these features with those of their Czech translation equivalents. Based on this analysis, we present the approach adopted in our dictionary. In addition, we include sample lexical entries for the words under study.

2 Upper Sorbian and Upper Sorbian dictionaries

Upper Sorbian is a West Slavic minority language characterized by a specific sociolinguistic situation: its relatively small speech community is almost universally bilingual in German.[3] As a consequence there is no monolingual Upper Sorbian dictionary. This function is therefore filled by the bilingual Upper Sorbian–German and German–Upper Sorbian dictionaries. Almost all the recent ones are now available online via the platforms *Soblexx* and *hornjoserbsce.de*. Other Upper Sorbian bilingual dictionaries are rare and often of a small size.

Another important source of information are Upper Sorbian language corpora, which are not yet comparable to those available for larger languages: The *Hotko* Upper Sorbian corpus (44.4 million tokens, neither lemmatised nor tagged), the corpus accessible via *hornjoserbsce.de* (approximately 95 million tokens, lemmatised), and the parallel corpus *InterCorp*[4] (586,000 tokens). Unfortunately, the parallel corpus is not of a sufficient size for a data- or corpus-based approach such as the translation equivalents method (cf. Aijmer et al., 2006).

3 Mudra 2.0

The Mudra 2.0 project originated from a torso of an Upper Sorbian-Czech dictionary, which was started by a Czech Sorabist Jiří Mudra at the end of the 20th century. His unfinished manuscript with about 20,000 edited entries was automatically converted and transferred to a Tshwane-Lex database, where the entries are linked to the corresponding index cards, manually corrected and expanded in terms of content (cf. Škrabal & Brankačkec, 2021, and the imprint of the dictionary website).[5] The database is available online as a growing dictionary and currently contains about 46,000 entries. In addition to references to other entries with a similar meaning, the edited entries are also externally linked to selected historical Upper Sorbian-German (Kral, 1927; Pful, 1866) and Upper Sorbian-Russian (Trofimovič, 1974) dictionaries, to the corresponding concordance in the *Hotko* corpus including the subcorpus of the Upper Sorbian-English (Stone, 2002) and the English-Upper Sorbian dictionary (Strauch & Wornar, 2007), the phraseological online dictionary on the *hornjoserbsce.de* platform and to the Upper Sorbian text corpus accessible on this platform. Information on inflection can be accessed via a link to the Upper Sorbian-German-Upper Sorbian online dictionary Soblexx and its Flex Tables. In addition, selected entries contain references to the corresponding pages on Wikipedia.

The latest version of Mudra 2.0 was released in April 2026 and includes completed entries "a – jadro" (approximately 10,500 entries). Within this section, individual entries still requiring a more complex lexicographic workflow remain designated as "work in progress".

[2] When glossing our examples, we use the English equivalents that are most appropriate, leaving a detailed analysis of English function words aside in this article.

[3] For the situation of Sorbian language see Faska, 1998; Brankačkec et al., 2019. Breu (2015, pp. 9ff.) describes the actual language situation as "full linguistic contact".

[4] For Hotko and InterCorp see http://www.korpus.cz/

[5] See <https://mudra.korpus.cz/_about>

Among those entries are, for example, auxiliaries such as *być* 'to be', but also function words, such as *docyła* 'completely'or *cyle* 'quite', which form the central focus of this paper.

4 Function words in dictionaries

Similarly to monolingual dictionaries, bilingual dictionaries primarily focus on content words, with function words being of secondary interest. One reason for this can be seen in the fact that the interest in function and pragmatic expressions has only recently arisen. Therefore, a standardised lexicographic treatment has not been agreed upon, often not even in monolingual dictionaries (cf. Štěpánková et al., 2024). For example, in Czech monolingual dictionaries, these words were often only treated using synonyms or through cross-references to their base adjectives (cf. intensifier *úplně* (1)). It is only in the emerging Akademický slovník současné češtiny dictionary (hereafter ASSČ) that all meanings are explained. On the contrary, German monolingual dictionaries provide a more detailed description of these words (see examples 11–13 below). However, in bilingual dictionaries, function words are usually described only through translation equivalents.

> (1) úplně: 'přísl. naprosto, zcela, docela: žít úplně sám; úplně s tebou souhlasím' 'completely: adv. totally, absolutely, quite: live all by myself; I completely agree with you' (Slovník spisovné češtiny, hereafter SSČ)

At the same time, function words often exhibit multifunctionality and polysemy, which are linked to their evolution from adverbs to expressions with significant pragmatic or grammatical functions, i.e., grammaticalization (cf. Traugott, 1989; Nevalainen & Rissanen, 2002). Diewald (2013) generalizes this phenomenon and describes two typical situations in terms of the synchronous behaviour of these expressions: (i) In some cases it is possible to distinguish between several meanings or even define different parts of speech of the expression;[6] (ii) in other cases it is difficult to interpret the usages of such multifunctional words unambiguously (Šindlerová & Štěpánková, 2021). Example (2) illustrates a usage of the Upper Sorbian multifunctional word *scyła* which can be interpreted either as an adverb or as a pragmatic particle.

> (2) Scyła wosebity grat na konje je jutrowny. (Ćišinski, 1875 – Hotko)'A very special horse harness is used at Easter.' (adverb)'In fact, the horse harness used at Easter is special.' (particle)

Given the polysemous nature of function words, it is generally acknowledged that basing their treatment solely on translation equivalents is inadequate. Moreover, Atkins and Rundell (2008, p. 472) argue that their semantic content alone is insufficient, emphasizing that "the function of the word (the role it plays in expressing or interpreting the meaning of a phrase or sentence), together with its collocational context," is equally important in bilingual dictionaries. A widely proposed solution is therefore to supplement dictionary entries with additional information such as definitions, usage notes, grammatical comments and illustrative examples (see Hoekstra, 2010, on grammatical information; Cacchiani, 2016, on the treatment of intensifiers in American dictionaries).

However, the implementation of this solution is complicated by a lack of terminological consistency and diverse theoretical approaches in this area. For example, labels such as 'particles', 'synsemantics', 'pragmatic markers', and 'sentence adverbs' are sometimes used interchangeably, whereas other approaches treat them as only partially overlapping categories (cf. Rozumko, 2016). Furthermore, there is no consensus on whether all the meanings or functions of a lexical item should be assigned to a single part-of-speech category (see the more restrictive position adopted by Felbr, 2024, p. 121).

[6] Cf. Bordet (2016) on the entries for English *very, really, so, totally* in Merriam Webster.

The aim of this paper is to describe the process through which we developed a method for the consistent lexicographic treatment of multifunctional function words in a bilingual dictionary. To illustrate this approach, we present a case study of the Upper Sorbian expressions *docyła* and *cyle*, demonstrating how we identified the information essential for their lexicographic description. To arrive at a comprehensive account of their use, we compared the treatment provided in the most detailed grammar of Upper Sorbian currently available (Faßke & Michalk, 1981) with corpus evidence and with the corresponding dictionary entries in Czech, German, and Upper Sorbian lexicographic sources (see Figure 1).

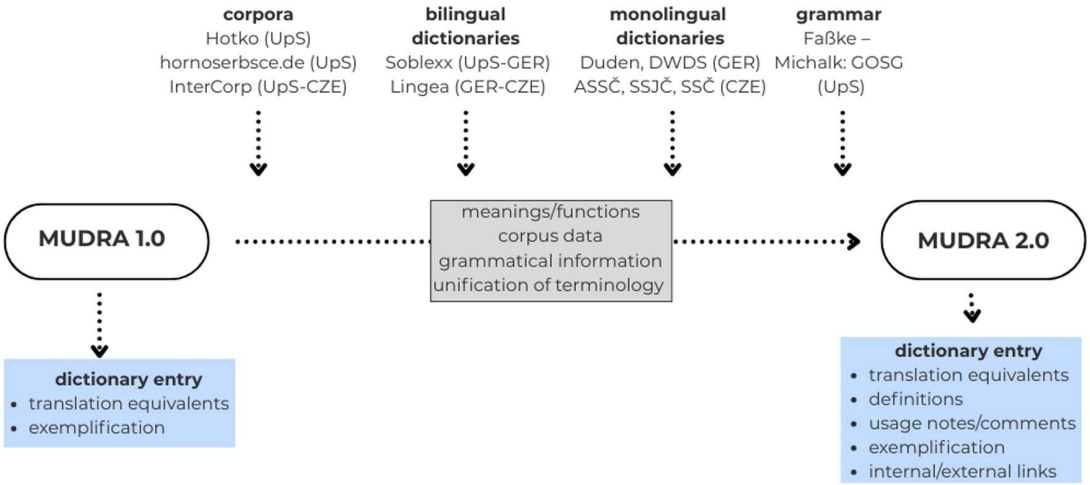


FIGURE 1 Scheme for creating function words entries in the MUDRA 2.0 dictionary

5 Case study

5.1 *Docyła*

In Upper Sorbian – German dictionaries, the entry for the polysemous function word *docyła* relies primarily on translation equivalents:

(3) docyła 1. indekl. ganz, völlig; 2. part. überhaupt
1. 'entirely, completely'; 2. 'at all' (Soblexx)

The Upper Sorbian-English learner's dictionary (Stone, 2002) adds an example and its translation to the entry:

(4) docyła adv. completely, totally, at all; to d. wěrno njeje that is completely untrue (Stone, 2002)

In the former dictionary, we see that the entry is divided into two meanings; the advantage of the latter one is that the translation equivalent is supported by an example. However, none of the dictionaries explicitly describes the functions of the given meanings. This can be an issue since translation equivalents are polysemous (see below). From the equivalents and the example in Stone's dictionary we can judge that *docyła* is used as an intensifier of totality and as an intensifier of negation. This interpretation is also supported by Faßke's Upper Sorbian Grammar (Faßke & Michalk, 1981), which lists two functions/meanings of the word *docyła*: (i) "pronoun of totality"[7] (ibid., p. 636), (ii) a word modifying particle and intensifier of negation (ibid., p. 771).

[7] Faßke's and Michalk's Grammar is written in German and uses the term *Totalitätspronomen* here.

The first meaning "*intensifier*" ('entirely, completely') is translated into Czech as 'docela, úplně'.[8] In Czech monolingual dictionaries (SSČ, Slovník spisovného jazyka českého, hereafter SSJČ) these expressions are described as adverbs expressing the greatest degree of something. However, they are subject to grammaticalization and are also used as intensifiers (similar to their German equivalents, cf. Šindlerová & Štěpánková, 2021). The Czech word 'docela' – similarily as German 'ganz' (see below) – can even have the meaning 'relative, incomplete degree' (cf. ASSČ), thus becoming a so-called 'false friend' for learners of Upper Sorbian.

For the second meaning "intensifier of negation" – 'at all' – the Czech equivalent 'vůbec' is used, which has at least two (SSČ), but as many as six (SSJČ) meanings in Czech monolingual dictionaries. The German equivalent *überhaupt* is similarly polysemous. In addition, this meaning (see (5)), has largely supplanted the original one in contemporary Upper Sorbian. The meaning 'completely' is attested mainly in older texts, as illustrated in (6). In present-day Upper Sorbian, this obsolete meaning has a partial synonym in *cyle* ('wholly, completely, entirely'), which occurs more frequently.

(5) Ale njejsym docyła zrudny, wšako je to rjany turnĕr z dobrej naladu.'But I am not sad at all, after all it is a nice competition with a good atmosphere.' (Serbske Nowiny 2005 – Hotko)
(6) Słowjanska powaha je docyła rozdźĕlna wot němskeje. 'The nature of Slavic is completely different from German.' (Ćišinski 1885 – Hotko)

As our dictionary is based on an existing dictionary torso, the information prepared by our predecessor, Jiří Mudra, serves as another point of departure. The entry in Mudra's manuscript provides the following:

(7) docyła: docela, naprosto; to ~ wĕrno njeje to vůbec není pravda; to je ~ njemóžne to je naprosto nemožné
'completely: entirely, altogether; that is completely untrue; that is completely impossible' (Mudra 1.0)

It is confusing that Mudra 1.0 illustrates the obsolete meaning with direct equivalents while providing the more common meaning only as an example. This does not meet the requirement for universality in basic translation equivalents (cf. Aktins & Rundell, 2008, p. 467).

5.2 Cyle

As mentioned above, the Upper Sorbian word *cyle* shows both functions as well[9] – parallelly to the German equivalent *ganz*. This word is presented by Mudra ´s torso as follows:

(8) cyle zcela, docela; ~ wĕśće ~ určitĕ; sym tu ~ zbytny jsem tu úplně zbytečný
'quite; for sure, absolutely; I am completely useless here' (Mudra 1.0)

Again, Mudra gives only one meaning as a direct equivalent but mentions the other meaning as an example. The Upper Sorbian-German dictionary Soblexx provides the following information:

[8] The translation equivalents were proposed based on German-Czech dictionaries and verified in the InterCorp parallel corpus (Czech and German versions), using its derivative tool Treq.
[9] However, Upper Sorbian *cyle* shows another function, which is presented in our dictionary as meaning 3, compare the proposed entry in Mudra 2.0 below.

(9) cyle indekl. (→ cyły) ganz, völlig; ~ jednorje ganz einfach
'completely, absolutely; simply, just'
cyle wěsće mod. ganz gewiss, ganz bestimmt
'for sure, definitely' (Soblexx)

In Stone's Upper Sorbian-English dictionary (2002), the entry looks like this:

(10) cyle adv. quite, completely, entirely, wholly c. jednorje, quite simply (Stone, 2002)

From these equivalents and examples, we can judge that both meanings are "intensifiers": the first of the type "pronoun of totality" (Faßke & Michalk, 1981, p. 652), the second function is a moderating one (Pintarić & Frleta, 2014, p. 33),[10] which is not mentioned in Faßke's and Michalk's grammar, but examples of such usage appear in the corpus (cf. Figure 3, the draft of the *cyle* entry).

In comparison to this, we find a more complex picture in German dictionaries for the equivalent *ganz* (which corresponds to Czech *docela* similarly as to Upper Sorbian *cyle*) and to *überhaupt* (which corresponds to Upper Sorbian *docyła* in its more frequent function).

The German monolingual dictionary Duden mentions five meanings for *ganz*. For our purpose, only two of them are relevant. The "moderating meaning" and the "pronoun of totality" (Faßke & Michalk, 1981, p. 652) are given in the second meaning of the Duden entry for *ganz*:

(11) a) ziemlich, einigermaßen; Grammatik – einschränkend bei Adjektiv; das Wetter war ganz schön; es gefällt mir ganz gut
'quite, to some extent; Grammar – restricting adjectives; the weather was quite fine; I like it quite well'
b) sehr; überaus; Grammatik – intensivierend bei Adjektiv; ein ganz kleiner Rest; er war ganz begeistert
'very; exceedingly; Grammar – intensifying adjectives; a very small remainder; he was very enthusiastic' (Duden)

In comparison to this, the monolingual dictionary Digitales Wörterbuch der deutschen Sprache (hereafter DWDS) gives six meanings, the two relevant usages for our study are given in separate meanings here:

(12) stets betont a) vollkommen; b) (ganz und gar) =vollständig; c) in abgeschwächter Bedeutung; sehr
'always with stress a) completely; b) completely, totally; c) in less strong sense; very' (DWDS)
(13) bezeichnet eine Einschränkung: ziemlich
'indicates a restriction: quite' (DWDS)

The DWDS dictionary provides a larger number of examples as well as more grammatical information than the Duden dictionary. It is interesting to note that the DWDS meaning in 12c) refers to our observation of the usage of Upper Sorbian *cyle* in the third meaning (see meaning 3 in our draft of Mudra 2.0 below), which is lacking for Czech *docela*.

Thus, both German dictionaries provide threefold information on the usage of these words: partial synonyms, grammatical information (e.g., part of speech) and/or formal definitions as well as rich exemplification. In our case, therefore, the dictionaries serve not only to provide

[10] ASSČ uses the term "relativization of degree" for both intensification and attenuation. However, it doesn't quite capture the completely abstract "attitudinal" usage (for Czech *docela*): see meaning 3 in ASSČ.

relevant information on these German function words but also became an important point of inspiration for the treatment of function words in Mudra 2.0.

While the principles of lexicographic presentation on these function words are very similar in both German monolingual dictionaries, they vary in the linguistic interpretation of the data. The Duden interprets at least part of the usages of *ganz* and *überhaupt* as a particle, the DWDS judges all of them as adverbs. On the one hand, this shows the different terminological approaches, but on the other hand, it can sometimes also reveal the ambiguity of some usages of these multifunctional words.

For the reasons outlined above, we regard part-of-speech categorisation as being of only secondary importance in the description of function words. Instead, primary emphasis is placed on the description of meaning and function, supported by clear and informative exemplification.

5.3 Draft of a Mudra 2.0 entry *docyła*

For our entry on *docyła* (Figure 2), we decided to give two basic meanings, while the more frequent one has an even finer-grained distinction: depending on the semantic and grammatical context in which it is used, the Czech equivalents can differ. There is a more general definition for the **meaning 1** "emphasizes the negative and treats the deviation as unexpected; in questions, it can modify a verb without negation, in which case it expresses the speaker's doubt about a positive answer". This meaning is divided into three shades of meaning:

(a) "(in negative sentences), it reinforces or emphasizes the negation"; Czech equivalents *vůbec, naprosto* 'completely, totally', *absolutně* 'absolutely';

(b) "(in a question) expresses doubt", Czech equivalents *vůbec* 'at all', *vlastně* 'actually';

(c) "indicates general applicability"; Czech equivalents *vůbec, celkově* 'on the whole', *obecně* 'generally'.

For **meaning 2** *úplně, zcela, naprosto, dočista* 'totally, wholly, absolutely, clearly', with the function of a maximizing intensifier, the attributes 'obsolete' and 'less frequent' are added. Furthermore, the meanings are illustrated with examples in the entry.

Finally, links to partial Upper Sorbian synonyms of the individual functions complete the entry. The latter are especially important for those functions where *docyła* is used less frequently or no more. However, it should be noted, that some of these entries for partial synonyms were not fully processed yet.

5.4 Draft of a Mudra 2.0 entry *cyle*

For our entry of *cyle* (Figure 3), we defined three basic meanings/functions: **Meaning 1** refers to a maximizing intensifier, often modifying an adjective or adverb 'completely, totally, on the far, at the very'. Again, there is a formal description of the function ("indicates the general validity of the content") besides links to partial Upper Sorbian synonyms, examples and, of course, the Czech equivalents. In the translation of the examples, the corresponding Czech function word is bolded for better orientation of the user. In addition, this meaning contains two fixed collocations: *cyle jednorje* 'quite simply' and *cyle wěsće* 'definitely'.

Meaning 2 expresses a function of a moderating intensifier ("diminishes the intensity, expresses an incomplete degree") and is illustrated likewise.

Meaning 3 describes a function of a boosting intensifier according to the scale offered by Pavić Pintarić and Frleta (2014, p. 33) with an expressive connotation. This function has a parallel in the German equivalent *ganz*.

6 Conclusion

The lexicographic treatment of function words constitutes a specific area of bilingual dictionary compilation. Whereas the treatment of content words primarily relies on selecting semantically appropriate translation equivalents supported by corpus evidence, function

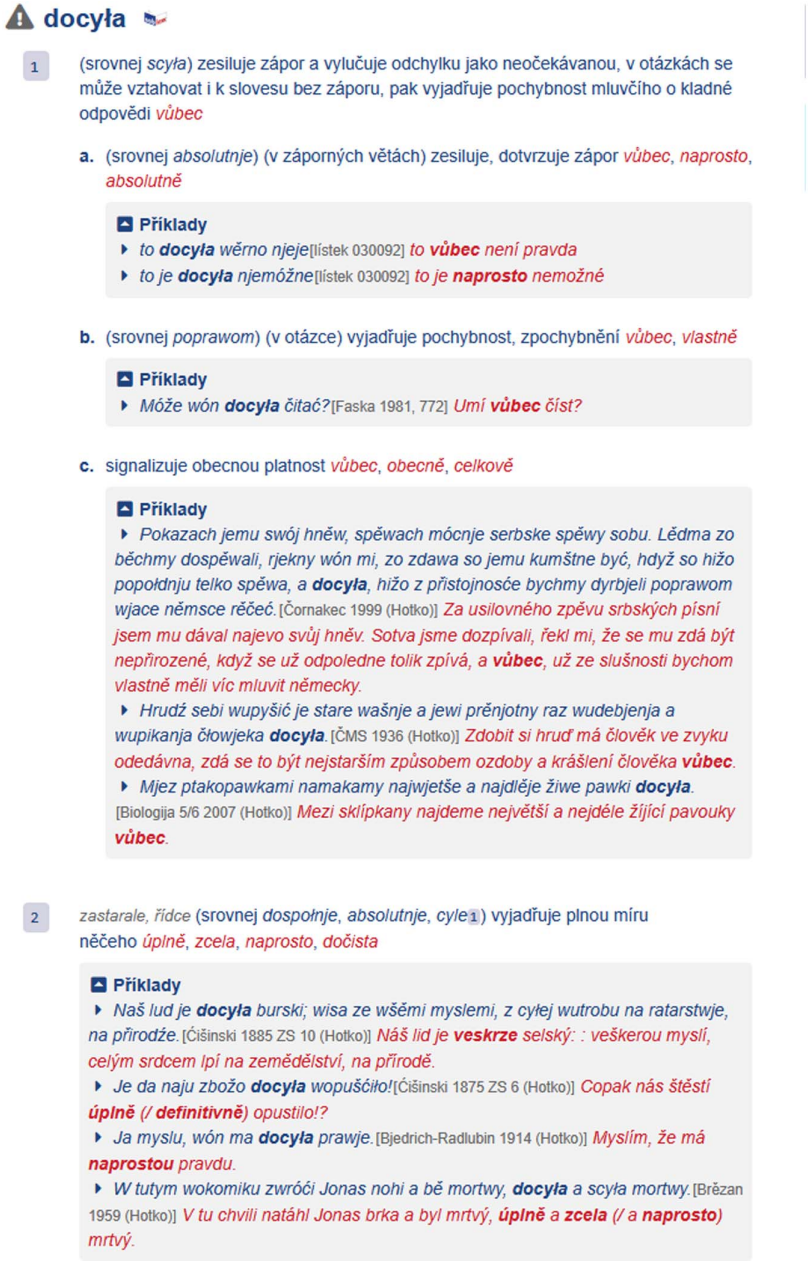


FIGURE 2 The entry *docyła* in Mudra 2.0

words additionally require explicit descriptions of their pragmatic and grammatical functions, especially due to their frequent polysemy. In this study, we highlighted the specific challenges of their lexicographical treatment in a bilingual dictionary of a minority language with limited linguistic resources.

Using the example of two Upper Sorbian intensifiers we demonstrated how the absence of a monolingual Upper Sorbian dictionary could be compensated by corpus analysis and cross-language comparison (based on German, Czech and Upper Sorbian grammars and Upper Sorbian bilingual dictionaries). Owing to the close language contact, German monolingual dictionaries represent an important source as well. We took into account the different terminological traditions and developed a comprehensive formal description for each function.

The cross-linguistic comparison led us to a deeper understanding of the individual functions of these lexemes. It appears that the boundaries between those individual functions do not always coincide with the traditional classification of parts of speech. Instead, a variety of



FIGURE 3 The entry *cyle* in Mudra 2.0

specific phenomena of a syntactic, semantic and pragmatic nature must be considered. Therefore, in cases where category boundaries are not clear, but rather fluid, we do not specify parts of speech in dictionary entries. In general, we emphasize providing clear definitions and comments that complement the translation equivalents, and illustrating them with appropriate exemplification. This is demonstrated by the sample entries of the expressions *docyła* and *cyle*. To a certain extent, the detailed treatment of function words in Mudra 2.0 compensates the lack of a monolingual Upper Sorbian dictionary and may serve as a foundation for future Upper Sorbian lexicographic projects.

References

Aijmer, K., Foolen, A. & Simon-Vandenbergen, A. M. (2006). Pragmatic markers in translation: a methodological proposal. In K. Fischer (Ed.), *Approaches to discourse particles* (pp. 101–114). Elsevier.

ASSČ: *Akademický slovník současné češtiny*. Retrieved May 26, 2026, from https://slovnikcestiny.cz/.

Atkins, S. B. T. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press.

Bordet, L. (2016). From vogue words to lexicalized intensifying words: the renewal and recycling of intensifiers in English. A case-study of very, really, so and totally. *Lexis*, 10, 1–16.

Brankačkec, K., Martínek, F. & Paap, A. (2019). *Lehnprägungen im Tschechischen und Sorbischen*. Lang.

Brankačkec, K., Štěpánková, B. & Škrabal, M. (2026). Hornolužickosrbský výraz hewak: mezi porovnáním jazykových dat a gramatických tradic. *Časopis pro moderní filologii*, in press.

Breu, W. (2015). Fremder Einfluss in der Grammatik: Der obersorbische und moliseslawische Sprachkontakt im Vergleich. In H. Bartels & S. Wölke (Eds.), *Einflüsse des Deutschen auf die grammatische Struktur slawischer Sprachen* (pp. 9–32). Domowina-Verlag.

Cacchiani, S. (2006). Lexico-Functional Categories in Noah Webster's American Dictionary of the English Language: The Case of Intensifiers. In E. Corino, C. Marello & C. Onesti (Eds.), *Proceedings of the 12th EURALEX International Congress* (pp. 41–47). Edizioni dell'Orso.

Diewald, G. (2011). Pragmaticalization (defined) as grammaticalization of discourse functions. *Linguistics*, 49(2), 365–390.

Diewald, G. (2013). "Same same but different" – Modal particles, discourse markers and the art (and purpose) of categorization. In L. Degand, B. Cornillie & P. Pietrandrea (Eds.), *Discourse Markers and Modal Particles: Categorization and description* (pp. 19–46). John Benjamins Publishing Company.

Degand, L. & Evers-Vermeul, J. (2015). Grammaticalization or pragmaticalization of discourse markers? More than a terminological issue. *Journal of Historical Pragmatics*, 16(1), 59–85.

Duden. Online-Wörterbuch, Cornelsen-Verlag. Retrieved March 12, 2026, from https://www.duden.de/.

DWDS: *Digitales Wörterbuch der deutschen Sprache. Das Wortauskunftssystem zur deutschen Sprache in Geschichte und Gegenwart*. Retrieved January 22, 2026, from https://www.dwds.de.

Faska, H. (Ed.) (1998). *Serbšćina*. Uniwersytet Opolski – Instytut Filologii Polskiej.

Faßke, H. & Michalk, S. (1981). *Grammatik der obersorbischen Schriftsprache der Gegenwart. Morphologie*. Akademie der Wissenschaften der DDR – VEB Domowina Verlag.

Felbr, L. (2024). Aktueller Stand der deutsch-tschechischen Lexikographie am Beispiel von Interjektionen in ausgewählten Wörterbüchern. *Brücken*, 31(2), pp. 113-137.

Filipec, J. & Čermák, F. (1985). *Česká lexikologie*. Academia.

Hanks, P. (2002). Mapping Meaning onto Use. M.-H. Corréard (Ed.), *Lexicography and Natural Language Processing. A Festschrift in Honour of B. T. S. Atkins* (pp. 156-198). Euralex.

Hoekstra, E. (2010). Grammatical information in dictionaries. In A. Dykstra & T. Schoonheim (Eds.), *Proceedings of the XIV Euralex International Congress* (pp. 1007-1012). Fryske Akademy.

Hornjoserbsce.de. Retrieved March 23, 2026, from https://hornjoserbsce.de/korpus/zorla/.

Serbski Institut Bautzen/Budyšin. HOTKO: Upper Sorbian text corpus, version 2 from 6 Mar 2021. ÚČNK FF UK. Retrieved from https://www.korpus.cz, https://wiki.korpus.cz/doku.php/cnk:hotko.

Kral, J. (1927). *Serbsko-němski słownik hornjołužiskeje rěče*. Maćica Serbska.

Mudra, J. & Petr, J. (1989). *Učebnice lužické srbštiny*. Státní pedagogické nakladatelství.

Mudra, J., Brankačkec, K., Škrabal, M. & Vondřička, P. (2023–). *Mudra 2.0: Hornjoserbsko-češki rosćacy słownik online*. Slovanský ústav AV ČR, v. v. i./Ústav lingvistiky FF UK. https://mudra.korpus.cz/.

Nevalainen, T. & Rissanen, M. (2002). Fairly pretty or pretty fair? On the development and grammaticalization of English downtoners. *Language Sciences* 24(3–4), 359–380.

Pavić Pintarić, A. & Frleta, Z. (2014). Upwards Intensifiers in the English, German and Croatian Language. *Journal for Foreign Languages* 6, 31–48.

Pful, K. B. (1866). *Łužiski serbski słownik*. Maćica Serbska.

Rosen, A., Šimčík, B., Vavřín, M. & Zasina, A. J. (2024). *Korpus InterCorp – hornolužická srbština*, verze 16ud ze 17. 9. 2024. Retrieved May 1, 2026, from http://www.korpus.cz/.

Rozumko, A. (2016). Linguistic Concepts across Languages: The Category of Epistemic Adverbs in English and Polish. *Yearbook of the Poznan Linguistic Meeting*, 2(1), 195–214.

Soblexx, online dictionary. Retrieved Jan 23, 2026, from https://soblexx.de.

SSČ: Filipec, J. (Ed.) (1978). *Slovník spisovné češtiny pro školu a veřejnost*. Academia.

SSJČ: Havránek, B. (Ed.) (1960–1971). *Slovník spisovného jazyka českého*. ČSAV.

Stone, G. (2002). *Hornjoserbsko-jendźelski słownik – Upper Sorbian-English Dictionary*. LND.

Strauch, M. & Wornar, E. (2007). *English-Upper Sorbian learner´s dictionary*. LND.

Šindlerová, J. & Štěpánková, B. (2021). Between adverbs and particles: A corpus study of selected intensifiers. *Jazykovedný časopis*, 72(2), 444–453.

Škrabal, M. & Brankačkec, K. (2021). Mudra's Upper Sorbian-Czech dictionary – what can be done about this lexicographic "posthumous child"? In I. Kosem, M. Cukr, M. Jakubíček, J. Kallas, S. Krek & C. Tiberius (Eds.), *Electronic lexicography in the 21st century (eLex 2021): Post-editing lexicography* (pp. 325–338). Lexical Computing CZ.

Štěpánková, B., Poláková, L., Šindlerová, J. & Novák, M. (2024). What Can Dictionaries Tell Us About Pragmatic Markers – Building the Lexicon of Epistemic and Evidential Markers in Czech. In K. Š. Despot, A. Ostroški Anić & I. Brač (Eds.), *Proceedings of the XXI EURALEX* (pp. 728–741). Institut za hrvatski jezik.

Traugott, E. C. (1989). On the rise of epistemic meanings in English: An example of subjectification in semantic change. *Language*, 65(1), 31–55.

Trofimovič, K. K. (1974). *Hornjoserbsko-ruski słownik*. LND.

Contact information

Katja Brankačkec
Institute of Slavonic Studies of the Czech Academy of Sciences
katja.brankackec@slu.cas.cz

Barbora Štěpánková
Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University
stepankova@ufal.mff.cuni.cz

Acknowledgements

The research has been supported by Ministry of Education, Youth and Sports of the Czech Republic under the LINDAT/CLARIAH-CZ project (LM2023062), as well as by the Foundation of the Sorbian People (Założba za serbski lud) under the project *Finishing the Upper Sorbian-Czech dictionary Mudra 2.0* and as part of the programme *AI: Artificial Intelligence for Science and Society – AI in bilingual lexicography. Possibilities and Limitations*.

Polina Iaroshenko & Natalia Loukachevitch

MORPHOLOGICAL FORMS IN EMOTION LEXICONS

Evidence from LLMs

Abstract Emotion expression at the lexical level has been extensively studied, whilst considerably less research addresses emotion expression through morphological devices. This study demonstrates that different morphological forms of the same lexical unit may convey different emotions, as revealed through experiments with large language models (LLMs). Material was drawn from RusEmoLex, a Russian emotion lexicon comprising 1,024 lexical units across five emotion categories (joy, sadness, anger, fear, surprise). For 473 verbs, 12 morphological forms were generated (5,676 in total); for 336 nouns, 6 case forms were generated (2,016 in total). Eight LLMs (7B–671B parameters) classified each form. Performance was evaluated using weighted F1-score. Results reveal substantial asymmetry between word classes. Noun case forms showed minimal morphological influence (mean difference: 3.6%), whilst verb forms demonstrated considerably greater variation (11.5%). Past tense feminine forms consistently achieved the best performance across all models; imperative forms systematically underperformed. A dominant error pattern emerged: sadness-labelled verbs in imperative forms were frequently misclassified as anger. Semantic analysis suggests this pattern may arise from interaction between grammatical form and lexical semantics, with imperative forms potentially amplifying violence-associated meanings in agentive verbs. Findings indicate that incorporating morphological variants may enhance emotion lexicon effectiveness for morphologically rich languages.

Keywords Emotions; large language models; lexicon; natural language processing

1 Introduction

In computational linguistics, the task of emotion analysis has undergone significant development, encompassing not only emotion recognition but also emotion understanding and even imitation. This trend reflects the increasing demand for AI-based dialogue systems in diverse fields, including education and healthcare (Cavicchio, 2025, pp. 5–13). Consequently, a substantial body of research has been devoted to studying the emotional intelligence of large language models (LLMs) and their capacity to reproduce human emotional reactions (see, for example, Huang et al., 2024; Dalal et al., 2025).

Emotion lexicons—lists of words and expressions associated with different emotional categories (Cavicchio, 2025, p. 34)—continue to be an important tool within this field (see, for example, Mohammad, 2023). Describing and systematising emotional vocabulary can contribute to improving the performance of artificial intelligence systems. Emotional lexis, in particular, can function as a marker indicating that a particular emotion is being expressed in a text.

Considerable research has focused on studying the expression of emotions at the lexical level. At the same time, significantly fewer studies address the connection between emotions and morphology. It has been noted that relatively little research has explored the means of expressing emotional content through morphological devices across languages, although a connection with various morphological processes has been documented for most basic emotions (Ponsonnet & Vuillermet, 2018). For example, diminutive and augmentative suffixes can influence the emotional evaluation of a word (see, e.g., Lewandowska-Tomaszczyk & Wilson, 2023; Hinojosa et al., 2022). Recent research on emotion classification with encoder models (BERT-based architectures) has demonstrated that morphological form can systematically influence

model predictions (Iaroshenko, Petrova & Loukachevitch, 2026). Additionally, certain morphological features serve as markers for detecting verbal aggression and incitement to violence, including imperative mood and second-person verb forms (Burnap & Williams, 2015; Etaywe, 2022).

The idea of a connection between the morphological characteristics of a word and emotions suggests that expanding emotion lexicons by incorporating various morphological forms, rather than lemmas alone as is typically done, may enhance the effectiveness of lexicon usage, as different word forms may potentially express different emotions.

To test this hypothesis, an experiment was conducted in which LLMs were tasked with classifying various morphological forms of Russian verbs and nouns into emotion categories. Russian was selected due to its relatively rich morphology. The research material was drawn from RusEmoLex (Iaroshenko & Loukachevitch, 2025), a Russian emotion lexicon created by integrating diverse emotional vocabulary resources, in which lexical units are represented as lemmas and categorised by emotion. Verbs and nouns were extracted from the lexicon, and various morphological forms were generated from these lemmas for experimental purposes. The following research question was formulated: will word form be a significant factor in determining the emotional class of a word by LLMs? The models were tasked with multi-class classification, and performance was evaluated using the F-measure.

2 Data

The research material was drawn from RusEmoLex (Iaroshenko & Loukachevitch, 2025), a Russian emotion lexicon created by integrating diverse emotional vocabulary resources: lexicographic sources (dictionaries and thesauri), corpus data, survey data reflecting native speakers' word–emotion associations, and machine learning datasets. RusEmoLex comprises 1,024 lexical units across different parts of speech, with each word assigned to one of five emotional categories: joy, sadness, anger, fear, and surprise.

For the present experiment, two word classes were selected: verbs and nouns. Table 1 presents examples of lexical entries from RusEmoLex with English translations.

TABLE 1 Examples from RusEmoLex

Word	Part of Speech	Emotion Class
bespokoit'sya (to worry)	Verb	Fear
bespokoystvo (worry)	Noun	Fear
voskhishchat'sya (to admire)	Verb	Joy
voskhishcheniye (admiration)	Noun	Joy
zlit'sya (to be angry)	Verb	Anger
zlost' (anger)	Noun	Anger
udivlyat'sya (to be surprised)	Verb	Surprise
udivleniye (surprise)	Noun	Surprise
grustit' (to be sad)	Verb	Sadness
grust' (sadness)	Noun	Sadness

Table 2 presents the distribution of lexical items selected for the experiment by emotion class.

TABLE 2 Distribution of Verbs and Nouns by Emotion Class

Part of Speech	Total Items	Joy	Sadness	Anger	Fear	Surprise
Nouns	336	106	115	67	34	14
Verbs	473	107	181	81	78	26

2.1 Nouns

All 336 nouns from the lexicon were included. Each noun was inflected across six case forms: nominative, genitive, dative, accusative, instrumental, and locative, yielding 6 forms per noun and 2,016 forms overall.

2.2 Verbs

All 473 verbs from the lexicon were used in the experiment. Morphological forms were generated encompassing both the imperative and indicative moods. The indicative mood forms included past and present tenses across all applicable persons and numbers, with gender distinctions generated where relevant (e.g., masculine and feminine forms for past tense singular). The complete set of verb forms comprised: imperative (singular and plural), infinitive, past tense (masculine, feminine, and plural), and present tense (all three persons in both singular and plural). In total, 12 morphological forms were generated per verb, yielding 5,676 forms overall.

3 Methods

During the experiment, verbs and nouns in various morphological forms were sequentially input into the LLMs, one lexical unit at a time. Each model was required to assign an emotion label to the given word based on a Russian-language prompt. The English translation of the prompt is provided below:

> Your task is to identify the emotion associated with the given word. Emotions: Joy, Fear, Surprise, Sadness, Anger. Your answer should consist of a single word indicating the relevant emotion class. No additional comments are required.

Eight LLMs were employed in the experiment, ranging from models with a substantial number of parameters, such as DeepSeek-V3 (671 billion parameters), to medium-sized models (32 billion parameters) and smaller models (7–8 billion parameters). The tested models also varied in origin: some were international models, while others were developed in Russia or specifically adapted for the Russian language. Table 3 presents the characteristics of the models used in the experiment.

TABLE 3 Characteristics of LLMs Used in the Experiment

Model	Size	Origin
DeepSeek-V3	671B	International
Qwen3-235B-A22B-Instruct-2507	235B	International
Llama-3	70B	International

TABLE 3 (*CONTINUED*)

Model	Size	Origin
RuadaptQwen2.5-32B-Pro-Beta	32B	Russian-adapted
T-pro-it-1.0	32B	Russian-adapted
YandexGPT-5 Lite	8B	Russian-developed
RuadaptQwen2.5-7B-Lite-Beta	7B	Russian-adapted
T-lite-it-1.0	7B	Russian-adapted

The models were configured with temperature set to 0.2 to ensure deterministic and consistent outputs. The responses generated by the models were compared with the gold-standard emotion labels from RusEmoLex, and performance was evaluated using the weighted F1-score to account for class imbalance in the emotion categories.

4 Results: Classification Performance Across Morphological Forms

The experiment revealed substantial differences between nouns and verbs in terms of morphological influence on emotion classification. The results for each part of speech are presented below. For evaluation purposes, the nominative case served as the baseline for nouns, and the infinitive as the baseline for verbs, as lexical units in emotion lexicons are typically represented in these lemma forms.

4.1 Classification of Noun Forms

For the majority of models, the morphological form of a noun exerted minimal influence on classification performance. The results of noun classification are presented in Table 4.

TABLE 4 Classification Performance for Nouns Across Different Case Forms

Model	Baseline (nominative)	F1 best	F1 worst
DeepSeek-V3	0.83	0.86 genitive	0.83 nominative accusative instrumental
Qwen3-235B-A22B-Instruct-2507	0.82	0.82 nominative accusative instrumental	0.81 genitive dative locative
Llama-3	0.83	0.84 instrumental	0.82 accusative genitive dative locative
RuadaptQwen2.5-32B-Pro-Beta	0.80	0.80 nominative	0.74 dative

TABLE 4 (*CONTINUED*)

Model	Baseline (nominative)	F1 best	F1 worst
T-pro-it-1.0	0.83	0.83 nominative accusative genitive instrumental	0.81 dative
YandexGPT-5 Lite	0.78	0.79 accusative	0.70 dative
RuadaptQwen2.5-7B-Lite-Beta	0.80	0.80 nominative locative	0.78 accusative dative
T-lite-it-1.0	0.81	0.81 nominative accusative	0.77 dative locative

The difference between the best- and worst-performing forms ranged from 1% (Qwen3-235B) to 9% (YandexGPT-5 Lite), with a mean of 3.6% across all models. It should be noted that the maximum value of 9% is atypical: for seven of the eight models, the difference did not exceed 6%. The advantage of the best-performing form over the baseline (nominative case) was also modest: for five of the eight models, the nominative case itself was among the best-performing forms, while for the remaining three models, the improvement ranged from 1% (YandexGPT-5 Lite) to 3% (DeepSeek-V3 and Llama-3).

A characteristic feature of the noun results is the frequent convergence of F1 scores across different case forms. For six of the eight models, several cases demonstrated identical performance values. This is most pronounced in Qwen3-235B and T-pro, where three and four cases, respectively, exhibited identical scores among both the best- and worst-performing forms.

The only case that systematically showed lower performance was the dative: it appeared among the worst-performing forms for six of the eight models (either as the sole worst-performing form or as one of several forms with identical scores).

4.2 Classification of Verb Forms

In contrast to nouns, the morphological form of a verb exerted substantial influence on classification performance. The results of verb classification are presented in Table 5.

TABLE 5 Classification Performance for Verbs Across Different Morphological Forms

Model	Baseline (infinitive)	F1 best	F1 worst
DeepSeek-V3	0.79	0.85 past_femn	0.72 impr_sing
Qwen3-235B-A22B-Instruct-2507	0.85	0.89 past_femn	0.80 impr_plur
Llama-3	0.78	0.83 past_femn	0.69 impr_plur

TABLE 5 (*CONTINUED*)

Model	Baseline (infinitive)	F1 best	F1 worst
RuadaptQwen2.5-32B-Pro-Beta	0.73	0.77 past_femn	0.66 impr_plur
T-pro-it-1.0	0.83	0.88 past_femn	0.77 impr_plur
YandexGPT-5 Lite	0.73	0.77 past_femn	0.66 impr_plur
RuadaptQwen2.5-7B-Lite-Beta	0.66	0.71 past_femn	0.62 impr_plur pres_sing_2per pres_plur_3per
T-lite-it-1.0	0.79	0.85 past_femn	0.72 impr_sing

The difference between the best- and worst-performing forms ranged from 9% (RuadaptQwen2.5-7B) to 14% (Llama-3), with a mean of 11.5% across all models—3.2 times greater than for nouns. Moreover, unlike nouns where multiple forms frequently showed identical F1 scores, verb forms predominantly produced distinct performance values, with most morphological forms yielding different classification results.

The past tense feminine form (past_femn) consistently yielded the highest classification performance across all eight models, surpassing the baseline (infinitive) by 4–6%. In contrast, imperative forms demonstrated the poorest performance: the plural imperative (impr_plur) was the worst-performing form for six models, whilst the singular imperative (impr_sing) showed the lowest scores for two models. Only for RuadaptQwen2.5-7B did three forms (impr_plur, pres_sing_2per, pres_plur_3per) exhibit equally low performance.

5 Semantic Patterns in Verb Form Misclassification

Given that verb forms demonstrated substantially greater F1 score variation than nouns, we conducted a detailed analysis of verb classification patterns. Figure 1 presents confusion matrices for Qwen3-235B-A22B-Instruct-2507, the best-performing model overall. The left matrix represents the verb form that best aligns with RusEmoLex labels (past_femn), while the right matrix represents the poorest-performing form (impr_plur).

The confusion matrix reveals the most pronounced direction of emotional shift: in the past tense feminine form, the model correctly classifies 86.7% of Sadness cases; however, in the imperative, accuracy drops to 59.1%, primarily due to Sadness-class words being assigned to the Anger category (34.1%).

This tendency for Sadness to shift towards Anger in the plural imperative (or, less frequently, the singular imperative) is the dominant pattern across all models in the experiment. To gain a more detailed understanding of the possible semantic nature of this pattern, verbs that were most frequently misclassified as Anger in the imperative mood were examined. The analysis focused on verbs with the true label Sadness that were classified as Anger in imperative form by all eight LLMs (37 cases), seven LLMs (24 cases), or six LLMs (23 cases). Table 6 presents examples of verbs that were unanimously classified as Anger in the imperative by all eight models.

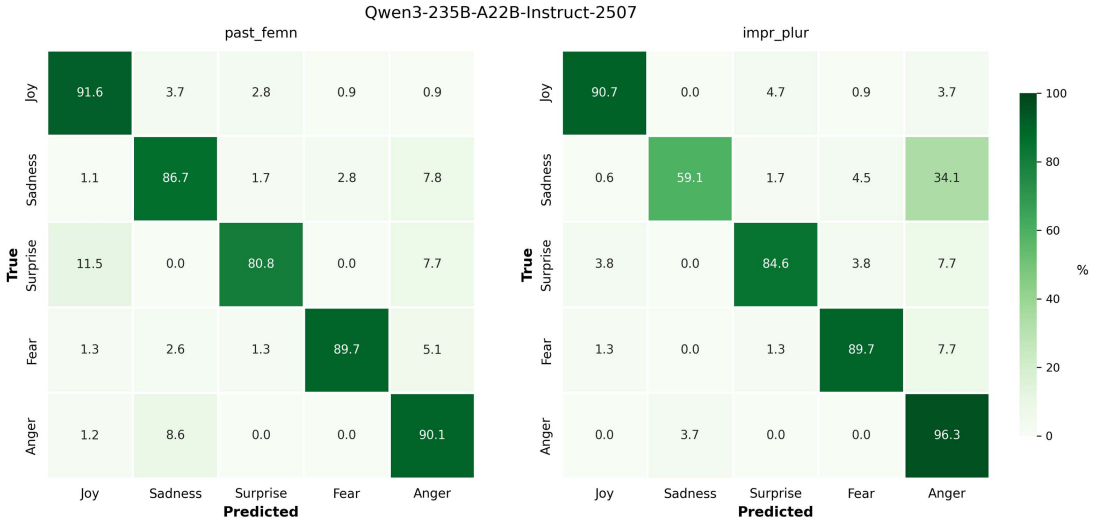


FIGURE 1 Confusion matrices for Qwen3-235B-A22B-Instruct-2507

TABLE 6 Verbs Classified as Anger in Imperative Forms by All Models

Verb	Form
vymatyvat' (to exhaust)	impr_sing, impr_plur
obidet' (to offend)	impr_sing, impr_plur
ubit' (to kill)	impr_sing, impr_plur
muchit' (to torment)	impr_plur
slomit' (to break)	impr_plur

Semantic analysis suggests a pattern: many verbs misclassified by the majority of models appear to denote active destructive actions or the infliction of suffering: *terzat' ('to torment'), zatravit' ('to hound'), razdavit' ('to crush'), slomit' ('to break'), davit' ('to press'), zhech' ('to burn')*. These verbs are characterised by high agentivity—they typically denote deliberate, agent-controlled actions—and govern a direct object.

For comparison, we can examine verbs that were correctly classified in imperative form by all eight models (12 cases total): *zagrustit' ('to become sad'), opechalit'sya ('to become saddened'), otchayat'sya ('to despair'), skuchat' ('to be bored'), khandrit' ('to be melancholic')*. These verbs directly express sadness without semantic components related to physical violence or causing harm. Broadening the analysis to include verbs for which the majority of models assigned the correct label (cases where at least five of eight LLMs classified correctly), we also observe verbs denoting physical manifestations of sadness: *rydat' ('to sob'), poplakat' ('to cry'), khmurit'sya ('to frown')*.

These findings suggest that the systematic classification pattern—assigning sadness-labelled verbs to anger in imperative forms—may arise from an interaction between grammatical form and lexical semantics. Specifically, verbs with semantic components related to inflicting physical or psychological harm appear to be particularly affected, with the imperative form amplifying these violence-associated meanings.

The consistent pattern observed across LLMs suggests that for certain verbs, assigning the anger label in imperative forms may reflect underlying semantic properties rather than

classification error. These verbs typically denote calls for destructive actions directed at an object, and in such contexts, imperative forms may legitimately function as anger markers rather than sadness indicators.

Regarding the potential expansion of emotion lexicons to include morphological variants, it should be noted that automatic label reassignment based solely on LLM predictions would be premature. A systematic approach would require, first, consensus across multiple models to identify candidate verbs, and second, manual semantic verification, as the appropriateness of anger or sadness labels depends on the specific lexical semantics of each verb. Such systematically identified cases, however, could serve as valuable material for enriching emotion lexicons by capturing form-dependent emotion associations that are not evident at the lemma level.

6 Conclusion

This study investigated whether morphological form influences emotion classification by large language models, using verb and noun forms from RusEmoLex. Results suggest substantial differences between word classes. Noun case forms showed minimal performance variation, while verb morphology demonstrated considerably greater impact—approximately three times more variation than nouns.

Two systematic patterns emerged for verbs. Past tense forms consistently outperformed the infinitive baseline across all models, while imperative forms systematically underperformed. A pronounced misclassification tendency was observed: sadness-labelled verbs in imperative forms were frequently assigned to the anger class.

Preliminary semantic analysis suggests this pattern may relate to an interaction between grammatical form and lexical semantics. Verbs most frequently misclassified appear to denote agentive, destructive actions, while correctly classified verbs tend to express sadness states or physical manifestations without violence-related semantic components. The imperative form may potentially amplify existing violence-associated meanings.

These findings indicate that for morphologically rich languages, incorporating morphological variants in emotion lexicons may warrant consideration, particularly for verbs. However, morphological influence appears part-of-speech dependent and may interact with lexical semantics in complex ways.

Further research should examine whether similar patterns occur in other morphologically rich languages and across different parts of speech. Understanding how morphology, semantics, and emotion interact may hold value for both theoretical linguistics and applied natural language processing, particularly in the domains of emotion recognition and emotion understanding.

References

Burnap, P. & Williams, M. (2015). Cyber hate speech on Twitter: An application of machine classification and statistical modeling for policy and decision making. *Policy & Internet*, 7(2), 85–103. https://doi.org/10.1002/poi3.85

Cavicchio, F. (2025). *Emotion detection in natural language processing*. Springer.

Dalal, D., Negi, G. & Picca, D. (2025). LLMs and emotional intelligence: Evaluating emotional understanding through psychometric tools. In *Proceedings of the 33rd ACM Conference on User Modeling,*

Adaptation and Personalization (UMAP '25) (pp. 323–328). ACM. https://doi.org/10.1145/3699682.3728315

Etaywe, A. (2022). Exploring the grammar of othering and antagonism as enacted in terrorist discourse: Verbal aggression in service of radicalisation. *Humanities and Social Sciences Communications*, 9, Article 177. https://doi.org/10.1057/s41599-022-01178-5

Hinojosa, J. A., Haro, J., Calvillo-Torres, R., González-Arias, L., Poch, C. & Ferré, P. (2022). I want it small or, rather, give me a bunch: The role of evaluative morphology on the assessment of the emotional properties of words. *Cognition & Emotion*, 36(6), 1203–1210. https://doi.org/10.1080/02699931.2022.2093840

Huang, J., Lam, M. H., Li, E. J., Ren, S., Wang, W., Jiao, W., Tu, Z. & Lyu, M. R. (2024). Apathetic or empathetic? Evaluating LLMs' emotional alignments with humans. *Advances in Neural Information Processing Systems*, 37, 97053–97087.

Iaroshenko, P. V. & Loukachevitch, N. V. (2025). RusEmoLex: Russian emotion lexicon. *Russian Journal of Linguistics*, 29(3), 659–687. https://doi.org/10.22363/2687-0088-44439

Iaroshenko P. V., Petrova A. A. & Loukachevitch N. (2026). V. Morphology and Emotion Recognition: How Verb Form Affects Language Model Predictions. In N. V. Lukashevich, S. G. Tatevosov & V. P. Selegey (Eds.), *Computational Linguistics and Intellectual Technologies: Proceedings of the International Conference Dialogue* 2026 (pp. 134–143).

Lewandowska-Tomaszczyk, B., & Wilson, P. A. (2023). Morphology and emotion. In G. L. Schiewer, J. Altarriba, & B. C. Ng (Eds.), *Language and emotion: An international handbook* (pp. 423–441). De Gruyter Mouton.

Mohammad, S. M. (2023). Best practices in the creation and use of emotion lexicons. In A. Vlachos & I. Augenstein (Eds.), *Findings of the Association for Computational Linguistics: EACL 2023* (pp. 1825–1836). https://doi.org/10.18653/v1/2023.findings-eacl.136

Ponsonnet, M. & Vuillermet, M. (2018). Morphology and emotions: A preliminary typology. *Studies in Language*, 42(1), 1–16.

Contact information

Polina Iaroshenko
Lomonosov Moscow State University
polina.iaroshenko@yandex.ru

Natalia Loukachevitch
Lomonosov Moscow State University
louk_nat@mail.ru

Acknowledgements

The research was supported by the Russian Science Foundation, project No. 25-11-00191, https://rscf.ru/project/25-11-00191/. The research was carried out using the MSU-270 supercomputer of Lomonosov Moscow State University.

Boris Kern

WORD-FORMATION CHAINS IN SLOVENIAN

Evaluating the Effectiveness of AI

Abstract This study investigates word-formation chains, i.e. sequences of derivationally related words in which each derivative serves as the base for the next. Due to their rich morphemic structure, such chains are particularly complex in Slavic languages. Their analysis is relevant for understanding word-formation and semantic potential, morphotactics, and lexicographic criteria for vocabulary inclusion. The research builds on the project *Formant Combinatorics in Slovenian* (2022–2025), which applied automatic extraction methods to data from the *Word-Family Dictionary of the Slovene Language* (BSSJ). Abstract suffix-chain patterns derived from this resource were extended to larger lexical datasets. Preliminary findings showed limited success of automatic extraction methods (24% for the most frequent suffix chain), whereas AI achieved higher accuracy (59%). However, performance declined markedly with increasing chain complexity. The present study aims to evaluate the effectiveness of ChatGPT 5.5 in identifying word-formation chains with nominal, verbal and adjectival bases. It focuses on the impact of chain complexity, the role of part-of-speech within chains, and problematic suffixes and zero-formant derivation patterns. Particular challenges include derivatives without overt suffixes and cases where zero-formant nouns are misclassified as unmotivated words. The analysis examines suffix chains of varying complexity and compare AI performance with previous automatic extraction approaches, with the goal of improving both extraction methods and AI prompting strategies.

Keywords Multistage word formation; morphotactics; formants; AI

1 Introduction

The rich derivational morphology of Slavic languages enables the formation of extensive networks of related lexemes through successive derivational processes. For example, the Slovenian word-formation chain *mir* 'peace' → *miriti* 'to calm' → *pomiriti* 'to calm down' → *pomirjen* 'calmed' → *pomirjenost* 'calmness' illustrates a four-stage derivational process.[1] Such chains consist of a sequence of lexemes in which each derivative serves as the base for the next derivative in the sequence until the derivational potential of a particular lexeme is exhausted (Skarżyński 2003; Stramljič Breznik, 2004; Kern, 2008, 2017, 2026). This complexity has given rise to the methodology of multistage word formation, which emerged as a methodological response to the limitations of traditional word-formation studies, which had largely focused on the binary relation between a motivating and a motivated word. Its aim is to capture broader derivational structures: all direct and indirect derivatives organised around a simplex lexical base (Burkacka, 2001, p. 11; Jadacka, 2003, p. 29). This hierarchically organised unit is referred to as a word-formation nest.

In the historical outline of the methodology, the development of the approach is linked above all to Russian scholarship, particularly to A. N. Tikhonov. In 1978, Tikhonov published the school word-formation dictionary of Russian, conceived as a pilot work, followed in 1985 by the two-volume *Word-Formation Dictionary of Russian*, which classified 144,808 words into

[1] The highest word-formation stage in Slovenian is the seventh stage, for example: *tip* 'touch' > *tipati* 'to touch' → *tipka* 'key, especially on a keyboard or typewriter' → *tipkati* 'to type; to press keys' → *pretipkati* 'to retype' → *pretipkovati* 'to retype repeatedly or systematically' → *pretipkovalec* 'retypist' → *pretipkovalčev* 'belonging to or relating to a retypist' (Kern, 2017).

18,118 word-formation nests. This work also defined the basic concepts of multistage word formation, including the word-formation nest,[2] word-formation chain, ect. (Tikhonov, 1985; Olejniczak, 2003, pp. 12–13).

In Polish linguistics, the development of multistage word formation was driven both by the need to study larger derivational units, rather than only the relation between a motivating and a motivated word, and by the achievements of previous researches (Olejniczak, 2003, pp. 12–28). The first plans for a Polish word-formation nest dictionary began in 1977 at the Institute of Polish Language, University of Warsaw, on the initiative of Jadwiga Puzynina. Eleven years later, under the direction of Hanna Jadacka, the pilot volume *Zeszyt próbny "Słownika gniazd słowotwórczych współczesnego języka ogólnopolskiego"* was published. Although modelled on Tikhonov's dictionary, the Polish approach introduced several methodological refinements (Jadacka 2003; Olejniczak, 2003, pp. 12–28). This tradition subsequently developed into an extensive research programme, including major lexicographic works such as *Słownik gniazd słowotwórczych współczesnego języka ogólnopolskiego* (2001, 2004).

In Slovenian linguistics, systematic research on multistage derivation became possible with the publication of the *Word-Family Dictionary of the Slovenian Language: Pilot Volume for Headwords Beginning with B* (Stramljič Breznik, 2004) and was further developed through subsequent theoretical and empirical studies (Kern, 2017, 2024, 2026).[3] This tradition was further expanded by the most extensive multilingual monograph in the field of multistage word formation, published in 2026 and edited by Boris Kern, which brings together contributions by 42 researchers and covers 14 languages: 13 Slavic languages and Slovenian Sign Language (Kern, 2026).

The paper is structured as follows. Section 2 outlines the morphotactic framework and previous approaches to the automatic extraction of word-formation chains. Section 3 presents the data, methodology and analysis of ChatGPT 5.5 performance, while Section 4 summarises the main findings.

2 Morphotactics

One of the most important research perspectives enabled by multistage word formation is *morphotactics*. In the context of multistage word formation, morphotactics refers to the combinability and distribution of formants within word-formation chains. It investigates which suffixes, prefixes, and other derivational morphemes can co-occur, in what order they appear, and which formal, semantic, or systemic constraints govern their combination (Burkacka, 2012; Kern, 2020; Kern & Stramljič Breznik, 2025). Morphotactic analysis therefore shifts attention from individual derivational operations to the structural organisation of derivational systems. In Slavic languages, morphotactic research is particularly relevant in the analysis of suffix combinatorics. This can be illustrated by the word-formation chain based on *mir* 'peace', already mentioned in the introduction, which displays four-degree derivation and a three-suffix chain:

[2] A **word-formation nest** is a hierarchically organised set of words according to their word-formation stage. These words are formed from the same root in the word-formation base and, in a synchronic analysis, often also from the same dictionary meaning. At the beginning of a word-formation nest is a simplex, a non-derived word, followed by all its direct and indirect derivatives; it is the basic unit for organising and systematising directly and indirectly derived lexemes formed from the same root in the word-formation base (Kern, 2010, 2017, 2026).

[3] The foundations of this perspective can already be observed in the work of Miklošič (1875) and later in the studies of Vidovič Muha (1988).

-iti + -n + -ost. Morphotactic research, however, is also possible at the level of prefixes. The word-formation chain *znati* 'to know' → *poznati* 'to know, to be acquainted with' → *prepoznati* 'to recognise' → *prepoznan* 'recognised' → *neprepoznan* 'unrecognised' displays four-degree derivation and a three-prefix chain: *ne- + pre- + po-*.

A morphotactic perspective reveals that derivational formants do not combine randomly. Certain combinations occur frequently and productively, whereas others are restricted or confined to the periphery of the system. For instance, the Slovenian word-formation chain *barva* 'colour' → *barven* 'colourful' → *barvnost* 'colourfulness' exhibits the highly productive suffix chain *-en + -ost*, similary the chain *beton* 'concrete' → *betonirati* 'to concrete' → *betoniranje* 'concreting' illustrates the productive combination *-irati + -anje*. Whereas the chain *bog* 'god' → *pobožen* 'pious' → *pobožnjak* 'religious hypocrite' → *pobožnjakar* 'religious hypocrite, with an additional expressive or pejorative nuance' → *pobožnjakarica* 'female religious hypocrite, with an additional expressive or pejorative nuance' contains the considerably rarer four-suffix chain *-en + -jak + -ar + -ica* (Kern, 2026). Such examples show that morphotactics operates at the level of recurrent formant patterns rather than isolated derivatives.

2.1 Morphotactics of Slovenian

In Slovenian linguistics, awareness of morpheme sequencing and multistage derivation is by no means new. It can already be found in Franz Miklošič's *Comparative Grammar of the Slavic Languages*, especially in the second volume, *The Study of Stem Formation* (1875), where suffixes are analysed both phonologically and semantically. By carefully distinguishing the functions of individual morphemes within the word, Miklošič also implicitly differentiated between word-formation stages. In this sense, later morphotactic research may be seen not as an entirely new concern, but as a systematic continuation of earlier insights into the internal organisation of complex derivatives.

In Slovenian linguistics, morphotactic questions were also anticipated before morphotactics had become established as a separate research perspective. An important point of reference is Ada Vidovič Muha's discussion of the predictability of further derivation in *Slovenian Syntactic Word Formation Based on Examples of Compounds* (1988). In her analysis of adjectival and nominal derivatives, she observed that adjectives formed with the suffix *-ski* can, in principle, motivate nouns denoting a property. At the same time, she identified an important restriction: nouns denoting properties are not normally formed from relational adjectives in *-ni* when these are based on concrete nouns, as shown by impossible formations such as **cestnost* 'road-ness' and **miznost* 'table-ness' (Vidovič Muha, 1988, p. 104). These observations are relevant for morphotactics because they show that suffixal chains are not freely combinable, but are constrained by the semantic and categorial properties of the base.

As a distinct field of investigation, however, morphotactic research in Slovenian word formation is relatively recent. Its first systematic application to Slovenian is presented in Kern's article *Combinatorics of Suffixes in Word-Formation Nests of Verbs of Sensory Perception* (Kern, 2020), which built on the material and methodology developed in the monograph *Multistage Word Formation: The Case of Verbs of Sensory Perception* (2017).[4] This line of research was then

[4] The research included 34 word-formation nests based on simplex verbs of sensory perception and confirmed the highly ramified nature of the Slovenian word-formation mechanism, which is also reflected in the diverse combinatorics of suffixes. Among the 751 derivatives analysed, 76 different suffixal chains were identified. (Kern, 2020, pp. 70, 76–77)

expanded within the project *Combinatorics of Word-Formation Affixes in Slovenian* (hereafter KOBOS) (2021–2025), where the systematic analysis of a suffixal chain became one of the central research aims. An important model for Slovenian morphotactic research was provided in particular by the Polish studies of Iwona Burkacka, especially her work on suffix combinatorics in Polish noun-based derivation (Burkacka, 2012). Alongside Kern's 2017 monograph, Irena Stramljič Breznik's *Word-Family Dictionary of the Slovenian Language: Pilot Volume for Headwords Beginning with B* (2004; hereafter BSSJ) provided an essential empirical basis for the development of Slovenian morphotactics.

The KOBOS project substantially advanced morphotactic research in Slovenian and resulted in 22 publications. These studies examined the combinatorics of affixes in a wide range of multistage derivatives, including derivatives based on verbs, nouns, adjectives, adverbs and interjections, as well as more specific topics such as feminatives in noun-based word-formation chains. The project also addressed methodological and lexicographic issues, including criteria for dictionary inclusion, the terminology of multistage word formation, the automatic identification of word-formation chains and improvements in automatic morphemisation for Slovenian. These studies have shown that morphotactic analyses are relevant not only to theoretical morphology but also to lexicography, terminology studies, corpus linguistics, and natural language processing. (Erjavec et al., 2023; Kern, 2024; Kern, 2026).

2.2 Automatic Extraction of Word-Formation Chains

Recent advances in computational linguistics and artificial intelligence have opened new possibilities for the automatic identification of derivational relationships. Automatic extraction methods based on formal patterns and large language models can generate large numbers of potential word-formation chains, but their effectiveness remains largely unexplored in the context of complex multistage word formation (Pranjić, Pollak, 2026).

Within the KOBOS project, the automatic extraction of word-formation chains was based on the *BSSJ Dictionary* by Irena Stramljič Breznik. As a first step, the dictionary data were converted into TEI Lex0 and subsequently into TSV format in order to make the material suitable for computational processing. The resulting dataset included information on the word family, lemma, word-formation stage, word-formation chain, morphemes and morpheme types (prefixes, suffixes, affixes) (Erjavec et al., 2023). On the basis of pairs consisting of a simplex word and its derivative, together with morphemic segmentation and part-of-speech annotation, abstract morphemic patterns were then automatically induced, such as VERB: X-*evati* → NOUN:X-*anje*. The procedure focused primarily on suffixal patterns.

These abstracted patterns were subsequently applied on a larger scale to lemmas with part-of-speech annotation from two major lexical/corpus resources, namely Sloleks and metaFida 1.0, as well as to the *Dictionary of Standard Slovenian (Slovar slovenskega knjižnega jezika)*. In this way, the automatic procedure extended the extraction of word-formation chains beyond the BSSJ Dictionary material to the wider vocabulary of Slovenian. However, the quality of the automatically extracted candidates differed considerably between the sources: Sloleks produced substantially less noise than metaFida, a finding that was also confirmed by manual evaluation of a sample of the automatically extracted chains (Erjavec et al., 2023).

A preliminary study illustrates both the potential and the limitations of this approach. For derivatives with the word-formation model S,Ad,S (i.e. a noun-based chain in which a noun motivates an adjective, which in turn motivates a noun) and the suffixal

combinatorics *-en/-ni + -ost*[5]—the most frequent suffixal combination among noun-based multistage derivatives—the tool identified 959 word-formation chain candidates. Of the 959 chain candidates, 233 were judged to be appropriate, corresponding to a success rate of approximately 24%. This figure should be interpreted with some caution, since certain examples allow more than one possible derivational interpretation. The assessment of suitability was based on the inclusion of the relevant derivatives in dictionary sources, excluding purely systemic formations listed in the *Slovenian Normative Guide (Slovenski pravopis)* (2001) without attested corpus use, and/or on the presence of at least one convincing occurrence in the largest corpus of Slovenian, metaFida 1.0.[6]

Against this background, the present study examines the ability of artificial intelligence to identify Slovenian denominal word-formation chains and evaluates the extent to which morphotactic complexity influences performance. Particular attention is devoted to the relationship between chain length and identification accuracy, the role of part-of-speech transitions within derivational chains, and the morphotactic patterns that repeatedly cause difficulties for automatic analysis. By comparing AI-generated chains with manually verified morphotactic data, the study contributes both to the linguistic description of Slovenian derivational morphology and to the development of computational methods for modelling complex word-formation systems.

3 Analysis of Automatically Extracted Word-Formation Chain Candidates

3.1 Source Data for the Analysis of Word-Formation Chains

BSSJ Dictionary serves as the source material for the analysis of word-formation chains examined within the KOBOS project. The BSSJ contains 9,458 derivatives in which different word-formation stages and combinations of suffixes can be observed.

The largest group consists of derivatives with a nominal word-formation base. There are 4,582 such derivatives, representing almost half of all derivatives included in the BSSJ Dictionary and analysed within the KOBOS project (48%). Of these, 1,291 exhibit combinations of two or more suffixes, accounting for 28% of all derivatives with a nominal base. In this group, 719 different suffix chains were identified, meaning that the average number of derivatives per suffix chain is 1.8.

The second-largest group consists of derivatives with a verbal word-formation base. This group comprises 3,348 derivatives, of which 2,108 exhibit combinations of two or more suffixes, corresponding to approximately 63% of all derivatives with a verbal base. Among derivatives with an adjectival word-formation base, a total of 1,145 derivatives were identified, of which 533 exhibit combinations of suffixes; this corresponds to approximately 47%.

These data show that suffix combinations are not evenly distributed across the different types of word-formation bases. Whereas among derivatives with a nominal base such combinations

[5] The automatic procedure could not reliably distinguish between the suffixes *-en* and *-ni*. This distinction is relevant because, in Slovenian, *-en* typically derives qualitative adjectives expressing a property of the referent, while *-ni* derives relational adjectives expressing that something belongs to, is associated with, or is classified by the base noun.

[6] The metaFida 1.0 corpus was selected because it is the most extensive corpus of Slovenian and also includes texts from social media, colloquial language and other less formal registers. A derivative was considered attested in metaFida if it occurred at least once in the corpus.

are observed in slightly more than one quarter of the examples, the proportion is substantially higher among derivatives with a verbal base, where it applies to almost two thirds of the material.

3.2 Criteria for the Selection of Word-Formation Chains

The selection of word-formation chains was based on several criteria. First, only chains with verbal, nominal and adjectival bases were included, since these represent the most productive word-formation bases in Slovenian. The second criterion was the frequency of individual suffixal chains, especially in the case of two-suffix chains. Accordingly, five two-suffix chains were selected, as these are also among the most numerous types identified in previous research (Kern, 2020, 2026; Stramljič Breznik, 2026). In addition, three three-suffix chains, one four-suffix chain and one five-suffix chain were included in the analysis.[7]

A further criterion was the problematic status of certain morphemes within suffixal chains. This is particularly relevant in the case of the zero morpheme, which has proved problematic in earlier research (Kern, 2026; Pranjić, Pollak 2026). Since a derivative formed with a zero morpheme is often shorter than its motivating word, it may appear formally less complex, which poses a considerable challenge for automatic recognition. The selection also took into account the similarity of suffixal chains. This made it possible to examine whether the effectiveness of automatic extraction depends on the part-of-speech category of the word-formation base, for example whether comparable suffixal patterns behave differently when they are based on nouns, verbs or adjectives.

3.3 Use of ChatGPT as an Exploratory Tool for Identifying Word-formation Chain Candidates

In the analysis of automatically extracted word-formation chain candidates, ChatGPT 5.5 was used as an additional exploratory tool for identifying possible suffixal chains. The searches focused on the following patterns, where the initial abbreviation indicates the part of speech of the word-formation base: N = noun, V = verb and Ad = adjective:

1) N + *-en*/*-ni* + *-ost*;
2) V + *-en*/*-ni* + *-ost*;
3) Ad + *-iti* + *-enje*;
4) Ad + *-eti* + *-enje*;
5) V + ø + *-ni*;
6) V + ø + *-en*/*-ni* + *-ost*;
7) N + *-en*/*-ni* + *-jak* + *-ar* + *-ica;*
8) Ad + *-ati* + *-eva-* + *-alen* + *-ica* + *-ški*.

[7] Previous studies on Slovenian have already indicated that the success of automatic identification decreases as chain complexity increases and that certain derivational mechanisms (Pranjić, Pollak, 2026; Kern, 2026).

The candidate word-formation chains were elicited using prompts of the following type: "List 100 examples[8] of derivatives based on a simplex noun from which an adjective in *-en* is formed, followed at the second degree of derivation by a noun in *-ost*, for example: *elita* > *eliten* > *elitnost, sok* > *sočen* > *sočnost*."[9] In the case of verb-based derivatives, the prompt explicitly stated that an intermediate prefixed verb may also occur within the chain. The example *buliti* > *izbuliti* > *izbuljen* > *izbuljenost* was consistently provided in order to guide the extraction of such candidates. For each suffix pattern, ChatGPT 5.5 was queried once using a separate fresh chat session. The prompts followed the same general structure, with only the part-of-speech category, suffix sequence and illustrative examples adapted to the pattern under investigation. An additional instruction was included requiring the proposed derivatives to be attested in actual language use, as preliminary testing showed that ChatGPT occasionally generated non-existent or unattested forms. No follow-up prompts or manual corrections were used during candidate generation. The generated candidates were subsequently evaluated manually according to the criteria described above. The complete set of prompts used in the analysis is available in the accompanying online repository.

The use of ChatGPT 5.5 in this part of the study pursued five main objectives.

1) Since a preliminary study on the automatic extraction of word-formation chains had already been carried out for the pattern N + *-en/-ni* + *-ost*, the aim was to determine whether the results obtained with ChatGPT 5.5 differed from those previously obtained with ChatGPT 5.2 (Kern 2026).

2) The analysis sought to establish whether the part-of-speech category of the word-formation base affects the proportion of suitable candidate word-formation chains identified by ChatGPT.

3) The study aimed to determine whether structurally similar suffix chains yield comparable levels of effectiveness when searched for using ChatGPT.

4) The study examined whether the presence of a zero suffix within the suffix sequence influences the effectiveness of candidate extraction.

5) It assessed the extent to which the complexity of suffix sequences affects the effectiveness of identifying candidate word-formation chains using ChatGPT.

The candidates were manually evaluated by a single annotator with expertise in Slovenian word formation. Inter-annotator agreement was therefore not calculated. Given the exploratory and relatively small-scale nature of the study, the primary aim was qualitative analysis of recurrent error types rather than statistical validation of an annotation scheme.

3.3.1 Two-suffix chains

a) **N + *-en/-ni* + *-ost***

e.g. ***glas*** 'voice; sound' > ***glasen*** 'loud; sonorous' > ***glasnost*** 'loudness; sonority'

[8] The numerical requirement of 100 examples was included only in cases where a relatively large number of such derivatives could be expected, especially in the case of two- and three-suffix chains.

[9] As the performance of LLMs may change across model versions, the results reported here should be understood as specific to ChatGPT 5.5 under the prompting and evaluation conditions described in this study.

The suffix chain N + *-en/-ni + -ost* is the most frequent suffix chain among word-formation chains with a nominal word-formation base.

A preliminary study, which also included data obtained through the automatic extraction of candidate word-formation chains, showed that among 959 candidates displaying this type of combinability, 233 were suitable, corresponding to a success rate of 24%.[10] It should be emphasised that, in the automatic search, it was not possible to reliably distinguish between derivatives in -en and those in -ni. A preliminary search using ChatGPT 5.2 produced candidate word-formation chains 59% of which were suitable. The present study, carried out with ChatGPT 5.5, confirms a comparable but slightly higher proportion of suitable candidates, namely 61%.[11]

Among the unsuitable candidates, the largest group consists of cases in which the first word-formation stage contains the suffix *-ičen* or *-alen*, rather than *-en*, as in *problem* > *problematičen* > *problematičnost and forma* > *formalen* > *formalnost*. These are followed by cases in which the word-formation base is incorrectly identified, for example *vid* > *viden* > *vidnost*. The smallest group consists of cases in which the second-stage derivative in *-ost* is systemically possible but not attested metaFida corpus, as in *sreča* > *srečen* > *srečnost*, or cases in which the derivative is systemically unfounded, as in *koža* > *kožen* > *koženost*.

b) **V + *-en/-ni* + *-ost***

e.g. ***gledati*** 'to look; to watch' **>** ***pregledati*** 'to inspect; to review' **>** ***pregledni*** 'clear; transparent; synoptic' **>** ***preglednost*** 'clarity; transparency; synoptic quality'

The suffix chain V + *-en/-ni + -ost* is likewise the most frequent suffix chain among word-formation chains with a verbal word-formation base. However, the proportion of suitable candidates is considerably lower than in the case of N + *-en/-ni + -ost*. The success rate amounts to only 24%, which is less than half of the success rate recorded for the corresponding denominal suffix chain.

Among the unsuitable candidates, the vast majority consist of cases in which the derivative at the second or third word-formation stage—the latter in chains involving prefixal affixes—does not contain the suffix *-en*, but rather the participial suffix *-n* or *-t*. There are also several cases in which the intermediate first-stage prefixed verbal derivative is missing from the proposed chain, for example *misliti* > *zamišljen* > *zamišljenost* instead of *misliti* > *zamisliti* > *zamišljen* > *zamišljenost*. Only one unsuitable candidate was found in which the word-formation base itself was incorrectly identified: *skrbeti* > *zaskrbeti* > *zaskrbljen* > *zaskrbljenost*. In this case, the appropriate word-formation base should be *skrb*, which is a noun.

c) **Ad + *-iti* + *-enje***

e.g. ***siv*** 'grey' **>** ***siviti*** 'to make grey; to turn grey' **>** ***sivenje*** 'greying'

For the suffix chain Ad + *-iti + -enje*, 72% of the candidates were assessed as suitable. Among the unsuitable candidates, the most frequent problem was the incorrect identification of the word-formation base. For example, *temen* 'dark' is not the word-formation

[10] For a more detailed description of the automatic extraction of word-formation chains in the KOBOS project, see Kern (2026, p. 467).

[11] The comparison with ChatGPT 5.2 is based on a comparable methodological procedure.

base of the corresponding chain; rather, the base is the noun *tema* 'darkness'. Similarly, the relevant bases are not *mračen* 'gloomy', *hladen* 'cold', *vlažen* 'moist' or *miren* 'calm', but the nouns *mrak* 'dusk, darkness', *hlad* 'coldness, coolness', *vlaga* 'moisture' and *mir* 'peace, calm', respectively. Some unsuitable candidates involved second-stage derivatives that are systemically possible but not metaFida corpus, such as *otopljenje* and *začrnjenje*. In a small number of cases, the proposed word-formation chain also contained a non-existent verb, for example *mehčiti*, instead of *mehčati*.

d) **Ad + *-eti* + *-enje***

e.g. ***bled*** 'pale' **>** ***bledeti*** 'to grow pale; to fade' **>** ***bledenje*** 'paling; fading'

Compared with the suffix chain Ad + *-iti* + *-enje*, the results for Ad + *-eti* + *-enje* were less reliable: 56% of the candidates generated by ChatGPT were assessed as suitable. Among the unsuitable candidates, the most frequent problem was the occurrence of first-stage verbs that do not exist in Slovene, such as *mladeti, mehketi* and *suheti*. Another recurrent type of error involved the incorrect identification of the word-formation base. For instance, *besen* 'furious' was treated as the base instead of the noun *bes* 'rage', and *temen* 'dark' instead of the noun *tema* 'darkness'. In addition, several proposed second-stage derivatives in *-enje* were not attested in the metaFida corpus, for example *zmodrenje* and *zazelenenje*. These findings suggest that ChatGPT is less effective in identifying candidate word-formation chains of the type Ad + *-eti* + *-enje* than those of the structurally similar type Ad + *-iti* + *-enje*, partly because the model tends to overgenerate non-existent first-stage verbs and unattested nominal derivatives.

e) **V + *-ø* + *-ni***

e.g. ***pisati*** 'to write' **>** ***dopisati*** 'to add in writing; to write additionally' **>** ***dopis*** 'written addition; letter' **>** ***dopisni*** 'relating to written addition/correspondence'

For the automatically extracted suffix chain V + *ø* + *-ni*, 57% of the candidates were assessed as suitable. Among the unsuitable candidates, the most frequent problem was the incorrect identification of the noun formed with a zero suffix. This applies, for example, to cases such as *posluh* and *prebir*, where the proposed chain did not adequately reflect the expected derivational relationship involving a deverbal noun with a zero suffix. Another recurrent error involved the incorrect identification of the word-formation base, as in cases where *meriti* 'to measure' was treated as the base instead of the noun *mera* 'measure'. In addition, some candidates contained adjectives in *-ni* that were not attested in the metaFida corpus.

3.3.2 Three-suffix chains

a) **V + *-ø* + *-en/-ni* + *-ost***

For the suffix chain V + *-ø* + *-en/-ni* + *-ost*, 52% of the candidates generated by ChatGPT 5.5 were assessed as suitable. Among the unsuitable candidates, the main problem was the lack of confirmation of the final word-formation stage in the metaFida corpus, as in *odhodnost* and *uvoznost*. In some cases, the final two stages were problematic, for example *zagledni* > *zaglednost*, where neither the adjectival derivative nor the corresponding noun in *-ost* was confirmed in the analysed material. Another recurrent error involved the incorrect identification of the word-formation base: the verb *nositi* 'to carry' appeared seven times as

the proposed base, although the appropriate simplex base is *nesti* 'to carry'. These results suggest that, in this suffix chain, ChatGPT 5.5 is moderately successful in identifying suitable candidates, but its reliability decreases particularly at the later word-formation stages and in cases where the distinction between related verbal bases is relevant.

3.3.3 Four-suffix chain

a) **N + *-en/-ni* + *-jak* + *-ar* + *-ica***

e.g. ***bog*** 'god' → ***pobožen*** 'pious' → ***pobožnjak*** 'religious hypocrite' → ***pobožnjakar*** 'religious hypocrite, with an additional expressive or pejorative nuance' → ***pobožnjakarica*** 'female religious hypocrite, with an additional expressive or pejorative nuance'

In the case of the suffix chain N + *-en* + *-jak* + *-ar* + *-ica*, ChatGPT 5.5 did not identify any fully suitable word-formation chains among the ten newly generated examples. The word-formation bases and the first-stage derivatives were mostly identified correctly, for example *greh* 'sin' > *grešen* 'sinful', *glas* 'voice' > *glasen* 'loud', and *strast* 'passion' > *strasten* 'passionate'. However, the second-stage derivatives proposed by the model were non-existent words, such as *grešnjak*, *glasnjak* and *strastnjak*. This shows that, although ChatGPT was generally able to recognise the initial derivational relation between the noun and the adjective in *-en*, it failed to generate attested continuations of the longer suffix chain.

The only partial exception was the chain *svet* 'world' > *posveten* 'secular' > *posvetnjak* 'a person who is not spiritually or religiously oriented' > *posvetnjakar* 'derrogatory for a person who is not spiritually or religiously oriented' > *posvetnjakarica* 'derrogatory for a woman who is not spiritually or religiously oriented'. In this case, the first four members of the chain were assessed as suitable, whereas the feminative in *-ica* was not confirmed in the metaFida corpus. A suitable chain *pamet* 'intelligence, reason' > *pameten* 'intelligent, clever' > *pametnjak* 'clever person; know-it-all' > *pametnjakar* 'smart aleck, know-it-all' > *pametnjakarica* 'female know-it-all' was not identified by ChatGPT 5.5.[12]

The results therefore indicate that ChatGPT 5.5 is markedly less reliable when dealing with highly complex and low-frequency suffix chains, especially where several successive nominal derivatives are required after the initial adjectival formation.

3.3.4 Five-suffix chains

a) **Ad + *-ati* + *-eva-* + *-alen* + *-ica* + *-ški***

e.g. ***boljši*** 'better' → ***boljšati*** 'to improve, to make better' → ***poboljšati*** 'to improve; to reform, to make morally better' → ***poboljševati*** 'to improve/reform repeatedly or over time' → ***poboljševalen*** 'reformatory, corrective, intended for reform' → ***poboljševalnica*** 'reformatory institution' → ***poboljševalniški*** 'relating to a reformatory institution'

For the suffix chain Ad *+ -ati + -eva- + -alen + -ica + -ški*, ChatGPT 5.5 did not identify any fully suitable word-formation chains. However, it is noteworthy that four out of the ten generated

[12] This word-formation chain had already been identified in a previous study using ChatGPT 5.2 (Kern, 2026, p. 432).

examples displayed an almost entirely suitable chain: in these cases, all members of the chain were assessed as suitable except for the final adjectival derivative in *-ški*, which was not attested in the metaFida corpus. This suggests that ChatGPT 5.5 was relatively successful in reconstructing the earlier stages of this highly complex suffix chain, but less reliable in predicting the final adjectival formation.

4 Conclusions

The results provide partial confirmation of the initial research aims. First, the comparison with the preliminary study shows that ChatGPT 5.5 produced slightly better results than ChatGPT 5.2 for the suffix chain **N + -en/-ni + -ost**: the proportion of suitable candidates increased from 59% to 61%. Second, the part-of-speech category of the word-formation base clearly affects the effectiveness of candidate identification, as noun-based chains such as **N + -en/-ni + -ost** yielded substantially better results (61%) than the corresponding verb-based chain **V + -en/-ni + -ost** (24%). Third, structurally similar suffix chains do not necessarily produce comparable results, as shown by the difference between **Ad + -iti + -enje** and **Ad + -eti + -enje** (72% vs. 56%). Fourth, the presence of a zero suffix appears to reduce reliability, particularly because ChatGPT often fails to identify the zero-derived noun or the appropriate word-formation base. Fifth, increasing suffix-chain complexity has a clear negative effect on the effectiveness of candidate identification: ChatGPT 5.5 was moderately successful with two- and three-suffix chains, but failed to identify any fully suitable examples for the most complex four- and five-suffix chains. The results for all analysed suffix chains are summarised in Table 1. Overall, ChatGPT can be useful as an exploratory tool for generating candidate word-formation chains, but the results require systematic linguistic verification. Future research will build on these findings by refining prompting strategies to address the recurrent error types identified in the analysis and by testing whether such modifications improve the identification of more complex word-formation chains. The analysis will also be extended to additional derivational patterns and word-formation bases.

TABLE 1 Percentage of suitable candidates by suffix chain

Suffix chain	Percentage of suitable candidates
N + *-en/-ni* + *-ost*	61%
V + *-en/-ni* + *-ost*	24%
Ad + *-iti* + *-enje*	72%
Ad + *-eti* + *-enje*	56%
V + *-ø* + *-ni*	57%
V + *ø* + *-en/-ni* + *-ost*	52%
N + *-en/-ni* + *-jak* + *-ar* + *-ica*	0%
Ad + *-ati* + *-eva-* + *-alen* + *-ica* + *-ški*	0%

References

Burkacka, I. (2001). *Porównawcza analiza gniazdowa wybranych leksemów neutralnych i nacechowanych ekspresywnie*. Dom Wydawniczy Elipsa.

Burkacka, I. (2012). *Kombinatoryka sufiksalna w polskiej derywacji odrzeczownikowej*. Wydział Polonistyki Uniwersytetu Warszawskiego.

Erjavec, T., Pranjić, M., Kern, B., Pollak, S., Pelicon, A. & Stramljič Breznik, I. (2023). Automating derivational morphology for Slovenian. In M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček & S. Krek (Eds.), *Proceedings of eLex 2023: Electronic lexicography in the 21st century* (pp. 449–465). Lexical Computing CZ. https://elex.link/elex2023/wp-content/uploads/92.pdf

Erjavec, T. (2023). Corpus of combined Slovenian corpora metaFida 1.0. *Slovenian language resource repository CLARIN.SI*. Retrieved May 31, 2026, from http://hdl.handle.net/11356/1747

Jadacka, H. (2003). Zagadnienie motywacji słowotwórczej w opisie gniazdowym. In M. Skarżyński (Ed.), *Słowotwórstwo gniazdowe: Historia – metoda – zastosowania*. Księgarnia Akademicka.

Kern, B. (2017). *Stopenjsko besedotvorje: Na primeru glagolov čutnega zaznavanja*. Založba ZRC. https://omp.zrc-sazu.si/zalozba/catalog/book/1082

Kern, B. (2020). Kombinatorika priponskih obrazil v besedotvornih sestavih glagolov čutnega zaznavanja. In M. Krajnc Ivič & A. Žele (Eds.), *Pogled v jezik in iz jezika: Adi Vidovič Muha ob jubileju* (pp. 67–79). Univerzitetna založba.

Kern, B. (2024). Considering word formation in compiling dictionaries. In K. Štrkalj Despot (Ed.), *Lexicography and Semantics: Proceedings of the XXI EURALEX International Congress* (pp. 438–448). Institut za hrvatski jezik. https://euralex.jezik.hr/wp-content/uploads/2021/09/Euralex-XXI-proceedings_1st.pdf

Kern, B. (2026). Izsamostalniški besedotvorni nizi v slovenščini z vidika morfotaktike. In B. Kern (Ed.), *Stopenjsko besedotvorje v slovanskih jezikih*. Založba ZRC. [In press.]

Kern, B., & Ledinek, N. (2025). Izprislovne stopenjske tvorjenke. In F. Marušič, M. Melinc Mlekuž, P. Mišmaš & R. Žaucer (Eds.), *Škrabčevi dnevi 13: Zbornik prispevkov s simpozija 2023* (pp. 16–28). Založba Univerze v Novi Gorici. https://ung.si/media/publishing/2025/03/12/08/24/17/Zbornik-SD13-2025-koncna.pdf

Kern, B., & Stramljič Breznik, I. (2025). Kombinatorika izhodiščnega obrazila -ati v slovenščini. *Slavistična revija*, 73(2), 329–346. https://doi.org/10.57589/srl.v73i2

Olejniczak, M. (2003). 30 lat słowotwórstwa gniazdowego: Rys historyczny. In M. Skarżyński (Ed.), *Słowotwórstwo gniazdowe: Historia – metoda – zastosowania* (pp. 12–28). Księgarnia Akademicka.

Pollak, S., Voršič, I., Kern, B. & Ulčar, M. (2023). Novel Slovenian COVID-19 vocabulary from the perspective of naming possibilities and word formation. In M. Medveď,, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček & S. Krek (Eds.), *Proceedings of eLex 2023: Electronic lexicography in the 21st century* (pp. 419–438). Lexical Computing CZ. https://elex.link/elex2023/wp-content/uploads/92.pdf

Skarżyński, M. (2003). Słowotwórcze gniazda odczasownikowe: Charakterystyka wstępna. In M. Skarżyński (Ed.), *Słowotwórstwo gniazdowe: Historia – metoda – zastosowania*. Księgarnia Akademicka.

Slovar slovenskega knjižnega jezika 2. Fran. Retrieved May 31, 2026, from https://www.fran.si

Slovenski etimološki slovar 3. Fran. Retrieved May 31, 2026, from https://www.fran.si

Stramljič Breznik, I. (2004). *Besednodružinski slovar slovenskega jezika: Poskusni zvezek za iztočnice na B*. Slavistično društvo.

Stramljič Breznik, I. (2024). Značilnosti izmedmetnih priponskih nizov v Besednodružinskem slovarju slovenskega jezika za iztočnice na B. *Slavia Centralis*, 17(1), 1–17.

Stramljič Breznik, I., & Ledinek, N. (2024). Kvantitativni podatki o besedotvornih modelih in priponskih nizih izmedmetnih tvorjenk v Besednodružinskem slovarju slovenskega jezika za iztočnice na B. Slovenski jezik. *Slovene Linguistic Studies*, 16, 59–85.

Vidovič Muha, A. (1988). *Slovensko skladenjsko besedotvorje ob primerih zloženk*. Znanstveni inštitut Filozofske fakultete, Partizanska knjiga.

Vidovič Muha, A. (2011). *Slovensko skladenjsko besedotvorje*. Znanstvena založba Filozofske fakultete.

Voršič, I., Žele, A. & Ledinek, N. (2024). Od skladenjskega do stopenjskega besedotvorja pri prikazu tvorbenih zmožnosti izbranih glagolov in prislovov. *Slavistična revija*, 72(4), 369–388. https://srl.si/ojs/srl/article/view/4221

Contact information

Boris Kern
Research Centre of the Slovenian Academy of Sciences and Arts, Fran Ramovš Institute of the Slovenian Language; University of Nova Gorica, School of Humanities
boris.kern@zrc-sazu.si

Acknowledgements

The article was written within the framework of the programme *Slovene in Synchronic and Diachronic Development* (P6-0038), led by Associate Professor Dr Kozma Ahačič. The project and programme are co-financed by the Slovenian Research and Innovation Agency.

Václava Kettnerová, Veronika Kolářová & Jiří Mírovský

WHERE THE BOUNDARY BETWEEN DERIVATIONAL AND INFLECTIONAL MORPHOLOGY MEETS LEXICOGRAPHY

Building the DeriVallex Lexicon

Abstract The lexicographic treatment of word-formation categories that lie at the boundary between inflection and derivation remains an unresolved issue. In this paper, we introduce the process of compiling *DeriVallex*, a valency lexicon containing automatically generated valency frames, which provides information on the valency behavior of Czech noun and adjectival deverbal derivatives being halfway between forms and lexemes. These word-formation categories inherit their valency frames and other valency-related phenomena (as, e.g., reflexivity, reciprocity, and other semantic alternations) from their base verbs. We first outline the automatic generation of valency frames of these noun and adjectival derivatives, which resulted in 17,145 valency frames assigned to lexical units, i.e., individual senses of these derivatives, represented by 11,248 noun and adjectival lemmas. The focus of this paper in on the process of structuring of the obtained data into a valency lexicon, providing rich information on valency and derivational properties of the selected nouns and adjectives, enabling human users to search and filter the data. The resulting data may be integrated into both valency lexicons and explanatory dictionaries, thereby increasing their descriptive comprehensiveness.

Keywords Valency; valency lexicons; derivation; inflection; word-formation categories; transposition; noun and adjectival derivatives; verb forms; noun and adjectival lemmas

1 Introduction

The boundary between inflection and derivation has been the subject of long-standing debate (see, e.g., Booij, 2000a; ten Hacken, 2014). Being halfway between forms and lexemes, nouns and adjectives falling under transpositional word-formation categories stand at this boundary (e.g., the Czech action noun *zařízení* 'arranging' < *zařídit* 'to arrange'). Transposition is defined by Dokulil (1962, p. 229) as the onomaziological category in which a syntactic function of a word is changed without changing its lexical content. What brings transposition close to inflection is semantic transparency, valency inheritance, and usually high productivity. What sets it apart is that the output is a new lexeme that may acquire additional meanings (e.g., *zařízení* 'device' and *zařízení* 'facility'; such lexicalized meanings, however, do not fall under transposition anymore).

The question arises of how these word-formation categories should be represented in dictionaries. For example, while Panocová and ten Hacken (2018) in their proposal of the representation of action nouns suggest that they should be recorded in run-on entries on condition that their existence is unpredictable, *Akademický slovník současné češtiny* (ASSČ 2017–2026) has adopted a different approach, treating these nouns within the lexical entry of their base verbs – similarly to one of their forms (see Kochová & Opavská, 2016).

In this paper, we introduce *DeriVallex*, a valency lexicon of automatically generated valency frames of Czech noun and adjectival derivatives falling predominantly under transpositional word-formation categories (see Kettnerová et al., 2026). We limit our attention to deverbal derivatives, namely to (i) action nouns ending in *-ní/-tí* (e.g., *rozdělování* 'distributing' and *natáhnutí* 'stretching'), (ii) active adjectives denoting either a simultaneous action (e.g., *prchající* 'fleeing') or (iii) an anterior action ending in *-lý* and *-(v)ší* (e.g., *uniklý* 'escaped',

přišedší 'having arrived', and *přeživší* 'having survived'), and (iv) passive action adjectives ending in *-ný/-tý* (e.g., *doporučený* 'recommended' and *zapomenutý* 'forgotten'); see Dokulil et al. (1986). In addition, we discuss potentiality adjectives (e.g., *čitelný* 'readable') as these adjectives exhibit several properties common to the transpositional word-formation categories. Drawing on the construction of *DeriVallex*, we introduce our findings on these categories and their implications for lexicographic practice. In line with Atkins and Rundell (2008, pp. 299–304), who emphasize the importance of valency descriptions for dictionary writing, we focus on valency and valency-related phenomena associated with the selected categories.

The paper is structured as follows. Section 2 describes the valency behavior of the above-given word-formation categories. Section 3 presents the automatic generation of valency frames and in particular the process by which data suitable for natural language processing tasks were transformed into a valency lexicon with an interface allowing human users to search and filter the data. Section 4 introduces possible uses of the data provided by *DeriVallex* for lexicographic purposes.

2 Valency: Between Forms and Lexemes

Valency can be defined as *"the capacity a verb (or noun, etc.) has for combining with particular patterns of other sentence constituents"* (Allerton, 2006, p. 301). In contrast to the valency of verbs, the valency of nouns and adjectives remains understudied. Among the most extensively studied are deverbal nouns and partly adjectives, the valency of which often corresponds to that of their base verbs (see esp. Chomsky, 1970; Grimshaw, 1990; Alexiadou, 2007; for Czech, see Kolářová et al., 2019; Kolářová et al., 2021). Those deverbal nouns and adjectives the valency of which most closely parallels that of their base verbs belong to transpositional word-formation categories. The valency correspondences observed in these nouns and adjectives support the view that their valency is inherited from base verbs (see esp. Booij, 2000b; Bierwisch, 2015; Grimshaw, 1990; Alexiadou, 2007; Wunderlich, 2012). Further, as the valency of verbs is subject to various valency operations such as diatheses, reflexivity, reciprocity, and other alternations (see, e.g., Levin, 1993; Kulikov, 2009; Haspelmath & Müller-Bardey, 2015), the question arises as to which of these valency-related properties are also inherited by the deverbal nouns and adjectives concerned and which are not.

In this section, we first describe valency inheritance in transpositional word-formation categories of deverbal nouns and adjectives in Czech (Section 2.1) and then specify which valency-related properties of verbs are relevant to the discussed nouns and adjectives and may also be considered inherited (Section 2.2). Our findings come from an analysis of the data provided by *VALLEX*, a valency lexicon of Czech verbs, and *NomVallex*, a valency lexicon of Czech nouns and adjectives.

2.1 Valency Inheritance

In Czech, transpositional word-formation categories include action nouns, action adjectives denoting simultaneous action and anterior action, and passive action adjectives. We can observe that these nouns and adjectives have the same number and type of valency complements as their base verbs, i.e., their deep valency can be considered inherited. What changes is the surface realization of their valency complements (i.e., surface valency) so that these deverbal derivatives can fulfil syntactic functions different from those of their base verbs. From this fact follows that the surface changes affect the valency complements expressed in prototypical verbal positions, particularly in the nominative subject and the accusative direct object.

Further, adjectives denoting passive (e.g., *nahraditelný* 'substitutable') and active potentiality (e.g., *rozpustitelný* 'dissolvable by itself') exhibit the same valency behavior as the adjectival derivatives belonging to the transpositional category, despite falling under the onomaziological category of mutation (Dokulil, 1962), where the lexical content of a word, and thus its semantic properties, changes. These adjectives thus illustrate that the boundary between transposition and mutation is not sharp.

As the changes in the surface expression of valency complements underlie the rules by which valency frames of the involved word-formation categories are generated (see Section 3.1), we summarize them in Table 1.

TABLE 1 Changes of the surface expression of valency complements in the discussed word-formation categories of deverbal nouns (N) and adjectives (A) (↑ indicates the antecedent of the valency complement elided from the surface)

Word-formation category	Surface Changes	Example
Action N	Nom > Instr, Gen, Poss Acc > Gen, Poss	*parník*$_{\text{Nom}}$ *drtil kry*$_{\text{Acc}}$ *na kusy* 'a steamer was crushing the ice floes into pieces' > ***drcení*** *ker*$_{\text{Gen}}$ *na kusy parníkem*$_{\text{Instr}}$ 'the crushing of the ice floes into pieces by a steamer'
Simultaneous action A	Nom > surface ellipsis[a]	> *parník*↑ ***drtící*** *kry na kusy* 'a steamer crushing the ice floes into pieces'
Passive action A	Nom > Instr, *od* 'from'+Gen Acc > surface ellipsis	> *kry*↑ ***drcené*** *parníkem*$_{\text{Instr}}$ *na kusy* 'the ice floes crushed into pieces by a steamer'
Anterior action A	Nom > surface ellipsis	*parník*$_{\text{Nom}}$ *rozdrtil kry*$_{\text{Acc}}$ *na kusy* 'a steamer crushed the ice floes into pieces' > *parník*↑ ***rozdrtivší*** *kry na kusy* 'a steamer having crushed the ice floes into pieces'
Passive potentiality A	Nom > Instr Acc > surface ellipsis	> *kry*↑ ***rozdrtitelné*** *parníkem*$_{\text{Instr}}$ *na kusy* 'the ice floes crushable by a steamer'
Active potentiality A	Nom > surface ellipsis	*nozdry*$_{\text{Nom}}$ *se roztáhly* 'the nostrils dilated' > ***roztažitelné*** *nozdry*↑ 'dilatable nostrils'

[a]In adjectives, one of valency complements present in the deep valency of the adjective is not expressed on the surface as an adjectival modification, i.e., it is elided from its surface valency. The antecedent of this elided complement is expressed on the surface either as the subject of the construction with the copular verb (e.g., *kry jsou drcené parníkem na kusy* 'ice floes are crushed into pieces by a steamer') or as the governor of the adjective (see examples in Table 1).

2.2 Inheritance of Other Valency-Related Phenomena

The valency of a verb is determined by the number and type of participants of the situation the verb denotes, as well as by the relations between these participants. Valency then governs the syntactic realization of participants in all syntactic constructions of the verb. In Czech, besides active constructions, these comprise diatheses (e.g., passive, impersonal or middle constructions), reflexivity, reciprocity, and other semantic alternations (particularly, lexical conversion, multiple realization of a single participant and splitting of a single participant). Table 2 displays which of these constructions extend from verbs to the discussed derivatives and can thus be seen as inherited from verbs. Their inheritance is then exploited in building *DeriVallex* (see Section 3.2.2).

TABLE 2 Valency-related phenomena, particularly, diatheses (Diat), reflexivity (Refl), reciprocity (Recipr), lexical conversion affecting either direct object (Obj conv) or subject (Subj conv), multiple realization of a participant (Multiple), and splitting of a participant (Split) (–/✓ indicate non-/applicability)

Word-formation category	Valency-related phenomenon						
	Diat	Refl	Recipr	Obj conv	Subj conv	Multiple	Split
Action N	–	✓	✓	✓	?✓[a]	✓	✓
Simultaneous action A	–	✓	✓	✓	✓	✓	✓
Passive action A	–	✓	✓	✓	–	✓	✓
Anterior action A	–	✓	✓	✓	✓	✓	✓
Passive potentiality A	–	✓	✓	✓	–	✓	✓
Active potentiality A	–	✓	✓	–	✓	✓	–

[a]Some constructions of subject-oriented conversions appear to be less stylistically acceptable (e.g., *?hemžení se zahrady myšmi* 'the garden filled with the scurrying of mice').

Most valency-related phenomena exhibited by verbs also apply to the discussed noun and adjectival derivatives. *Diatheses*, restricted to verbs (see Table 2), represent an exception, as they are conditioned by changes in the morphological category of verbal voice and deverbal derivatives lack this category.[1] In contrast, *reflexivity* and *reciprocity* are retained in all the discussed deverbal derivatives (cf., the reflexive structures in (1a) and (1b) and the reciprocal structures in (2a) and (2b)).

(1) a. *hrdinka* ***překonala***$_{V}$ *sebe samu* 'the heroine has surpassed herself'
b. *hrdinka* ***překonavší***$_{\text{anter-act-A}}$ *sebe samu* 'the heroine who has surpassed herself'

(2) a. *ceny, které* ***mohou*** *kupující* ***srovnávat***$_{V}$ *mezi sebou* 'prices that buyers can compare with each other'
b. *ceny* ***srovnatelné***$_{\text{pass-pot-A}}$ *mezi sebou* 'prices comparable with each other'

Both *object-* and *subject-oriented lexical conversion* are generally available in the noun and adjectival derivatives in question as well (cf., the lexical conversion affecting the direct object in (3a) and (3b) and the conversion involving the subject in (4a) and (4b)). There are, however, three restrictions (see the columns Obj conv and Subj conv in Table 2). First, object-oriented lexical conversion is not relevant for active potentiality adjectives as these adjectives are derived from intransitive verbs, which do not have a valency complement expressed as a direct object. Conversely, subject-oriented conversion does not occur in passive action and passive potentiality adjectives as they are derived from transitive verbs, which only rarely form converse constructions with the affected subject.

(3) a. *Sadaři* ***natírají***$_{V}$ *vápno na kmeny stromů* 'Fruit growers are smearing lime on tree trunks' — *Sadaři* ***natírají***$_{V}$ *kmeny stromů vápnem* 'Fruit growers are smearing tree trunks with lime'
b. *vápno* ***natírané***$_{\text{pass-act-A}}$ *na kmeny stromů* 'lime smeared on tree trunks' — *kmeny stromů* ***natírané***$_{\text{pass-act-A}}$ *vápnem* 'tree trunks smeared with lime'

(4) a. *na stěnách* ***září***$_{V}$ *barvy* 'colors are glowing on the walls' — *stěny* ***září***$_{V}$ *barvami* 'the walls are glowing with colors'
b. *barvy* ***zářící***$_{\text{sim-act-A}}$ *na stěnách* 'colors glowing on the walls' — *stěny* ***zářící***$_{\text{sim-act-A}}$ *barvami* 'the walls glowing with colors'

[1] However, active and passive syntax of base verbs is reflected in constructions with these categories (Kolářová, 2024).

The possibility of realizing a single participant of a situation as two distinct types of valency complements in verbs (see the column Multiple in Table 2) as well as the possibility of expressing a single participant once as one valency complement and another time as two complements in verbs (see the column Split in Table 2) are also preserved in the discussed noun and adjectival derivatives; see examples (5) and (6), respectively. There is only one exception: splitting of a single participant is not possible in active potentiality adjectives (see Table 2), as this operation affects an accusative complement in verbs; however, these adjectives are derived from intransitive verbs, which lack such a complement.

(5) a. *běžec* ***přeskočil***$_{V}$ *přes překážku* 'the runner jumped over the hurdle' — *běžec* ***přeskočil***$_{V}$ *překážku* 'the runner cleared the hurdle'
b. *běžcovo* ***přeskočení***$_{act\text{-}N}$ *přes překážku* 'the runner's jumping over the hurdle' — *běžcovo* ***přeskočení***$_{act\text{-}N}$ *překážky* 'the runner's clearing of the hurdle'

(6) a. *astronom* ***pozoruje***$_{V}$*, jak Venuše putuje přes Slunce — atronom* ***pozoruje***$_{V}$ *Venuši, jak putuje přes Slunce* 'an astronomer observes Venus transiting the Sun' lit. an astronomer observes how Venus is transiting the Sun — an astronomer observes Venus how it is transiting the Sun
b. *astronomovo* ***pozorování***$_{act\text{-}N}$*, jak Venuše putuje přes Slunce — astronomovo* ***pozorování***$_{act\text{-}N}$ *Venuše, jak putuje přes Slunce* 'the astronomer's observing of Venus transiting the Sun' lit. the astronomer's observing how Venus is transiting the Sun — the astronomer's observing Venus how it is transiting the Sun

3 Building DeriVallex

The development of valency lexicons is labor-intensive. In the case of valency lexicons dealing with nouns and adjectives, it is further complicated by the fact that noun and adjectival valency is still under-researched (see Section 2). It is thus not surprising that the description of nouns and adjectives in valency lexicons is still limited. Let us mention at least some valency lexicons providing such information: for English, see Herbst et al. (2004), *Walenty* for Polish, for Czech the printed dictionary *Slovník slovesných, substantivních a adjektivních vazeb a spojení* (Svozilová et al., 2005), and electronic lexicons *PDT-Vallex* and *NomVallex*.

Unlike the three aforementioned resources for Czech, *DeriVallex* has not been built manually but automatically with the aim of filling the gap in the description of noun and adjectival valency in Czech. It covers the productive and systemic core of the Czech lexicon. In its development, the data provided by three lexical resources have been used: the valency lexicon of Czech verbs, *VALLEX 4.5*, the valency lexicon of Czech nouns and adjectives, *NomVallex 2.6*, and the lexical network of word-formation relations in Czech, *DeriNet 2.3*. This section focuses on the process of automatically generating data intended for natural language processing tasks (Section 3.1) and on converting them into a structured valency lexicon with an interface that supports user-oriented search and filtering (Section 3.2).

3.1 Data Processing Pipeline: From Analysis to Generation

In the used valency lexicons, valency is captured in the form of valency frames, modeled as a sequence of valency slots, each standing for one complement, supplemented with a list of morphosyntactic forms indicating their surface realization (see Figure 1 for the valency frame of the verb *léčit* 'to treat'). As the automatic generation of valency frames of the discussed word-formation categories is described in detail in Kettnerová et al. (2026), we limit ourselves here to a summary of the key steps of the entire process. First, based on the semiautomatic comparison of the valency frames of verbs in *VALLEX 4.5* and the valency frames of their noun and adjectival derivatives contained in *NomVallex 2.6*, action nouns, simultaneous action adjectives,

passive action adjectives, anterior action adjectives, and potentiality adjectives were identified as those word-formation categories that exhibit systemic correspondences in valency (see Table 1 in Section 2.1). Further, based on this comparison, formal rules describing changes in the valency frames of the selected word-formation categories were formulated. Each rule specifies the conditions under which it applies to the valency frame of a base verb in *VALLEX*. The rules thus filter out those valency frames on which the prescribed changes operate. For each selected valency frame, the verb lemma(s) representing the lexical unit of the verb to which the valency frame is assigned were extracted from *VALLEX* too.[2] For each of these lemmas, the corresponding noun or adjectival derivative was then sought in *DeriNet*.[3] If such a derivative was found, it was assigned the valency frame derived from that of the base verb in accordance with the relevant rule. The manual analysis of the obtained valency frames showed that 97.8% of the generated frames were correctly determined (see Kettnerová et al., 2026).

See Figure 1, illustrating the pathway from the valency frame of the verb *léčit* 'to treat' to the valency frame of the passive potentiality adjective *léčitelný* 'treatable'. As the valency frame of the verb *léčit* 'to treat' satisfies the conditions prescribed in the rule describing passive potentiality adjectives (i.e., it is not assigned to a lexical unit represented by a reflexive verb and it contains ACTor in the nominative and PATient in the accusative), an adjective ending in *-telný* 'able' derived from the lemma *léčit* 'to treat' was extracted from *DeriNet*. Then the rule changed the valency frame of the verb: first, it replaces all the possible forms of the ACTor (indicated by the asterisk) with the instrumental, and second, it overwrites all the possible forms of the PATient by the upward arrow, indicating its surface ellipsis. The output of these operations is then included in *DeriVallex* as a lexical entry of the adjective *léčitelný* 'treatable'.

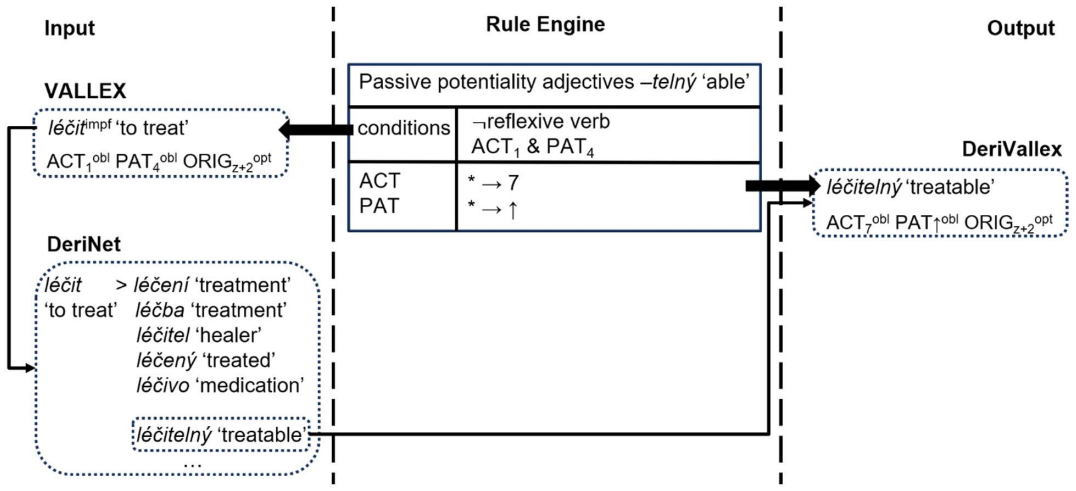


FIGURE 1 The pathway from the valency frame of the verb *léčit* 'to treat' in *VALLEX* through the rule engine to the valency frame of the passive potentiality adjective *léčitelný* 'treatable' in *DeriVallex*

2 The concepts of lemmas and lexical units are explained in detail in Section 3.2.1.

3 For action nouns, simultaneous action adjectives, passive action adjectives, and anterior action adjectives, the inventory of derivational means based on which their lemmas were detected in *DeriNet* is exhaustive. In the case of potentiality adjectives, only a central derivational means, the suffix *-telný* 'able' (e.g., *dělitelný* 'divisible'), was used for their identification. These adjectives are peripherally derived by the suffixes *-ný* and *-lý* as well. However, both these suffixes are not productive in forming them and at the same time they exhibit a high degree of ambiguity (cf., the potentiality adjective *rozpust-ný* 'soluble' with the passive action adjective *rozpuště-ný* 'dissolved' and further the potentiality adjective *jed-lý* 'eadable' with the anterior action adjective *spad-lý* 'having fallen').

The coverage of the noun and adjectival derivatives in *DeriVallex* is worth commenting on. As the lemmas contained in *DeriNet* were originally sampled from the *Czech National Corpus* and later substantially expanded using *MorfFlex*, a morphological dictionary of Czech comprising more than 1 million lemmas, including those that conform to Czech derivational patterns despite being corpus-unattested (see Žabokrtský et al., 2016), the resulting inventory of noun and adjectival derivatives obtained from *DeriNet* can thus be considered largely exhaustive (for further details see Section 3.2). However, noun and adjectival lemmas that are derivationally well-formed and semantically interpretable, yet corpus-unattested, were excluded from the *DeriVallex* data to maintain a distinction between systemic derivational potential and empirically attested usage. The representativeness of the *DeriVallex* data is further reinforced by the fact that the base verbs contained in *VALLEX* from which the derivatives are derived achieve a high corpus coverage (particularly, the verbs in this lexicon account for more than 96% of verb occurrences in the sub-corpus SYN2000 of the *Czech National Corpus*). In sum, *DeriVallex* includes a comprehensive list of noun and adjectival derivatives of base verbs that have high corpus coverage.

A further noteworthy observation resulting from the data is that not every valency frame of a verb satisfying the conditions imposed by a given rule has a corresponding noun or adjectival derivative in *DeriNet* (e.g., *proslout* 'to become famous' > **proslutí* 'becoming famous' and *vyskočit* 'to jump out' > **vyskočilý* 'having jumped out'). Consequently, no valency frame was generated in this case. See Table 3, displaying the proportion of the generated valency frames of nouns and adjectives (in total, 17,145 valency frames) relative to the valency frames of base verbs filtered out by the conditions specified in the rules.

TABLE 3 The number of the valency frames of base verbs (col. Base frames) and the number of the automatically generated valency frames of nouns and adjectives (col. Generated frames). The percentages correspond to the proportion of the generated frames relative to the base frames

Word-formation category	Base frames	Generated frames
Action N	5,138	4,946/96.3%
Simultaneous action A	5,317	4,425/83.2%
Passive action A	3,722	3,128/84%
Anterior action A	5,317	2,263/42.6%
Passive potentiality A	3,053	2,025/66.3%
Active potentiality A	500	358/71.6%

3.2 From Automatically Generated Data to a User-Oriented Lexicon

As the next step, the obtained data were structured into a valency lexicon allowing human users to search and filter the data. A key issue in this process concerned the choice of the headword lemma(s) introducing the lexical entry (Section 3.2.1) and the overall structuring of information on valency and derivational properties of derivatives provided in the lexical entry (Section 3.2.2). The structure of the lexical entry was adopted from the *VALLEX* lexicon and further modified for the purpose of the description of noun and adjectival derivatives in question, inspired by the *NomVallex* lexicon.

3.2.1 Headword Lemma

Each lexical entry in *DeriVallex* contains a lexeme, an abstract unit associating all lexical forms of a noun or adjectival derivative on the one hand and its lexical units (LUs), individual senses, on the other. Each lexical entry is introduced by a headword lemma or more lemmas,

as *aspectual counterparts* and *lemma variants* are treated together in the same entry (this approach is motivated by the fact that aspectual counterparts and lemma variants share valency properties). As noun and adjectival derivatives were returned from *DeriNet* as separate lemmas, it was necessary to group together their aspectual counterparts (e.g., *dokončovaný*^{impf} and *dokončený*^{pf} 'completed') and lemma variants (e.g., *odstrkávání* and *odstrkování* 'pushing away').[4]

Further, *non-reflexive* and *reflexive lemmas* are treated in separate lexical entries in *DeriVallex*. As *DeriNet* does not cover reflexive lemmas, the reflexive morpheme representing part of noun or adjectival lemmas had to be added where relevant. There were three cases. First, the reflexive morpheme is inherited from base verbs, including its obligatoriness or optionality (see the row Full inheritance in Table 4).[5] Second, the reflexive is inherited from base verbs, however, it becomes optional in the derivative regardless of its obligatoriness in the base verb (the row Partial inheritance). Finally, the reflexive present in the base verb lemma is dropped in derivation (the row Non-inheritance). For statistics on the non-reflexive and reflexive lemmas see Table 5.

TABLE 4 Inheritance of the reflexive morpheme in the discussed noun and adjectival derivatives[a]

	Word-formation category	Type of the reflexive	Examples
Full inheritance	Simultaneous action A	Obligatory	*vznášet se* 'to hover' > *vznášející se* 'hovering' *libovat si* 'to take pleasure' > *libující si* 'taking pleasure'
		Optional	*koukat (se)* 'to watch' > *koukající (se)* 'watching' *číst (si)* 'to read' > *čtoucí (si)* 'reading'
	Anterior action A ending in *–(v)ší*	Obligatory	*dotknout se* 'to touch' > *dotknuvší se* 'having touched' *postesknout si* 'to complain' > *postesknuvší si* 'having complained'
		Optional	*skončit (se)* 'to end' > *skončivší (se)* 'having ended' *vyžádat (si)* 'to request' > *vyžádavší (si)* 'having requested'
Partial inheritance	Action N	Only optional	*vznášení (se)* 'hovering', *libování (si)* 'taking pleasure', *koukání (se)* 'watching', *čtení (si)* 'reading', *dotknutí (se)* 'touching', *postesknutí (si)* 'complaining', *skončení (se)* 'completion', and *vyžádání (si)* 'requesting'
Non-inheritance	Anterior action A ending in *-lý*	—	*rozpadnout se* 'to fall apart' > *rozpadlý* 'fallen apart', *propadnout (se)* 'to sink' > *propadlý* 'sunken'[b]
	Passive action A	—	*dožadovat se* 'to request' > *dožadovaný* 'requested' *uvědomit si* 'to realize' > *uvědomovaný* 'recognized'
	Active potentiality A	—	*rozpustit se* 'to dissolve' > *rozpustitelný* 'dissolvable

[a]Passive potentiality adjectives are not included in Table 4, as they are derived only from non-reflexive verbs (with a few exceptions, such as, for example, *nedotknutelný* 'untouchable' < *nedotknout se* 'not to touch', which are not, however, contained in the data).
[b]The data do not attest any adjectives ending in *-lý* derived from verbs with obligatory or optional reflexive morpheme *si*.

[4] As the valency frames were originally generated for LUs represented by separate lemmas, the grouping of aspectual counterparts and lemma variants consequently cut down the number of the generated valency frames (cf., Kettnerová et al., 2026).

[5] Under the optionality of the reflexive morpheme, we understand the situation where a verb can be used with or without the reflexive without changing the meaning (e.g., *Jan spoléhá na bratra* and *Jan se spoléhá na bratra* 'John relies on his brother').

Finally, *DeriVallex* distinguishes *homographs*, i.e., the lemmas with the same spelling but unrelated meanings (e.g., *lekání*$_{I}$ *ryb* 'fish dying' and *lekání*$_{II}$ *malých dětí* 'scaring children') and captures them in separate lexical entries. *Derivallex* contains a total of 11,248 lemmas representing the noun and adjectival derivatives in question; see Table 5.

TABLE 5 The number of lemmas and LUs of the nouns and adjectives in *DeriVallex*, distinguishing non-reflexive and reflexive instances and homographs (the total number of non-reflexive lemmas is not their cumulative sum: as active potentiality adjectives are derived from reflexive verbs, which are derived from non-reflexive verbs that form passive potentiality adjectives, they are represented by the same non-reflexive lemmas as the corresponding passive potentiality adjectives; homographs are counted separately as they can be both non-reflexive and reflexive)

	Non-reflexive lemmas/ LUs	Reflexive lemmas/ LUs	Homographs/ LUs
Action N	2,835/3,722	1,327/1,224	160/205
Simultaneous action A	1,543/3,220	762/1,205	117/186
Passive action A	2,439/3,128	0/0	109/125
Anterior action A	789/1,668	355/595	37/61
Passive potentiality A	1,198/2,025	0/0	45/72
Active potentiality A	348/358	0/0	9/7
Total	8,804/14,121	2,444/3,024	477/656

3.2.2 Valency Properties

Each lexeme contains a set of LU(s) of a noun or adjectival derivative (17,145 in total; see Table 5 in Section 3.2.1). Each LU, characterized by a gloss,[6] is assigned the automatically generated valency frame, describing its valency behavior. In addition, each LU is specified by the attributes introducing its word-formation characteristics, particularly its derivational type[7] and its word-formation category, supplemented with the link to the valency frame of its base verb. Moreover, information on differences between the valency frame of the LU of the derivative and that of the LU of its base verb is provided in the *valdiff* attribute.

In addition to the obligatory attributes introduced above, LUs are accompanied with information on other valency-related characteristics where relevant, particularly, on reflexivity and reciprocity, and on conversion, multiple realization and splitting of participants. See Figure 2, displaying the lexical entry of the passive action adjective *ošetřovaný*^{impf}, *ošetřený*^{pf} 'treated', comprising two LUs related by the converse relation Affliction-Patient (cf., *rána ošetřovaná*^{impf}*/ošetřená*^{pf} *Petrovi zdravotní sestrou* 'a wound treated by a nurse for Peter' and *Petr ošetřovaný*^{impf}*/ošetřený*^{pf} *zdravotní sestrou* 'Peter treated by a nurse'). Further, the second LU of this adjective exhibits reflexivity (e.g., *Petr ošetřovaný*^{impf}*/ošetřený*^{pf} *sám sebou* 'Peter treated by himself') and reciprocity (e.g., *Petr a Karel ošetření sebou navzájem* 'Peter and Karel treated by each other').

[6] The glosses were automatically generated from the original glosses describing the meaning of LUs of base verbs, using LLMs.

[7] The attribute *type* combines information on the derivational category (deverbal nouns or deverbal adjectives in this case) and the specific means (esp. affixes) used in their derivation. For example, in the value A-DV-2, A indicates the part-of-speech (an adjective in this case), DV indicates the derivational category (deverbal in this case), and the Arabic numeral labels the derivational type (the adjectives ending in *-(v)ší* in this case).

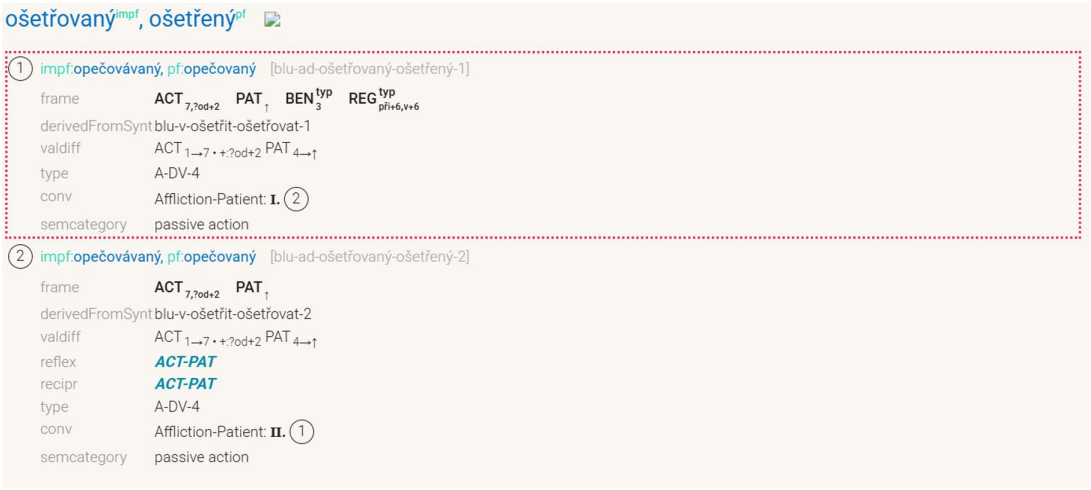


FIGURE 2 The lexical entry of the passive action adjective *ošetřovaný*^{impf}, *ošetřený*^{pf} 'treated' in *DeriVallex* with the attributes *frame* (the value of which is the valency frame), *derivedFromSynt* (the link to the respective lexical unit of the base verb), *type* (providing information on the derivational category and type of the LU), *semcategory* (introducing the word-formation category), *valdiff* (the comparison of the valency frame of the LU of the adjective and that of the LU of its base verb), *reflex* and *recipr* (introducing the pairs of the valency complements involved in reflexivity and reciprocity), and *conv* (providing information on the type of a converse relation) [https://ufal.mff.cuni.cz/derivallex/1.0/#/lexeme/lxm-ad-osetrovany-osetreny/0]

4 Conclusion

In this paper, we show how richly linguistically annotated data, such as the valency lexicons *VALLEX* and *NomVallex*, can be exploited to construct a lexical resource covering underrepresented areas. Particularly, we describe the process of automatic compiling a valency lexicon of word-formation categories standing at the boundary between inflection and derivation. The resulting data provide several opportunities for further theoretical research and applications in lexicography. First and most obviously, they may enhance lexical resources providing information on syntactic properties of words (particularly, valency lexicons) with information on noun and adjectival valency, which is not sufficiently covered in lexical resources yet. Second, given the fact that changes in valency serve as one of the reliable clues for identifying meaning shifts, the resulting data may facilitate the disambiguation of individual senses of the involved noun and adjectival deverbal derivatives. Lastly, although the formation of these derivatives is very regular, our experiment showed that it is difficult to determine their absence in corpus data. Their form-like representation, as adopted in *ASSČ* for action nouns, can thus be regarded as well justified, particularly if potential users also include language learners, who need substantially more grammatical information (see Heuberger, 2018) and for whom this representation allows such gaps to be identified. The data from *DeriVallex* then make it possible to integrate information on lemmas and the valency of these categories into explanatory dictionaries, thereby increasing their descriptive comprehensiveness.

References

Alexiadou, A. (2007). Argument structure in nominals. In A. Alexiadou, L. Haegemen & M. Stavrou (Eds.), *Noun Phrases in the Generative Perspective* (pp. 477–546). Walter de Gruyter.

Allerton, D. J. (2006). Valency grammar. In *Encyclopedia of Language & Linguistics* (pp. 301–314). Elsevier.

ASSČ. Akademický slovník současné češtiny (2017-2026). Retrieved May 20, 2026, from https://slovnikcestiny.cz

Atkins, S. B. T. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press.

Bierwisch, M. (2015). Word-formation and argument structure. In P. O. Müller, I. Ohnheiser, O. Susan & F. Rainer (Eds.), *Word-Formation: An International Handbook of the Languages of Europe, Vol. 2* (pp. 1057–1099). Walter de Gruyter.

Booij, G. (2000a). Derivation and Inflection. In G. Booij, Ch. Lehmann & J. Mugdan (Eds.), *An International Handbook on Inflection and Word-Formation, Vol. 1* (pp. 360–368). Walter de Gruyter.

Booij, G. (2000b). Inheritance. In G. Booij, Ch. Lehmann & J. Mugdan (Eds.), *An International Handbook on Inflection and Word-Formation, Vol. 1* (pp. 857–865). Walter de Gruyter.

Chomsky, N. (1970). Remarks on nominalization. In I. Jacobs & P. Rosenbaum (Eds.), *Readings in English Transformational Grammar*, *Vol. 2* (pp. 184–221). Ginn & Co.

Czech National Corpus. Retrieved May 25, 2026, from https://www.korpus.cz

DeriNet 2.3. Retrieved May 25, 2026, from https://ufal.mff.cuni.cz/derinet

DeriValllex. Retrieved May 25, 2026, from https://ufal.mff.cuni.cz/derivallex_1.0

Dokulil, M. (1962). *Tvoření slov v češtině 1. Teorie odvozování slov*. Nakladatelství Československé akademie věd.

Dokulil, M., Horálek, K., Hůrková, J. & Knappová, M. (1986). *Mluvnice češtiny I*. Academia.

Grimshaw, J. (1990). *Argument Structure*. The MIT Press.

Hacken, P. ten (2014). Delineating Derivation and Inflection. In R. Lieber & P. Štekauer (Eds.), *The Oxford Handbook of Derivational Morphology* (pp. 10–25). Oxford University Press.

Haspelmath, M. & Müller-Bardey, T. (2015). Valency Change. In G. Booij, Ch. Lehmann & J. Mugdan (Eds.), *An International Handbook on Inflection and Word-Formation*, *Vol. 2* (pp. 1130–1145). Walter de Gruyter.

Herbst, T., Heath, D., Roe I. F. & Götz, D. (2004). *A Valency Dictionary of English. A Corpus-Based Analysis of the Complementation Patterns of English Verbs, Nouns and Adjectives*. Mouton de Gruyter.

Heuberger, R. (2018). Dictionaries to Assist Teaching and Learning. In P. A. Fuertes-Olivera (Ed.), *The Routledge Handbook of Lexicography* (pp. 300–316). Routledge.

Kettnerová, V., Kolářová, V., Mírovský J. & Olbrich, M. (2026). Using Valency Inheritance in Building a Valency Lexicon. In S. Piperidis, N. Bel, H. van den Heuvel, N. Ide, S. Krek & A. Toral (Eds.), *Proceedings of the Fifteenth Language Resources and Evaluation Conference* (pp. 11506–11515). European Language Resources Association.

Kochová, P. & Opavská, Z. (Eds.) (2016). *Kapitoly z koncepce Akademického slovníku současné češtiny*. Ústav pro jazyk český.

Kolářová, V. (2024). Active and passive syntax of Czech deverbal and deadjectival nouns. *Lingua*, 307, 1–27. https://doi.org/10.1016/j.lingua.2024.103686

Kolářová, V., Vernerová, A. & Klímová, J. (2021). Systemic and non-systemic valency behavior of Czech deverbal adjectives. *Jazykovedný časopis*, 72(2), 371–382. 10.2478/jazcas-2021-0034

Kolářová, V., Vernerová A. & Verner, J. (2019). Non-systemic valency behavior of Czech deverbal nouns based on the NomVallex lexicon. *Jazykovedný časopis*, 70(2), 424–433. https://doi.org/10.2478/jazcas-2019-0071

Kulikov, L. (2011). Voice Typology. In J. J. Song (Ed.), *The Oxford Handbook of Linguistic Typology* (pp. 368–398). Oxford University Press.

Levin, B. (1993). *English Verb Classes and Alternations*. The University of Chicago Press.

MorfFlex CZ. Retrieved May 25, 2026, from http://ufal.mff.cuni.cz/morfflex

NomVallex. Retrieved May 25, 2026, from https://ufal.mff.cuni.cz/nomvallex/2.5

NomVallex 2.6. Retrieved May 25, 2026, from https://quest.ms.mff.cuni.cz/vallex/?version=data_release-NomVallex2.6

Panocová, R. & Hacken P. ten (2018). Process Nouns in Dictionaries: A Comparison of Slovak and Dutch. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (Eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts* (pp. 713–722). Ljubljana University Press.

PDT-Vallex. Retrieved May 25, 2026, from https://ufal.mff.cuni.cz/pdt-vallex-valency-lexicon-linked-czech-corpora

Svozilová, N., Prouzová, H. & Jirsová, A. (2005). *Slovník slovesných, substantivních a adjektivních vazeb a spojení*. Academia.

VALLEX 4.5. Retrieved May 25, 2026, from https://ufal.mff.cuni.cz/vallex/4.5

Walenty. Retrieved May 25, 2026, from https://zil.ipipan.waw.pl/Walenty/

Wunderlich, D. (2012). Operations on argument structure. In C. Maienborn, K. von Heusinger & P. Portner (Eds.), *Semantics. An International Handbook of Natural Language Meaning, Vol. 3* (pp. 2224–2259). De Gruyter Mouton.

Žabokrtský, Z., Ševčíková, M., Straka, M., Vidra, J. & Limburská, A. (2016). Merging Data Resources for Inflectional and Derivational Morphology in Czech. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk & S. Piperidis (Eds.), *Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016* (pp. 1307–1314). European Language Resources Association (ELRA).

Contact information

Václava Kettnerová
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
kettnerova@ufal.mff.cuni.cz

Veronika Kolářová
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
kolarova@ufal.mff.cuni.cz

Jiří Mírovský
Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
mirovsky@ufal.mff.cuni.cz

Acknowledgements

The work on the *DeriVallex* lexicon reported here was supported by the Ministry of Education, Youth and Sports of the Czech Republic, Project No. LM2023062, *LINDAT/CLARIAH-CZ* Research Infrastructure (https://lindat.cz). The work described herein has been using data and tools provided by this research infrastructure too.

Veronika Kolářová, Václava Kettnerová,
Michal Olbrich & Jiří Mírovský

NOMVALLEX

Valency of Czech Nouns and Adjectives in Networks of Derivationally Related Words

Abstract NomVallex is a valency lexicon of Czech nouns and adjectives, aimed primarily at describing the valency of individual noun and adjectival derivational categories in response to their insufficient theoretical description and coverage in lexical resources. Annotation has proceeded from the more extensively studied derivational categories (deverbal nouns and adjectives) through deadjectival nouns to the least researched denominal derivatives and compounds. Since valency is often affected by word-formation relations, NomVallex has implemented features for capturing derivational relations between derivationally related lexical units, and for comparing their valency. Although the annotation in NomVallex often encompasses more than one derivative of a particular base word, only a limited portion of the word-formation relations in Czech is covered there. The lexicon therefore newly provides links to DeriNet, where users can view a given word in the broader context of its derivational relations. Moreover, to display derivational relations and changes in valency associated with the word-formation processes in whole networks of derivationally related words recorded in NomVallex, version 2.7 has introduced a DeriNet-like graphical representation of derivational relations, which is available in a user-friendly format on the lexicon's website, facilitating further linguistic research into the influence of word formation on the valency of derived words.

Keywords Valency lexicon; nouns; adjectives; changes in valency; derivational relation; word-formation network; graph representation

1 Introduction

Derivational relations between words play a primary role in the word-formation domain but they are often reflected also in syntactic behavior of particular derivational types, especially the noun and adjectival ones, including their valency. We can observe that derived words usually share some of their valency properties with their base words (the phenomenon referred to as the so-called argument inheritance, see e.g., Booij, 2000). Valency dictionaries containing information about noun or adjectival valency (for English, see e.g., Herbst et al., 2004, for Slavic languages, e.g., Polish, see Alberski et al., 2018) usually only list the valency complements with which the nouns and adjectives combine and the forms of these complements. Concerning Czech, valency properties of nouns and adjectives are covered in several valency lexicons, namely in the printed dictionary by Svozilová, Prouzová and Jirsová (2005), and in two electronic valency lexicons, PDT-Vallex (Urešová et al., 2024) and NomVallex (Kolářová & Vernerová, 2022). While the printed dictionary and PDT-Vallex only describe the valency of selected nouns and adjectives, NomVallex goes a step further by not only recording the valency of nouns and adjectives but also capturing derivational relations between them and their base words. Moreover, it provides facilities enabling the comparison of valency frames of the derivationally related words, thus making it possible to explore changes in their valency.

Until now, the NomVallex' website has only allowed users to view differences in valency between two derivationally related lexical units, with that information displayed within the lexical entry in the attributes designated for this purpose (Section 3). However, research into the interface between word formation and valency shows that it is important to consider not only the two-sided relationship between a base word and its derivative, but also the position of the derivative within the whole word-formation network, including its indirect derivational

relationships. Consequently, there was a need to make information about valency properties of derived nouns and adjectives accessible in such a way that they could be tracked across the entire network of derivationally related words, with a particular focus on changes in valency. Extensive networks of derivationally related Czech words are captured and visualized, for example, in DeriNet (Section 2), which has thus become not only a reference source for identifying word-formation relations (Section 3.4), but also an inspiration for the graph representation of derivational relations employed in NomVallex (Section 4). In addition to the technical aspects of the used graph representation (Section 4.1), we describe its main advantages from the perspective of a lexicon user, including connecting multiple levels in a graph, user-friendly and visualized display, possibility to generate statistics and potential applications in future linguistic research (Section 4.2).

2 Derivational relations in dictionaries and lexical databases

Various monolingual dictionaries capture selected derivationally related words in the form of a word-formation tab (for Slovenian, see, e.g., Kern, 2024) or in run-on entries of their base words (for Czech, see, e.g., *Slovník spisovného jazyka českého*; for example, the entry for the adjective *uražený* 'offended' lists two of its derivatives, namely the noun *uraženost* 'offendedness' and the adverb *uraženě* 'offendedly'). However, such derivational relations are usually limited in both the types of derivatives and the number of stages of derivation kept, and moreover, they are not visualized using any graph representation.

In contrast, extensive networks of derivationally related words and their visual representation are the main topic of projects aimed at automatic modeling of word-formation relations, such as DeriNet (Olbrich et al., 2025). Figure 1 exemplifies the network of word-formation relations for the Czech noun *uraženost* 'offendedness' in DeriNet.

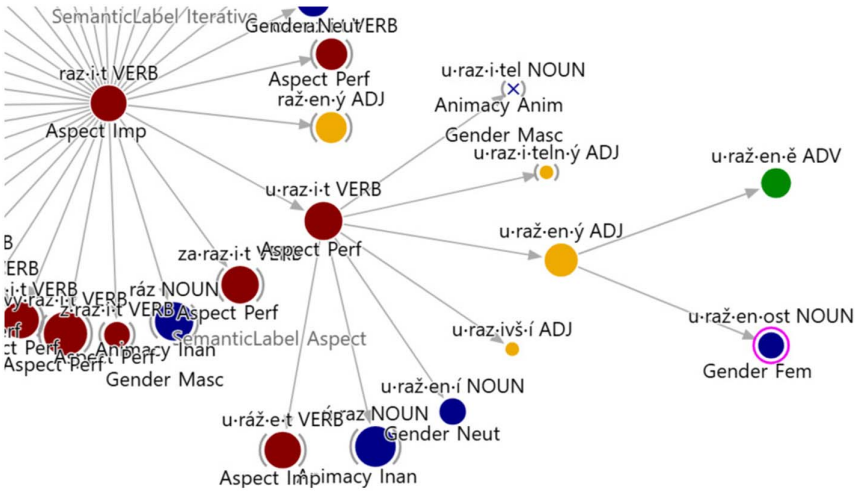


FIGURE 1 The network of word-formation relations for the noun *uraženost* 'offendedness' in DeriNet [https://quest.ms.mff.cuni.cz/derisearch2/v2/databases/]

3 Derivational relations in NomVallex

In this section, we describe the annotation strategy adopted in NomVallex with regard to capturing derivational relations and comparing the valency properties of derivationally related words.

3.1 The NomVallex lexicon

NomVallex is a manually annotated valency lexicon of Czech nouns and adjectives. Its newest version, NomVallex 2.5 (available as publicly accessible webpages and as downloadable and machine-readable data; Kolářová et al., 2024), comprises 1,337 lexical units, i.e., individual senses of nouns (N) and adjectives (A), contained in 730 lexical entries (lexemes).[1] As for derivational categories, it covers deverbal (DV), deadjectival (DA) and denominal (DN) nouns, and deverbal, denominal, deadjectival and primary (P) adjectives. Our presentation of the new functionalities implemented in NomVallex is based on a working version of the data, which also includes compounds (C) and primary nouns; the working version is available on the website (see NomVallex 2.7). Table 1 provides statistics on the number of lexical units and lexical entries in NomVallex.

TABLE 1 The statistics on lexical units and lexical entries (lexemes) in NomVallex, grouped according to their part-of-speech and derivational categories

Part-of-speech category	Derivational category		NomVallex versions					
		simplified type	version 2.5		version 2.7: all		version 2.7: only with a link to a base word	
			lexical units	lexemes	lexical units	lexemes	lexical units	lexemes
Noun	deverbal	N-DV	682	328	712	342	635	304
	deadjectival	N-DA	301	197	350	222	344	220
	denominal	N-DN	9	3	66	39	54	31
	primary	N-P	-	-	18	6	-	-
	compound	. . . +N (C)	-	-	15	13	15	13
Adjective	deverbal	A-DV	244	151	246	152	213	129
	deadjectival	A-DA	8	8	12	10	12	10
	denominal	A-DN	31	15	65	34	42	24
	primary	A-P	62	28	63	26	-	-
	compound	. . . +A (C)	-	-	10	8	10	8
Total			1,337	730	1,557	852	1,325	739

NomVallex has adopted and further modified, where necessary, the annotation scheme of the valency lexicon of Czech verbs VALLEX (Lopatková et al., 2022). The lexicon entry contains a lexeme, an abstract unit associating lexical forms with their lexical units (LUs). NomVallex applies the valency theory of the Functional Generative Description (Panevová, 1980): valency is captured in the form of valency frames assigned to individual lexical units. Each frame is modeled as a sequence of valency slots, each standing for one complement. Valency complements are supplemented with a list of morphosyntactic forms indicating their surface realization (i.e., surface valency). The following types of complements may be a part of valency frames: obligatory or optional actants (i.e., ACTor, PATient, ADDRessee, EFFect, and ORIGin, e.g., *Petrův*$_{\text{ACT}}$ *smysl pro humor*$_{\text{PAT}}$ 'Peter's sense of humor', *výrobek prodejný mládeži*$_{\text{ADDR}}$ 'product marketable to young people', *vyhláška pochybná svou legalitou*$_{\text{EFF}}$ 'regulation questionable as to

[1] NomVallex 2.5 was released under the following licence: Creative Commons - Attribution-NonCommercial-ShareAlike 4.0 International (CC BY-NC-SA 4.0).

its legality'), and obligatory free modifications (e.g., *muž povolaný do armády*$_{DIR3}$ 'a man drafted into the army'). For example, the deadjectival noun *uraženost* 'offendedness' and the deverbal adjective *uražený* 'offended' are assigned the valency frames (1) and (2), respectively. For comparison purposes, we also provide in (3) the valency frame of their motivating verb *urazit*^{pf} *se* – *urážet*^{impf} *se* 'to take offence' from VALLEX. In NomVallex, valency properties of a lexical unit are illustrated with examples from the *Czech National Corpus* (corpora SYNv14 and Araneum Bohemicum Maximum). For the noun *uraženost* 'offendedness', see Figure 2.

(1) *uraženost* 'offendedness': $ACT_{2,poss}$ $ADDR_{k+3,na+4}$ $PAT_{nad+7,z+2,že}$

(2) *uražený* 'offended': $ACT_{\uparrow}$ $ADDR_{na+4}$ $PAT_{pro+4,že}$

(3) *urazit*^{pf} *se* – *urážet*^{impf} *se* 'to take offence': ACT_{Nom} $PAT_{Ins,pro+4,že}$

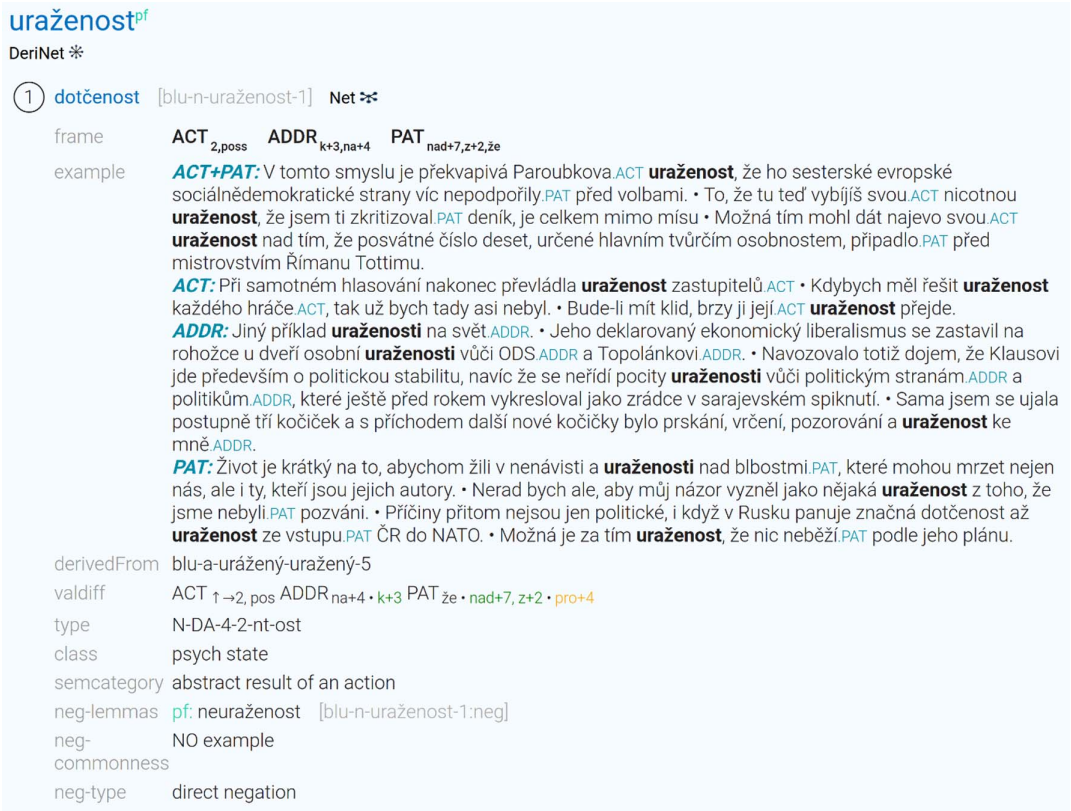


FIGURE 2 The entry for the noun *uraženost* 'offendedness' in NomVallex (derived from the adjective *uražený* 'offended', with changes in its valency specified in the valdiff attribute) [https://ufal.mff.cuni.cz/nomvallex/2.7/]

3.2 Word-formation relations reflected in valency

Derivationally related words belonging to different parts of speech often exhibit systemic changes in their surface valency. In Czech, the systemic changes are manifested primarily in the morphosyntactic forms of valency complements expressed in the base verbal constructions in the nominative or prepositionless accusative (see Figure 3 below; Kolářová et al., 2023). For example, while Actor of the verb (V) *urazit*^{pf} *se* – *urážet*^{impf} *se* 'to take offence' is expressed in the nominative, see (4), in the deverbal adjective *uražený* 'offended', this complement is subject to a systemic surface ellipsis (Kolářová et al., 2025), see (5), and in the deadjectival noun *uraženost* 'offendedness', it can be expressed either in the genitive or in the possessive form, see (6).

(4) *Petr*$_{\text{ACT-Nom}}$ *se urazil, že ho nepozvali*$_{\text{PAT}}$ 'Peter took offense at not being invited'

(5) *Petr*$_{(\uparrow)}$ *(je) uražený, že ho nepozvali*$_{\text{PAT}}$ 'Peter (is) offended that he wasn't invited'

(6) *Petrova*$_{\text{ACT-poss}}$ *uraženost, že ho nepozvali*$_{\text{PAT}}$ 'Peter's offendedness that he wasn't invited'

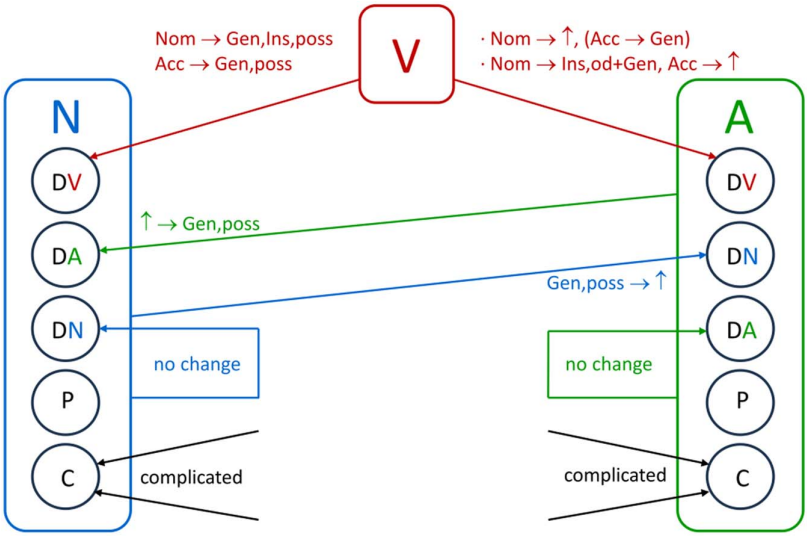


FIGURE 3 Systemic changes in surface valency reflecting word-formation relations

In addition to systemic changes in the valency of nouns and adjectives, various non-systemic changes may also occur. For example, the adjective *uražený* 'offended' has an Addressee in its valency frame presented in (2) which does not appear in the valency frame of the base verb *urazit*^{pf} *se* – *urážet*^{impf} *se* 'to take offence' in (3). Moreover, the morphosyntactic forms of this actant differ slightly between the adjective and the derived noun *uraženost* 'offendedness', cf. (2) and (1).

3.3 Derivational relations and valency changes in NomVallex

NomVallex provides lexicographic software allowing for linguistic research into various phenomena related to noun and adjectival valency, including comparison of valency properties of derivationally related lexical units. To enable analysis of the relationship between the valency behavior of base words and their derivatives, lexical units of nouns and adjectives in NomVallex are linked to their respective base lexical units (contained either in NomVallex itself or, in the case of verbs, in the VALLEX lexicon; approximately 85% of the lexical units in NomVallex 2.7 are linked, see Table 1 for statistics).[2] The links are captured by two attributes, keeping both directions, namely:

(i) the attribute derivedFrom (providing the link(s) from a particular LU to its base LU)

(ii) the attribute derivedLUs (capturing a set of the links to all LUs derived from the base LU)

[2] There are several reasons why the link is missing for some lexical units, e.g., it is a primary noun or adjective (e.g., *síla* 'strength'), the base verb of a noun or an adjective is not frequent enough to be captured in the VALLEX lexicon where the verbs were selected based on their corpus frequency (e.g., *dychtit* 'crave' > *dychtivý* 'avid'), or the base noun is avalent, so it is not treated in the NomVallex lexicon (e.g., *skála* 'rock, cliff' > *úskalí* 'difficulty').

In addition, each lexical unit of a noun or an adjective with the link to its respective base LU is automatically supplemented with information on differences between valency frames of the two LUs (specified in the valdiff attribute); namely, the number and types of valency complements and their morphosyntactic forms are automatically compared.[3]

When comparing the valency of derivationally related lexical units, it is not only information about their part of speech that proves relevant, but also information about their specific derivational type, reflecting the word-formation process involved (Section 4.2). For example, when describing valency of Czech abstract deverbal nouns, two types are distinguished, namely (i) stem nouns (ending in *-ní / -tí* and containing a theme suffix of the base verb, e.g., *uražení*^{pf} – *urážení*^{impf} 'offending'), and (ii) root nouns (derived from verbs by various suffixes, including the zero suffix, but not containing a theme suffix, e.g., *urážka*$^{no\text{-}aspect}$ 'offence'). In NomVallex, stem nouns are assigned the derivational type N-DV-1-stem and root nouns are labeled N-DV-2-root, which is specified in the type attribute.

The attributes capturing derivational type, derivational relations and changes in valency are illustrated with the entry for the deadjectival noun *uraženost* 'offendedness' in Figure 2.

3.4 Linking NomVallex to DeriNet

DeriNet models derivational relations between whole lexemes, using semi-automatic procedures (based on a combination of rule-based methods and machine learning), followed by a manual check (Ševčíková et al., 2016). In contrast, NomVallex manually captures derivational relations at the level of lexical units, not lexemes, i.e., between lexical units of base words on the one hand and lexical units of derived nouns and adjectives on the other. However, in the case that the derivational history of lexical units is not clear (and even the scholarly literature does not provide a definitive answer as to what the base word of these units is), pragmatic decisions were made in the annotation process aimed at ensuring data consistency, particularly, a certain direction of derivation was selected consistently across individual derivational types so that changes in their valency could be observed. As examples, we can cite suffixless nouns that are related to verbs, where it is unclear whether the noun was derived from the verb or vice versa (e.g., *boj* 'fight' – *bojovat* 'to fight'; Ševčíková, 2021), or nouns ending in *-ství/-ctví* (e.g., *zahradnictví* 'garden centre'; Šimandl, 2016), where it is unclear whether they are motivated by a noun (e.g., *zahradník* 'gardener', *zahrada* 'garden') or an adjective (e.g., *zahradní* 'garden'). As a result, the annotation describing the derivational relationships in NomVallex may differ from those recorded in DeriNet. For example, the noun *boj* 'fight' is annotated in NomVallex as the deverbal noun derived from the verb *bojovat* 'to fight', while in DeriNet the relationship between the noun and the verb is the opposite, with the verb treated as denominal, i.e., derived from the noun *boj* 'fight'. Another example includes the noun *vlastnictví* 'ownership', which is annotated in NomVallex as denominal (derived from the noun *vlastník* 'owner'), while in DeriNet it is recorded as deadjectival (derived from the adjective *vlastnický* 'property'). Moreover, though capturing derivational relations as described in Section 3.3, NomVallex still covers just a limited portion of a derivational network.

It therefore appears reasonable to link NomVallex to DeriNet so that the lexicon users can see the network of derivational relations within which a word in DeriNet is situated and assess its position within that network – not only in terms of the extent of the network in which the word

[3] Compounds are an exception, as their valency frames cannot be automatically compared with those of their base lexical units (see Section 4.2.2).

is contained, but also, where applicable, in terms of its otherwise conceived and recorded word-formation history.

Providing links from NomVallex to DeriNet, the following issues need to be taken into account: (i) the difference in the treatment of polysemous nouns and adjectives in NomVallex and in DeriNet (in NomVallex, their individual lexical units are differentiated, while in DeriNet they are not distinguished; this issue is briefly discussed in Section 4.2.4), and (ii) the fact that nouns and adjectives in NomVallex contained in a single lexical entry can be represented by more than one lemma (this includes especially noun or adjectival aspectual counterparts, e.g., *pociťování*^{impf} – *pocítění*^{pf} 'feeling', or orthographic and other variants of the lemma that are interchangeable without changing the meaning, e.g., *dotyk* 'touch' and *dotek* 'touch').

Links from NomVallex to DeriNet are one of the new features of NomVallex 2.7, with links handled as follows: Each lemma representing individual lexical units of nouns or adjectives in NomVallex includes a link to a search query for the respective lemma in DeriNet (see the icon labeled DeriNet in Figure 2).

4 Graph representation of word-formation relations in NomVallex

In order to display derivational relations and changes in valency associated with the word-formation processes in whole networks of derivationally related words in NomVallex, we compile (i) the attributes by which NomVallex represents derivational relations and valency changes (i.e., derivedFrom and valdiff, respectively; Section 3.3), and (ii) the DeriNet-like graphical representation of word-formation relations (Section 2). Technical aspects of the graph representation are described in Section 4.1 and its benefits for lexicon users are discussed in Section 4.2.

4.1 The graph representation in technical terms

The technical aspects of the new graph representation of word-formation relations in NomVallex and related valency changes concern (i) data formats and availability, and (ii) conversion of NomVallex to graph format.

(i) Data formats and availability

The graph data are stored in a JSON format following the node–edge convention of the Cytoscape.js visualization framework. Each lexical unit is represented as a compound node consisting of a main node carrying core attributes (identifier, lemma, and part of speech) and three child nodes encoding the gloss, the valency frame and the type, respectively. This structure allows the web interface to render each of these components as a separate, visually distinct panel. Edges are directed and typed, with the two primary relation types being DERIVED_FROM, capturing the derivational hierarchy, and HAS_LEXEME, linking a lexical unit to its parent lexeme.[4]

[4] To maximize accessibility for a broad audience, the data will be released in two formats: the full Cytoscape.js-compatible JSON file, intended for use with the accompanying web interface, and a pair of plain CSV files — one for nodes with their attributes and one for the typed, directed edge list — suitable for use with general-purpose tools such as Pandas or R.

(ii) Conversion of NomVallex to graph format

The source data of NomVallex are distributed in two complementary formats: lexemes and lexical units are provided in XML, while valency frames are available in plain-text files only. The conversion pipeline begins by parsing the XML into an intermediate Python dictionary that captures the hierarchical structure of lexemes and their associated lexical units along with key attributes, including the gloss, the derivedFrom reference, and the valdiff attribute encoding differences in valency. In the following step, this structured representation is flattened into a list of lexical units, which are then enriched with valency frames loaded from the text files. Reverse derivation links are subsequently computed by inverting the derivedFrom relation, yielding a hasDerivations index. Finally, the complete node–edge structure is assembled and serialized both to the Cytoscape.js JSON and to the two CSV files described above.

4.2 Key advantages of the graph representation

The new graph representation offers NomVallex users the following key benefits: connecting multiple levels in a graph (Section 4.2.1), user-friendly visualization (Section 4.2.2), an easier way to generate statistics for different graph variants (Section 4.2.3), and the possibility of using graphical representations or statistics in further linguistic research (Section 4.2.4).

4.2.1 Connecting multiple levels in a graph

Until now, NomVallex has only allowed users to follow a two-sided relationship between a base lexical unit and a derived lexical unit, via the attributes derivedFrom and derivedLUs (see Section 3.3). The new graph representation connects sequences of direct derivational relations into a network of these relationships. This presentation of the lexicon data is a significant advancement, as it allows nouns and adjectives to be viewed within entire networks of derivationally related words recorded in the lexicon, regardless of whether there is a direct derivational relationship between them (cf., the deadjectival noun *uraženost* 'offendedness' and the deverbal noun *uražení*^{pf} *se* – *urážení*^{impf} *se* 'taking offence' in Figure 4). In addition, users can compare these networks with those in DeriNet (Section 3.4), which further enhances their understanding of a given derivative's position within networks of derivationally related words.

4.2.2 User-friendly visualization

In NomVallex 2.7, it is possible to visualize the graphs in a user-friendly interface directly on the lexicon's website. The graph appears after clicking on the icon labeled 'Net' (located to the right of the identifier of the lexical unit; Figure 2), and the user can further customize it by dragging the nodes. In the graphs, nodes in a direct derivational relationship are connected by an edge, with the direction of the derivation indicated by an arrow. Each node represents one lexical unit, distinguished by color indicating its part of speech (verbs – red, adjectives – green, nouns - blue). Further, each lexical unit is supplemented with its near paraphrase (the gloss), its derivational type and its valency frame. The comparison of the valency frames of the lexical units involved in the direct derivational relation is attached to the edge connecting these units (legend: = no change, -> systemic change, + a new form, - a missing form; for the noun *uraženost* 'offendedness', see Figure 4).[5]

[5] A facility allowing users to search and display a base lexical unit together with all lexical units directly derived from it was already available in the Blue Search Engine (BlueSE, the 'vallex-like lexicons search tool'; see also Kolářová et al., 2023), but not as a structured network.

The graphs also display instances where two or more base lexical units are listed in the attribute derivedFrom. This situation systematically arises in the case of compounds and in the cases where two base lexical units are considered and it is not possible to choose between them unambiguously (e.g., some reflexive and non-reflexive base verbs). As for compounds, the links to the base lexical units corresponding to the compound's constituents are provided (e.g., *ohleduplný* 'thoughtful', providing links to the noun *ohled* 'consideration' and the adjective *plný* 'full'). However, it has turned out that the relationship between the valency of the compounds and the valency of their constituents is rather complicated. Therefore, in the case of compounds, the comparison of the involved valency frames cannot be made automatically, and at the edges of the graph, where changes in valency are typically indicated, only 'NA' is listed.

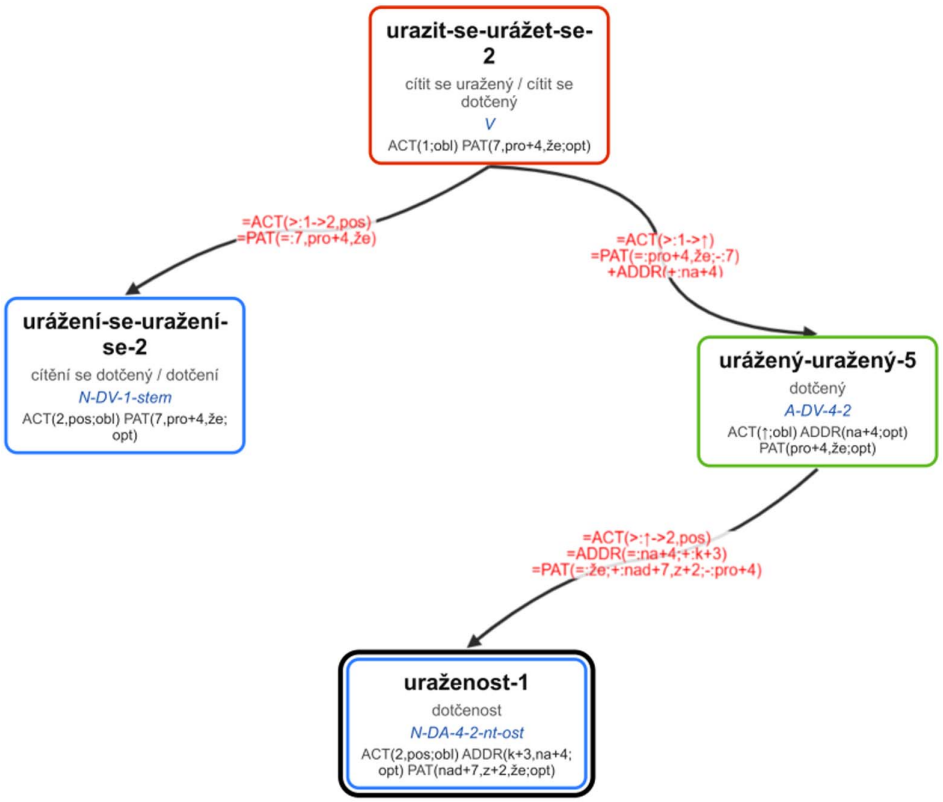


FIGURE 4 The graph representation of the network of derivational relations for the noun *uraženost* 'offendedness' [https://ufal.mff.cuni.cz/nomvallex/2.7/]

4.2.3 Statistics for different graph variants

The new graph representation makes it easier to generate statistics for the derivational networks covering the NomVallex data, at various levels of detail. Statistics based on the occurrence of nodes in the networks (Table 2) show that nouns and adjectives from NomVallex appear in a total of 626 graphs, which together contain 1,998 lexical units (including relevant lexical units of the base verbs obtained from the VALLEX lexicon). The networks contain on average 3 nodes (there are 133 graphs consisting of only 1 node, while the most extensive network covers 46 nodes).[6]

[6] The reason why the largest network is so extensive is that it includes compounds one constituent of which is the adjective *hodný* 'worthy', e.g., *důvěryhodný* 'trustworthy', *obdivuhodný* 'admirable', and covers also the networks for the individual constituents of the compounds, e.g., the noun *důvěra* 'trust' and the adjective *hodný* 'worthy' in the case of the compound *důvěryhodný* 'trustworthy'.

TABLE 2 Statistics on the networks in NomVallex

Total number of the networks	626
Total number of lexical units from NomVallex or VALLEX	1,998
Average number of nodes in a network	3.19

Networks covering the NomVallex data contain various graph variants in terms of the derivational types represented. For example, the graph shown in Figure 4, i.e., the network *urazit*^{pf} *se* – *urážet*^{impf} *se* 'to take offence' (*uražený* 'offended' (*uraženost* 'offendedness'), *uražení*^{pf} *se* – *urážení*^{impf} *se* 'taking offence'), with direct derivational relations enclosed in parentheses, represents the following combination of derivational types: V (A-DV-4-2 (N-DA-4-2-nt-ost), N-DV-1-stem). The statistics on the most frequent combinations of derivational types in individual graphs, as presented in Table 3, reflect the proportion of each derivational category described in the lexicon (e.g., while deverbal nouns are the most frequent, denominal derivatives, for example, are still represented to a lesser extent; see Table 1).[7]

TABLE 3 Statistics on the combinations of derivational types in the graphs (the combinations that appeared at least nine times in the data)

Frequency	Derivational types in a graph	Examples of nouns or adjectives on the lowest level of the networks
107	V (N-DV-1-stem)	zeptání-se-1 'asking'
45	N-DV-1-stem	podezření-2 'suspicion'
45	V (N-DV-1-stem, N-DV-2-root)	pociťování-pocítění-1 'feeling', pocit-1 'feeling'
27	N-DV-2-root	důvtip-1 'ingenuity'
22	A-P	prostý-1 'simple'
20	V (N-DV-1-stem, N-DV-1-stem)	přání-2 'wishing', přání-3 'congratulations'
20	A-P (N-DA-O)	nervozita-1 'nervousness'
15	V (N-DV-1-stem, N-DV-2-root, N-DV-2-root)	chválení-1 'praising', chvála-1 'praise', chvála-3 'praise(s)'
12	V (N-DV-2-root)	boj-3 'fight'
10	A-DV-3 (N-DA-3-lost)	vzteklost-1 'furiousness'
9	A-DV-6 (N-DA-6-vn-ost)	lačnost-1 'greed'
9	N-P	mysl-2 'opinion'
9	V (A-DV-6 (N-DA-6-vn-ost))	odolnost-1 'resistance'
9	A-P (N-DA-O, N-DA-O)	chudoba-1 'poverty', chudost-3 'poverty'

[7] Labels for deverbal adjectives (A-DV) use numbers to distinguish the type of the verb form from which they originate; for instance, number 3 marks adjectives originating from active participles, e.g., *vzteklý* 'furious' < *vztekal (se)* 'he was furious' gets the label A-DV-3. Moreover, labels for nouns derived from deverbal adjectives are supplemented by the segment they typically end in (e.g., the nouns labeled as N-DA-3-lost are derived from deverbal adjectives of type 3 and they mostly end in *-lost*, i.e., *vzteklost* 'furiousness'). Nouns derived from adjectives other than the deverbal ones, e.g., *chudoba* 'poverty' < *chudý* 'poor', get the label N-DA-O.

Focusing primarily on valency, NomVallex does not aim to capture complete networks of derivationally related lexical units; therefore, unlike DeriNet networks, the NomVallex' graphs necessarily contain gaps. As a result, they do not provide an accurate picture of how frequently individual combinations of derivational categories occur in the language. For example, some of the graphs containing in NomVallex only a verb and a root noun, i.e., V (N-DV-2-root), e.g., the verb *bojovat* 'to fight' and the root noun *boj* 'fight', could also involve a stem noun (e.g., *bojování* 'fighting'), which, however, has not been included in the lexicon yet.

4.2.4 A data source for further linguistic research

In contrast to DeriNet, the word-formation relations in NomVallex are captured at the level of lexical units, not lexemes (lemmas), which ensures that individual meanings of derivatives and those of their base words that correspond to each other are linked. Moreover, unlike DeriNet, NomVallex specifies the derivational type of each derivative (Section 3.3). The NomVallex data show that different meanings of a base word may differ in the number of derivational types of derivatives associated with them. For example, the verb *urazit* has three meanings, namely 'to knock off', 'to offend', and 'to travel'. All these meanings give rise to the stem noun *uražení* 'knocking off, offending, traveling' and the adjective *uražený* 'knocked off, offended, covered'. However, the second derivational type of a deverbal noun associated with this verb, i.e., the root noun *urážka* 'offense', only corresponds to the verbal meaning 'to offend'. In addition to the observation that while stem nouns are usually formed for all verb meanings, root-nouns are formed only for some of them, the NomVallex data allow us to examine in which specific cases this occurs and whether it is related in any way to valency or to other valency-related phenomena.

Another example of how the graph representation of derivational relations in NomVallex can be used for further linguistic research involves cases of rare combinations of derivational types in derivatives motivated by nouns. For example, there are various derivatives motivated by the deverbal noun *vláda* 'government' that are formed using different word-formation processes, namely denominal nouns *nadvláda* 'dominance' (formed by prefixation) and *bezvládí* 'anarchy' (formed by circumfixation), and compounds, e.g., *hrůzovláda* 'reign of terror' and *samovláda* 'autocracy' (formed by composition, with the first constituent being either a noun, e.g., *hrůza* 'horror, terror', or a pronoun, i.e., *sám* 'oneself'). In addition to the specification of particular derivational types of the derivatives in NomVallex, including the word-formation processes that apply to individual derivatives (e.g., the denominal noun *nadvláda* 'dominance' formed by prefixation gets the type N-DN-pref), the main contribution of the graph representation of derivational networks in NomVallex lies in the fact that it allows users to examine and compare the valency behavior of these diverse derivational types.

5 Conclusion

In this paper, we presented new features in the NomVallex lexicon related to the representation of derivational relations and the possibility of comparing the valency of derivationally related words. Derivational relations that have already been recorded in NomVallex can now be visualized using a graph representation which makes it possible to view nouns and adjectives within entire networks of derivationally related words recorded in the lexicon, regardless of whether there is a direct derivational relationship between them. Using the newly provided links to DeriNet, users can also compare the position of a given word within the broader context of word-formation networks, including any differing interpretations of its derivational history. The graph representation of derivational relations employed in NomVallex also makes it easy to identify and examine changes in valency associated with word-formation relations;

overcoming the original limitation to lexical units that share a direct derivational relationship, the new feature significantly expands the possibilities for studying changes in valency across different parts of speech and derivational types.

References

Alberski, B. et al. (2018). *Walenty*. Institute of Computer Science, Polish Academy of Sciences. Retrieved from http://hdl.handle.net/11321/586

Blue Search Engine (BlueSE). Retrieved May 30, 2026, from https://quest.ms.mff.cuni.cz/vallex/

Booij, G. (2000). Inheritance. In G. Booij, Ch. Lehmann & J. Mugdan (Eds.), *An International Handbook on Inflection and Word-Formation, volume 1* (pp. 857–865). Walter de Gruyter.

Czech National Corpus. Retrieved May 30, 2026, from https://www.korpus.cz/

Herbst, T., Heath, D., Roe, I. F. & Götz, D. (2004). *A Valency Dictionary of English. A Corpus-Based Analysis of the Complementations Patterns of English Verbs, Nouns, and Adjectives*. Walter de Gruyter.

Kern, B. (2024). Considering Word Formation in Compiling Dictionaries. In K. Š. Despot, A. Ostroški Anić & I. Brač (Eds.), *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress* (pp. 463–475). Institute for the Croatian Language.

Kolářová, V., Kettnerová, V., Klímová, J., Mírovský, J. & Vernerová, A. (2024). *NomValex 2.5*. LINDAT/CLARIAH-CZ digital library at ÚFAL, MFF UK. Retrieved from http://hdl.handle.net/11234/1-5826.

Kolářová, V., Kettnerová, V. & Mírovský, J. (2023). Through Derivational Relations to Valency of Non-verbal Predicates in the NomVallex Lexicon. *Jazykovedný časopis*, 74(1), 182–192. https://doi.org/10.2478/jazcas-2023-0036

Kolářová, V. Kettnerová, V. & Mírovský, J. (2025). Advanced Syntactic Phenomena in the NomVallex Lexicon. *Jazykovedný časopis*, 76(1), 322–332. https://doi.org/10.2478/jazcas-2025-0029

Kolářová, V. & Vernerová, A. (2022). NomVallex: A Valency Lexicon of Czech Nouns and Adjectives. In *Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022)* (pp. 1344–1352). European Language Resources Association.

Lopatková, M., Kettnerová, V., Mírovský, J., Vernerová, A., Bejček, E. & Žabokrtský, Z. (2022). *VALLEX 4.5*. LINDAT/CLARIAH-CZ digital library at ÚFAL, MFF UK. Accessible at: http://hdl.handle.net/11234/1-4756.

NomVallex 2.7. Retrieved May 30, 2026, from https://ufal.mff.cuni.cz/nomvallex/2.7/

Olbrich, M., Brezinová, V., Dohnalová, Š., John, V., Kyjánek, L., Papáček, A., Svoboda, E., Ševčíková, M., Vidra, J. & Žabokrtský, Z. (2025). *DeriNet 2.3*. LINDAT/CLARIAH-CZ digital library at ÚFAL, MFF UK. Accessible at: http://hdl.handle.net/11234/1-5846.

Panevová, J. (1980). *Formy a funkce ve stavbě české věty*. Academia.

Slovník spisovného jazyka českého. Retrieved May 30, 2026, from https://ssjc.ujc.cas.cz/

Svozilová, N., Prouzová, H. & Jirsová, A. (2005). *Slovník slovesných, substantivních a adjektivních vazeb a spojení*. Academia.

Ševčíková, M. (2021). Action nouns vs. nouns as bases for denominal verbs in Czech: A case study on directionality in derivation. *Word Structure*, 14(1), 97–128. https://doi.org/10.3366/word.2021.0181

Ševčíková, M., Žabokrtský, Z., Vidra, J. & Straka, M. (2016). Lexikální síť DeriNet: elektronický zdroj pro výzkum derivace v češtině. *Časopis pro moderní filologii*, 98(1), 62–76.

Šimandl, J. (Ed.) (2016). *Slovník afixů užívaných v češtině*. Univerzita Karlova, Karolinum.

Urešová, Z., Bémová, A., Fučíková, E., Hajič, J., Kolářová, V., Mikulová, M., Pajas, P., Panevová, J. & Štěpánek, J. (2024). *PDT-Vallex: Czech Valency lexicon linked to treebanks 4.5*. LINDAT/CLARIAH-CZ digital library at ÚFAL, MFF UK. Retrieved from http://hdl.handle.net/11234/1-5814.

Contact Information

Veronika Kolářová
Institute of Formal and Applied Linguistics, Charles University, Faculty of Mathematics and Physics
kolarova@ufal.mff.cuni.cz

Václava Kettnerová
Institute of Formal and Applied Linguistics, Charles University, Faculty of Mathematics and Physics
kettnerova@ufal.mff.cuni.cz

Michal Olbrich
Institute of Formal and Applied Linguistics, Charles University, Faculty of Mathematics and Physics
olbrich@ufal.mff.cuni.cz

Jiří Mírovský
Institute of Formal and Applied Linguistics, Charles University, Faculty of Mathematics and Physics
mirovsky@ufal.mff.cuni.cz

Acknowledgements

The research reported in the paper was supported by the Czech Science Foundation under the project 25-16716S. The work described herein has been using data and tools provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

Benedikt Perak, Dragana Špica & Irena Srdanović

A GRAPH-BASED LEXICOGRAPHIC WORKFLOW FOR JAPANESE ADJECTIVES AND THEIR LEXICAL RELATIONS IN AN AI TUTOR

Abstract This study presents a graph-based workflow for extending the lexicographic description of Japanese adjectives with large language model (LLM) relation candidates and expert review. JMdict provides the principal dictionary evidence. Lexical-item nodes carry written form, reading, grammatical category, English glosses, learner information, and provenance; relations record a proposed type together with semantic, register, domain, scalar, provenance, and review information. Written form plus hiragana reading serves as an operational lookup key that distinguishes homographs such as 辛い (*karai* 'spicy') and 辛い (*tsurai* 'painful'). The pilot stored 1,960 proposal observations from *gpt-4.1-mini* and 2,890 from *gpt-5.4-mini* across 400 source records representing 305 distinct items; the latter total includes a 101-observation preliminary run. An exploratory run added 76 observations. Three linguists supplied 420 judgments on 214 proposals: 314 judgments combined acceptance with an overall score of at least four, and 180 proposals received at least one such judgment. The results describe the reviewed subset and establish the feasibility of preserving dictionary evidence, model observations, target-resolution outcomes, and expert judgments as distinct components of a lexicographically interpretable resource.

Keywords Japanese adjectives; lexicography; lexical relations; JMdict; reading-aware identity; expert review; lexical graph

1 Introduction

Japanese adjectives pose several familiar lexicographic questions in an especially concentrated form. Inflecting *i*-adjectives and copula-dependent *na*-adjectives do not constitute a simple binary inventory: many *na*-adjectival expressions also function as nouns, while some forms used attributively are of verbal origin (Backhouse, 2004; Nishiyama, 1999; Takahashi, 2003). The Japanese writing system introduces a further distinction between graphic form and lexical identity. The spelling 辛い, for example, is ambiguous because it can represent 2 distinct lexical items - *karai* 'spicy' and *tsurai* 'painful'. Consequently, a relation recorded for the written form alone may be assigned to the wrong lexical item.

Japanese orthography combines kanji, characters of Chinese origin associated with lexical meaning, together with phonetic scripts, hiragana and katakana, which capture the sound. Words written with kanji can generally also be transcribed phonetically in hiragana according to their reading. Adjective 辛い is a homograph of 2 distinct lexical items: からい (*karai*, 'spicy') and つらい (*tsurai*, 'painful'). In the present study, the term *hiragana reading* refers to this phonetic representation of a lexical item in hiragana. Recording the reading is therefore particularly important both for learners of Japanese and for the reliable identification of lexical items.

Kanji and hiragana are frequently combined within the same word. Hiragana is used, among other functions, to represent inflectional endings and grammatical elements such as particles. The word 辛い illustrates this combination: 辛 is the kanji component, while is the inflectional ending characteristic of *i*-adjectives. In this function, the hiragana character functions as *okurigana*.

The present study integrates LLM candidate generation into a lexicographic workflow. Dictionary records supply the identity and grammatical description of lexical items; an LLM proposes possible relations; written form- and reading-based resolution associates proposed targets with the lexical inventory; and expert linguists evaluate the semantic relation and its classification. The graph connects these layers while preserving their different evidential status (i.e. how they were chosen).

This evidence-layered design enables an editor to trace a relation from its dictionary-derived source record, through the model's proposal and the matching of its proposed target against the lexical inventory, to the final expert judgment. It also keeps semantic assessment separate from formal target identification (e.g. POS[1]), which is particularly important for homographs, polysemous items, and expressions with overlapping grammatical uses.

The study has three aims. First, it describes a reading-sensitive representation of Japanese lexical items and explains the lexicographic function of the information stored in the nodes and relations of the lexical graph. Second, it characterizes a pilot in which two LLMs proposed relations for a coverage-oriented set of adjectives and three experts reviewed part of the resulting material. Third, it identifies the editorial steps required to prepare reviewed relations for learner-oriented use. The analysis is descriptive and centres on the operation of the workflow and the reviewed subset.

2 Lexicographic Sources and Scope

2.1 Sources and grammatical scope

JMdict is a maintained multilingual Japanese dictionary distributed as structured XML (Breen, 1995). The inventory used in this study was imported from the JMdict_english_with_examples.zip asset in the automated GitHub release dated 15 March 2026 (JMdict.2026-03-15; Yomidevs, 2026) and loaded into the project database on 25 March 2026. Entries contain written forms, readings, senses, part-of-speech and usage labels, cross-references, and antonym references. It supplies the principal evidence for lexical identity and grammar, while source relations remain distinguishable from model-generated proposals.

JLPT (Japanese-Language Proficiency Test), VDRJ (Vocabulary Database for Reading Japanese) (Matsushita, 2011), and the learner-oriented database described by Sunakawa et al. (2012) contribute proficiency, frequency, dispersion, and usage information for sampling and pedagogical filtering. The pilot covers N5–N1 proficiency levels and the project's *i*-adjectival and *na*-adjectival/*keijōshi* strata. JMdict (Japanese-English) and UniDic POS information are retained on each record so that inflecting adjectives, nominal uses, suffixes, and other adjective-like expressions can be distinguished without imposing one theoretical analysis (Backhouse, 2004; Nishiyama, 1999; Srdanović, 2019; Takahashi, 2003). Because no official JLPT vocabulary lists have been published since the 2010 revision, the JLPT levels used in this resource are obtained from Wiktionary and treated as heuristic learner metadata.

[1] Determining the POS of a word in Japanese is not trivial for many reasons. We will focus only on one aspect of adjectives here for the lack of space. E.g. as a near synonym for an -i adjective 丸い *marui* 'round' a noun+genitive marker 球状の *kyūjō no* 'spheric' was offered as a candidate. It is not uncommon to classify certain nouns followed by genitive particle *no* 'of' as adjectives (Nakamura, 2023). Depending on the position, this target can be viewed as fully adequate (if one includes some nouns followed by no into adjectives) or as semantically adequate, but formally non-adequate because the target word is neither *-i* nor *-na* adjective.

2.2 Lexical graphs and typed relations

WordNet-style resources represent lexical meaning through structured relations rather than isolated entries (Fellbaum, 1998). Japanese WordNet provides an important point of comparison (Isahara et al., 2008), while BabelNet demonstrates the scale obtainable by linking heterogeneous multilingual resources (Navigli & Ponzetto, 2012). Earlier work described LLM-supported construction of a Japanese multi-relation network of nouns (Špica & Perak, 2025). The contribution examined here is a reading-aware editorial workflow that preserves dictionary evidence, original model observations, target-resolution outcomes, and expert judgments as separately interpretable records for Japanese adjectives.

A property graph was chosen because both lexical items and their relations require structured description (Perak & Ban Kirigin, 2023). Nodes represent lexical records, while edges represent typed relations and carry the qualifications needed for their interpretation. This is particularly useful for Japanese, where the often non-transparent relationship between written form and reading, together with grammatical behaviour, sense, register, and provenance, informs lexical analysis.

Table 1 summarizes the node attributes most relevant to lexicographic interpretation in line with the key learner needs identified in Srdanović & Špica (2021). Their population depends on the contributing source, so individual records contain the fields available for that item. Internal field names connect the lexicographic description to the implemented representation.

TABLE 1 Lexicographic information represented on lexical-item nodes

Information category	Typical attributes	Lexicographic function
Headword form	standard orthography, lemma	identifies the citation form shown to the user and the normalized lemma used for retrieval
Reading and transliteration	hiragana reading, katakana reading, romanization	distinguishes homographs, supports pronunciation, and permits searches across Japanese scripts and romanization
Grammatical description	JMdict POS,UniDic lemma and POS components	records the source analysis and supports distinctions such as *i*-adjective, *na*-adjectival expression, noun, and suffix
Meaning	primary English gloss, additional translations, definition tags	supplies a concise orientation to the lexical item and the sense material needed during review
Learner information	JLPT level, VDRJ learner level, frequency, dispersion, and usage range	supports learner-sensitive selection
Lexical characteristics	word origin and word type	distinguishes native, Sino-Japanese, borrowed, and other lexical strata where the source provides this information
Provenance	JMdict identifier, source name, source identifier, import version, and revision	allows an editor to trace each piece of information to the resource from which it was obtained

Relations are described independently from their endpoint nodes. Table 2 groups the implemented proposal and relation fields by lexicographic purpose. Field use follows the relation category: scale positions apply to gradable antonyms and intensity relations, for example, while orthographic variants primarily require form and provenance information. Model confidence, relation weight, and resolution confidence are system-generated numeric indicators; reviewed status records the editorial workflow state, and symmetry records whether direction affects interpretation.

TABLE 2 Lexicographic and evidential information represented on relation edges

Information category	Typical attributes	Lexicographic function
Relation classification	relation type and category	states whether the proposed relation is synonymy, near-synonymy, antonymy, scalar intensity, register variation, collocation, or another controlled category
Semantic account	shared meaning, semantic distinction, usage context, and conditions favouring either item	explains both the common semantic ground and the contrast that prevents two related expressions from being treated as freely interchangeable
Register and domain description	source and target register, formality contrasts, stylistic labels, domain labels and their weights	specifies the concrete usage properties of each item and explains any register or domain contrast, whether or not the model labelled the proposal as a register variant
Scalar structure	scale dimension and the relative positions of source and target	represents oppositions such as 低い (*hikui* 'low')–高い (*takai* 'high') on a specified dimension
Directionality	symmetry and source–target orientation	distinguishes symmetric relations from relations in which direction or ordering matters
Strength and status	confidence, weight, resolution confidence, and reviewed status	separates model confidence and successful target identification from subsequent expert approval
Provenance	dictionary evidence, source sense references, model, provider, prompt version, and date	shows whether a relation derives from dictionary data, model generation, or more than one source and makes the proposal traceable

Complete proposal observations and judgments are retained in the experiment records. The graph presents an aggregated directed edge for an ordered source–target–type combination with summary provenance. Reverse-direction proposals and conflicting semantic accounts are handled through editorial normalization. Agreement between models adds provenance, while expert validation remains a separate editorial attribute.

2.3 *From a dictionary entry to a graph lookup*

The operational lookup key is the pair (standard orthography, reading in hiragana). Table 3 illustrates why both fields are needed.

TABLE 3 Reading-aware representation of a homograph

Written form	Reading and transliteration	POS evidence	English gloss
辛い	からい (*karai*)	*i*-adjective	'spicy; salty'
辛い	つらい (*tsurai*)	*i*-adjective	'painful; hard to bear'

The written form 辛い corresponds to distinct lexical items, including からい (*karai* 'spicy; salty') and つらい (*tsurai* 'painful; hard to bear'). A spelling-only lookup conflates these items and can attach a proposed relation to the wrong meaning. The workflow therefore uses written form plus hiragana reading as its operational lookup key and presents grammatical and sense information during review.

For relation work, both the source and proposed target retain written form and reading. The implemented resolver normalizes orthographic and reading variants, retrieves candidates matching either form or reading, filters them by relevant grammatical category, and ranks the remaining candidates. A "resolved" outcome denotes a unique top-ranked candidate. The original target form is preserved alongside the resolver outcome, allowing an editor to compare the proposal with its interpretation in the resource.

3 Lexicographic Proposal and Review Procedure

The workflow follows the sequence of decisions ordinarily required when a candidate relation is considered for a lexical resource. First, a source item is selected from the dictionary-derived inventory. Second, the model proposes one or more related expressions and assigns a provisional relation category. Third, the proposed target is checked against the lexical inventory using both writing and reading. Finally, a linguist evaluates the target, the relation type, the reading, and the usefulness of the explanation. Figure 1 shows how dictionary evidence, model proposals, formal checks, and expert review remain connected but distinct.

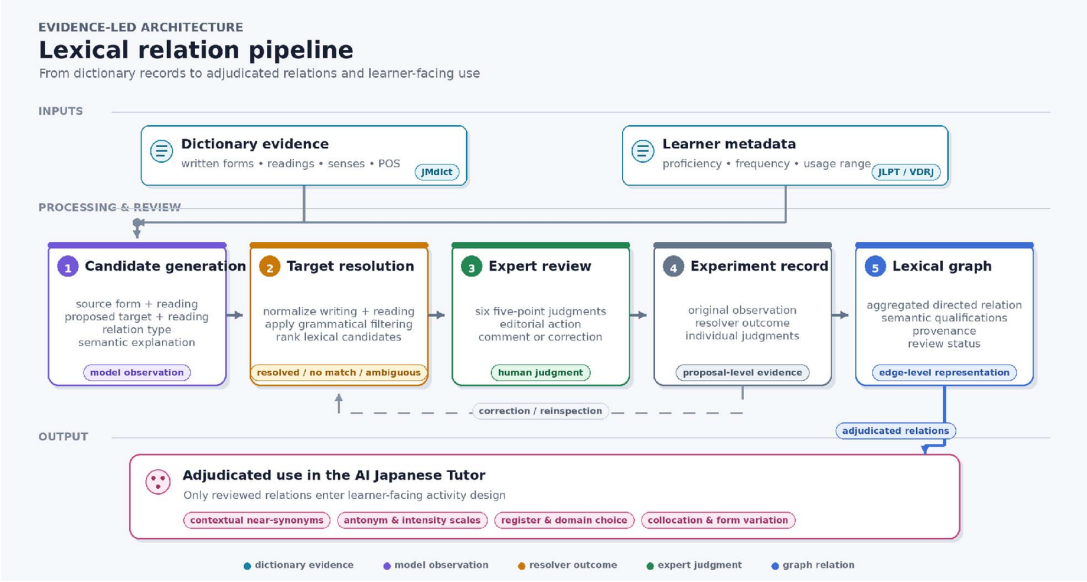


FIGURE 1 Lexical relation pipeline architecture. Dictionary and learner-oriented sources feed candidate generation and target resolution; complete observations and expert judgments remain in the experiment records; adjudicated relations are represented in the lexical graph and support tutor activities

3.1 Relation inventory

The pilot prompt covered relations useful for adjective description, including synonymy, near-synonymy, gradable and complementary antonymy, scalar intensity, register and domain variation, morphological variation, orthographic variation, and collocation. These labels organize candidates for expert assessment.

Synonymy and near-synonymy concern semantic overlap (Cruse, 1986); antonymy and scalar intensity describe opposition and degree; register, domain, morphological, and orthographic variants preserve restrictions or form differences; and collocation records conventional co-occurrence. Reviewers could retain or revise each proposed label. Table 4 gives glossed examples.

TABLE 4 Examples of proposed relation categories

Category	Example	Lexicographic interpretation
Gradable antonymy	低い (*hikui* 'low') – 高い (*takai* 'high')	opposite positions on a height scale
Near-synonymy	危ない (*abunai* 'dangerous') – 危険な (*kiken na* 'dangerous; risky')	overlapping meaning with distributional differences
Register variation	危ない (*abunai* 'dangerous') – やばい (*yabai* 'dangerous; risky', colloquial)	related meaning with a register contrast
Scalar intensity	丸い (*marui* 'round') – まんまるい (*manmarui* 'perfectly round')	greater degree of roundness
Synonymy or orthographic variation	丸い (*marui* 'round') – 円い (*marui* 'round; circular')	proposed as synonymy, but the identical reading and alternative kanji motivate orthographic reclassification

The examples reproduce the proposed forms; thus 危険な *kiken na* 'dangerous' is an attributive form whose lexical headword is *kiken*. The final row illustrates overlapping classification: 丸い (*marui* 'round') and 円い (*marui* 'round; circular') are semantically near-synonymous, while their identical reading and alternative kanji also support classification as orthographic variants.

3.2 Generation and provenance

The pilot used two batches of 200 source records containing written form, hiragana reading, grammatical category, learner level where available, and a stable identifier. The models returned a proposed target in Japanese writing and hiragana, a relation category, confidence, and a short semantic explanation.

The focal models, *gpt-4.1-mini* and *gpt-5.4-mini*, ran at temperature 0.0. Model, configuration, date, source item, original proposal, resolver outcome, and editorial status were retained. The *gpt-5.4-mini* archive contains 101 preliminary observations and subsequent runs of 1,389 and 1,400 observations.

3.3 Target resolution and graph representation

The resolver normalizes form and reading variants, retrieves candidates by orthographic or reading match, applies grammatical filtering, and ranks the remainder. “Resolved” records a top candidate; “no match” and “ambiguous” record the other outcomes. The original proposal remains available for expert assessment of reading, sense, phrase boundary, and relation type.

Complete observations remain in the experiment records. The graph summarizes an ordered source–target–type combination as one directed edge with aggregated provenance, while reverse proposals and successive semantic accounts remain available for editorial comparison. Model agreement can prioritize review; editorial judgment assigns reviewed status.

Figure 2 shows the resulting neighbourhood view for 危ない *abunai* ‘dangerous’; detailed provenance and review status remain available in the relation records.

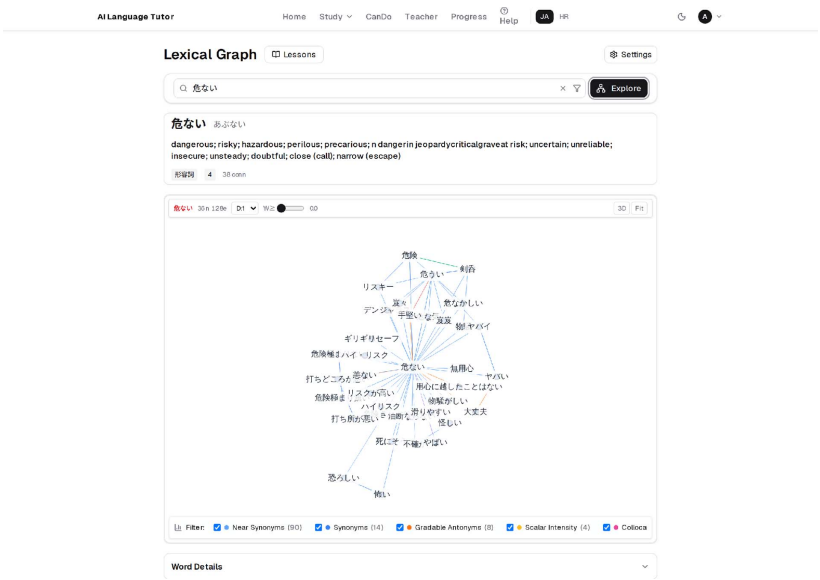


FIGURE 2 Lexical neighbourhood displayed for 危ない *abunai* ‘dangerous’ in the development application. The centre is the selected item, surrounding nodes are relation candidates, and edge colours distinguish displayed relation categories

Below we also show the relationship details for the graph edge formed by nodes 危ない *abunai* ‘dangerous’ and 危うい *ayaui* ‘dangerous’, with relationship type NEAR_SYNONYM. Their synonymy weight is 0.9, and shared meaning ‘dangerous, precarious, on the verge of trouble’. Figure 3 below shows explanations on distinction, usage contexts, formality, when to prefer one over the other.



FIGURE 3 Relationship details for 危ない *abunai* 'dangerous' and 危うい *ayaui* 'dangerous'

4 Pilot Design

4.1 Pilot material and generation

The two batches each contained 200 distinct written-form–reading pairs selected across adjective strata and available JLPT levels, with usage rank used for ordering. Their 95-item overlap yields 400 source records representing 305 distinct lexical items.

The counts of stored generation observations reconstructed from the experiment records were:

TABLE 5 Stored proposal observations

Model	Stored observations	Source records covered	Distinct written-form–reading items
gpt-4.1-mini	1,960	400	305
gpt-5.4-mini	2,890	400	305
Total	4,926	—	—

4.2 Expert review and analysis

Three linguistic experts familiar with Japanese lexical description provided the judgments. Two accounts belonging to one evaluator were combined, and four administrative test judgments were excluded. Review followed operational queues rather than a blinded comparative design.

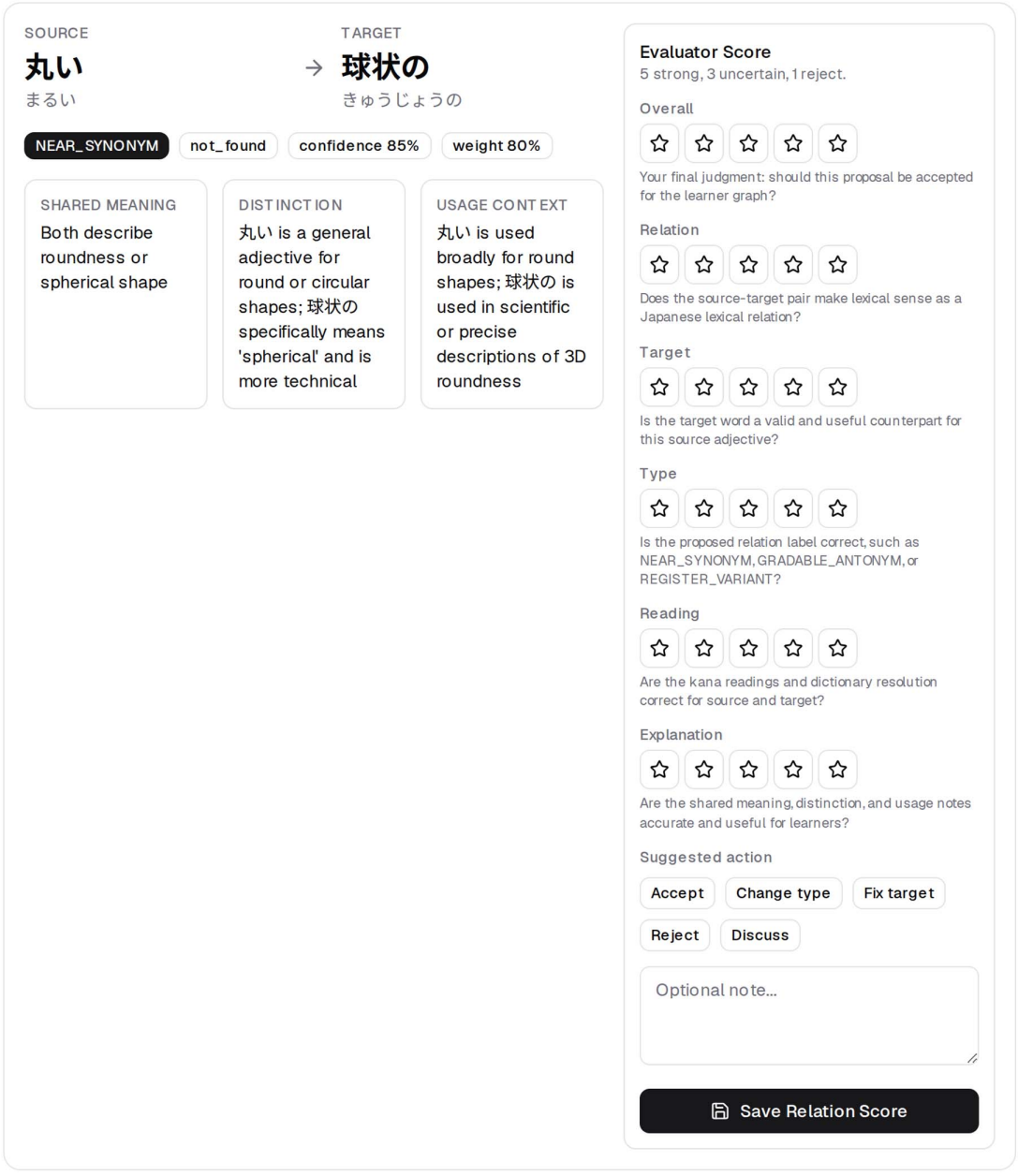


FIGURE 4 Lexical relation review interface used in the pilot. The proposal panel presents source and target writing and reading, relation type, resolution status, model confidence and weight, and the generated semantic account. The scoring panel records six five-point judgments, an editorial action, and an optional comment

For each proposal, the review form displayed the source and target forms and readings, proposed relation category, target-resolution status, confidence, weight, shared meaning, semantic distinction, and usage context. Reviewers assigned scores from one to five for the overall

proposal, lexical relation, target, relation type, reading, and explanation. They then chose one editorial action—accept, change the relation type, correct the target, reject, or discuss further—and could add a free-text comment. Figure 4 shows the complete scoring card for one proposal, source 丸い *marui* 'round' and target 球状の *kyūjō no* 'spheric' (see footnote 2). The displayed *not_found* status records the resolver outcome described in Section 5.1; the reviewer can still assess whether the proposed expression is linguistically valid and whether the relation is useful.

The evidence comprises 420 judgments on 214 proposals (4.4% of 4,850 focal observations). The convenience sample and assignment differed by evaluator and model. Sixty-eight proposals received one judgment, 86 received two, and 60 received three.

A judgment is *qualifying* when the evaluator chose "accept" and assigned an overall score of four or five. At proposal level, at least one qualifying judgment indicates a positive assessment, while final category adjudication remains separate. The analysis distinguishes source records, lexical items, stored proposal observations, proposals, and individual judgments.

5 Results

5.1 Proposal categories and target resolution

Near-synonymy dominated the stored focal observations, accounting for 1,214 of 1,960 *gpt-4.1-mini* observations and 1,767 of 2,890 *gpt-5.4-mini* observations. The concentration reflects the prompt inventory and the broad semantic space covered by "similar with usage differences".

TABLE 6 Stored observations by proposed relation label

Relation label	gpt-4.1-mini	gpt-5.4-mini
Near-synonym	1,214	1,767
Gradable antonym	262	383
Morphological variant	161	208
Synonym	116	146
Collocation	27	175
Register variant	85	82
Other labels	95	129
Total	1,960	2,890

The largest numerical difference concerns collocations: 175 *gpt-5.4-mini* observations and 27 *gpt-4.1-mini* observations. Table 10 shows why this category requires separate lexicographic assessment.

Target-resolution outcomes provide a second view of the generated records. Table 7 reports three processing categories assigned by the normalized, POS-constrained resolver. *Resolved to top candidate* means that the resolver selected one lexical record as the best available match; the relation and sense still proceed to expert review. *No match* means that the resolver could not associate the proposed target with a suitable lexical record. This may indicate an incorrect or misspelled proposal, a valid item absent from the imported inventory, a multiword expression represented outside the lexical-item inventory, or a difference in spelling, reading, or grammatical classification. *Ambiguous* means that several plausible records remained and the

resolver could not select one confidently. These categories describe the outcome of lexical lookup, not the linguistic validity of the proposed expression.

TABLE 7 Outcomes of the implemented target resolver

Model	Resolved to top candidate	No match	Ambiguous	Total
gpt-4.1-mini	648	1,235	77	1,960
gpt-5.4-mini	940	1,795	155	2,890
Total	1,631	3,063	232	4,926

The resolver assigned 1,631 targets to a top candidate. The outcomes describe target processing and identify records requiring editorial inspection.

5.2 Expert-review results

Of the 420 individual judgments, 314 (74.8%) met the qualifying rule. Considered by proposal, 180 of the 214 reviewed proposals (84.1%) received at least one qualifying judgment. The first percentage describes reviewer decisions; the second identifies proposals that received at least one positive assessment within their assigned one to three reviews.

TABLE 8 Expert judgments under the operational acceptance rule

Model recorded on proposal	Judgments	Distinct reviewed proposals	Qualifying judgments	Proposals with ≥1 qualifying judgment
gpt-4.1-mini	232	111	184	95
gpt-5.4-mini	188	103	130	85
Total	420	214	314	180

Table 9 complements the qualifying rule with the distribution of editorial actions. The 99 actions other than acceptance (321) comprise 30 target corrections, 35 requests for discussion, 26 rejections, and eight type changes. They expose errors in target form, reading, phrase boundary, and category.

TABLE 9 Expert action distribution at judgment level

Suggested action	gpt-4.1-mini	gpt-5.4-mini	Total
Accept	189	132	321
Fix target	11	19	30
Needs discussion	18	17	35
Reject	9	17	26
Change type	5	3	8
Total	232	188	420

5.3 Reviewed examples and interpretation

Proposals with three qualifying judgments included 太い (*futoi* 'thick')–細い (*hosoi* 'thin'), 古い (*furui* 'old')–新しい (*atarashii* 'new'), 低い (*hikui* 'low')–高い (*takai* 'high'), and 丸い (*marui* 'round')–まんるい (*manmarui* 'perfectly round'). Table 10 illustrates cases requiring correction or discussion.

TABLE 10 Examples requiring correction or discussion

Model	Source and target	Proposed label	Recorded actions	Lexicographic issue
gpt-5.4-mini	低い (*hikui* 'low') – 低い位置 (*hikui ichi* 'low position')	domain variant	discussion; reject	phrase containing the source, not a lexical variant
gpt-5.4-mini	冷たい (*tsumetai* 'cold') – 冷たい目 (*tsumetai me* 'cold look')	collocation	discussion; reject	potentially valid collocation, but not an adjective-to-adjective relation
gpt-4.1-mini	小たさい (*chiisai* 'small') – ちいさい (*chiisai*)	synonym	reject	script variation presented as synonymy
gpt-5.4-mini	多い (*ōi* 'many') – 多い (*ōi* 'many')	collocation	reject	exact self-target

Together, the examples show how structured review separates usable contrasts from script variation, self-relations, phrases, and category errors. Candidate generation broadens the material available for inspection; editorial review turns it into lexicographic evidence.

6 Discussion

Candidate generation presented possible contrasts, register alternatives, and scalar expressions beyond explicit JMdict relations. The reviewed subset combined established-looking contrasts with self-targets, script variants, phrases, and remote associations. This confirms the value of separating candidate discovery from editorial status: model confidence can prioritize queues, while contextual evidence and expert judgment determine the description (Atkins & Rundell, 2008).

Writing and reading together distinguish relations involving homographic forms and let the interface present transliteration, grammar, and gloss with each proposal. The pair operates as a lookup key; entry- and sense-level provenance addresses polysemy and duplicate records. Future POS-variant links can connect nominal and adjectival analyses where dictionary and corpus evidence support that decision.

The coverage-oriented, overlapping batches and operational review queues support descriptive analysis of workflow feasibility. They do not provide a balanced model comparison, population estimate, or inter-rater agreement study. The study used JMdict release JMdict.2026-03-15. Resolver outcomes should not be interpreted as direct measures of model accuracy because they may also reflect inventory coverage, spelling or reading variation, grammatical filtering, and candidate ranking. The incomplete expanded run was excluded from the reported results.

Adjudicated relations can support contextual near-synonym choice, intensity ordering, register selection, collocation tasks, and deliberate comparison of homographs and orthographic variants. Learner-facing use requires editorial status, reading, POS, category, restrictions, a learning objective, context, and feedback design (Council of Europe, 2020; Woolf, 2010). Pedagogical effectiveness remains a separate learner-study question.

7 Conclusion

We have presented a graph-based workflow for Japanese adjective lexical relations that integrates Sino-Japanese script readings, supports homograph disambiguation, resolves POS ambiguity, and enables quality-controlled relation building. The workflow separates dictionary evidence, model proposals, resolver outcomes, and expert judgments while using written form plus hiragana reading to distinguish homographs such as 辛い (*karai* 'spicy') and 辛い (*tsurai* 'painful'). The pilot stored 1,960 proposal observations from *gpt-4.1-mini* and 2,890 from *gpt-5.4-mini* across 400 source records. In the reviewed subset, 314 of 420 judgments met the qualifying rule and 180 of 214 proposals received at least one qualifying judgment; correction and rejection examples show why editorial control remains central. The pilot expert evaluation confirmed that GPT-5.4-mini produces superior candidates when constrained by graph-structural validation.

This graph-based representation enlarges the candidate space while preserving provenance and semantic qualification. A controlled follow-up will extend coverage, balance review, complete category adjudication, and measure effects on lexicographic effort and descriptive quality.

References

Atkins, B. T. S. & Rundell, M. (2008). *The Oxford guide to practical lexicography*. Oxford University Press.

Backhouse, A. E. (2004). Inflected and uninflected adjectives in Japanese. In R. M. W. Dixon & A. Y. Aikhenvald (Eds.), *Adjective classes: A cross-linguistic typology* (pp. 50–73). Oxford University Press.

Breen, J. W. (1995). *Building an electronic Japanese-English dictionary* [Invited paper]. Japanese Studies Association of Australia Conference, Brisbane.

Council of Europe. (2020). *Common European framework of reference for languages: Learning, teaching, assessment—Companion volume*. Council of Europe Publishing.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge University Press.

Fellbaum, C. (Ed.). (1998). *WordNet: An electronic lexical database*. MIT Press.

Isahara, H., Bond, F., Uchimoto, K., Utiyama, M. & Kanzaki, K. (2008). Development of the Japanese WordNet. In N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis & D. Tapias (Eds.), *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*. European Language Resources Association.

Matsushita, T. (2011). *The Vocabulary Database for Learners of Japanese* (Version 1.11). Retrieved from http://www17408ui.sakura.ne.jp/tatsum/database.html#vdrj

Nakamura, S. (2023) A Study of the Characteristics of No-adjectives in Japanese. *Tama University Research Bulletin "Management Information Research"*, 27 2023.

Navigli, R. & Ponzetto, S. P. (2012). BabelNet: The automatic construction, evaluation and application of a wide-coverage multilingual semantic network. *Artificial Intelligence*, 193, 217–250. https://doi.org/10.1016/j.artint.2012.07.001

Nishiyama, K. (1999). Adjectives and the copulas in Japanese. *Journal of East Asian Linguistics*, 8(3), 183–222.

Perak, B. & Ban Kirigin, T. (2023). Construction Grammar Conceptual Network: Coordination-based graph method for semantic association analysis. *Natural Language Engineering*, 29(3), 584–614. https://doi.org/10.1017/S1351324922000274

Srdanović, I. (2019). Adjectives on *-i* in Japanese language corpora: Distributions, patterns and lexical constraints. In I. Srdanović & A. Bekeš (Eds.), *The Japanese language from an empirical perspective*. University of Ljubljana Press.

Srdanović, I. & Špica, D. (2021). Participation of Japanese language students in lexicographic activities and planning of a bilingual e-dictionary. *Tabula*, (18), 175-200.

Sunakawa, Y., Lee, J. & Takahara, M. (2012). The construction of a database to support Japanese learners' dictionaries. *Acta Linguistica Asiatica*, 2(2), 97–115. https://doi.org/10.4312/ala.2.2.97-115

Špica, D. & Perak, B. (2025). Automating lexical graph construction with large language models: A scalable approach to Japanese multi-relation lexical networks. *Knowledge*, 5(4), 24.

Takahashi, T. (2003). *Dōshi kyūshō* [Verbs in nine chapters]. Hituzi Syobo.

Wiktionary contributors. (2025, November 1). Appendix: JLPT. In Wiktionary. Retrieved July 24, 2026, from https://en.wiktionary.org/wiki/Appendix:JLPT

Woolf, B. P. (2010). *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning*. Morgan Kaufmann.

Contact information

Benedikt Perak
University of Rijeka, Faculty of Humanities and Social Sciences, Croatia
bperak@uniri.hr

Dragana Špica
Juraj Dobrila University of Pula, Croatia
dspica@unipu.hr

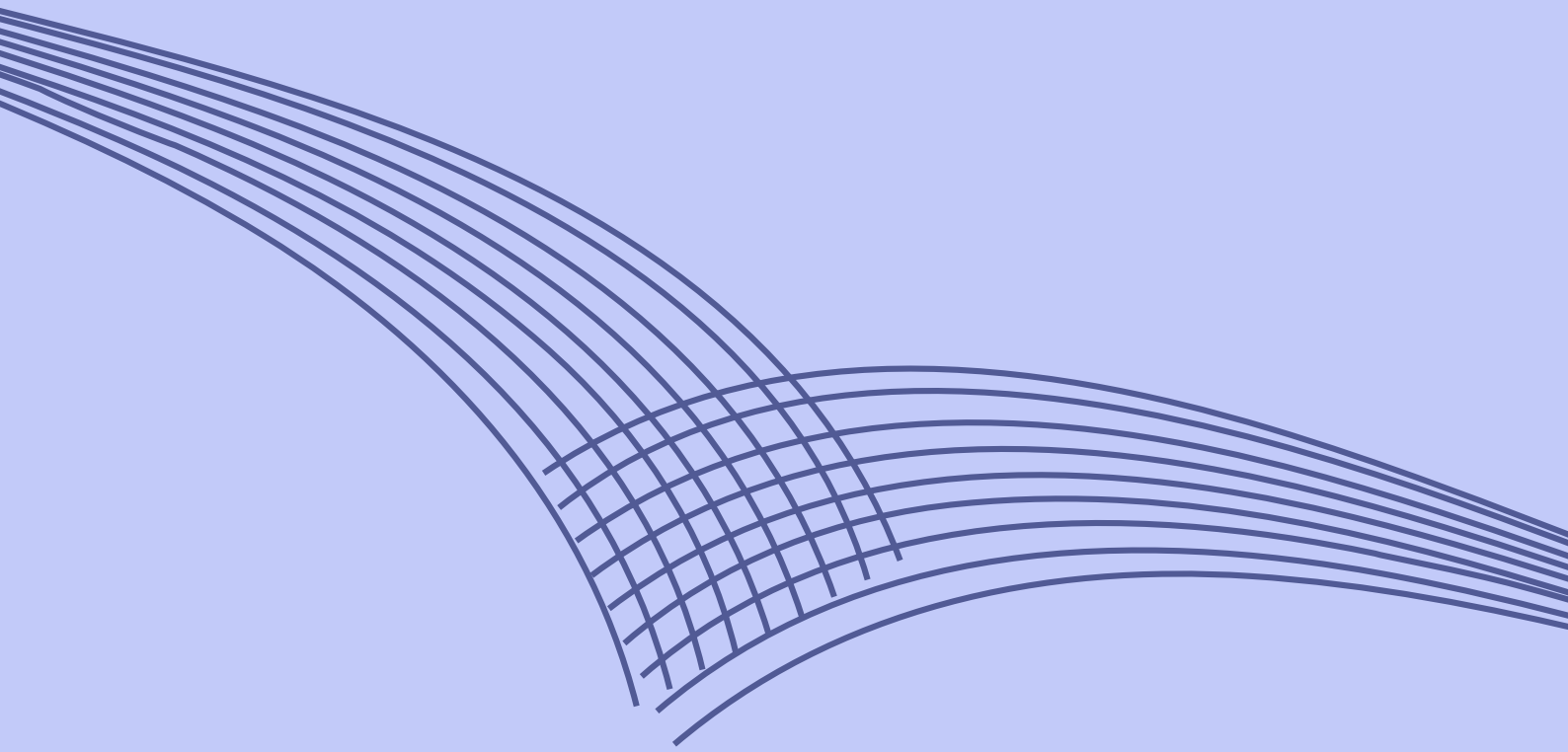
Irena Srdanović
Juraj Dobrila University of Pula, Croatia
irena.srdanovic@unipu.hr

Acknowledgements

This work was funded by the STUDIA project (UNIRI-IZ-25-248), co-financed by the National Recovery and Resilience Plan (NPOO) and the University of Rijeka (PI: Benedikt Perak); the NextGenerationEU programme under the Juraj Dobrila University of Pula institutional research project "From Dynamic Vocabulary to Virtual Interaction: Development of a Foreign Language Learning Model in Tourism Using Corpora, Artificial Intelligence, and Extended Reality (VIRAI – Language for Real and Virtual Worlds)" (IIP_UNIPU_010138; PI: Irena Srdanović); the FORMALS2 project, supported by the Croatian Science Foundation under grant HRZZ-IP-2024-05-3882 (PI: Tin Perkov); and the project "Epistemic Democracy in the Digital Age", supported by the Croatian Science Foundation under grant HRZZ-IP-2024-05-4113 (PI: Ivan Cerovac).

MULTILINGUALISM, LANGUAGE VARIETIES & CONTACT LEXICOGRAPHY

EURALEX 2026



Luke Omoyemi Akinremi & Elena Martín-Cancela

BILINGUAL LEXICOGRAPHY UNDER RESOURCE ASYMMETRY

Designing Yoruba–German Dictionary Articles in Dialogue with Galician–German

Abstract This paper addresses the challenge of designing bilingual dictionary articles under resource asymmetry by focusing on the Yoruba–German language pair and drawing methodological insights from the more developed Galician–German ecosystem. It proposes a resource-aware model of dictionary article composition that explicitly accounts for limited evidence by making uncertainty visible, using heterogeneous sources (corpora, pivot resources, expert input). Our approach involves dictionary-critical analyses of both Yoruba–German and Galician–German lexicographic resources and the transfer of microstructural design strategies from Galician to Yoruba. Case studies include Yoruba–German prototype articles (e.g. the verb *ṣe* and the nouns *ayé* and *akọ́nimọ̀ọ́gbá*) developed from these findings, incorporating sense divisions, translation equivalents, examples, grammatical/constructional notes and cross-references. The study also demonstrates a human-led use of AI tools to enrich example sentences and collocations in the prototypes. Main contributions are the concept of a resource-aware dictionary article and a cross-pair methodological dialogue, operationalised in prototype articles. These contributions provide practical guidelines for lexicographers of under-resourced languages and suggest scalable, AI-assisted workflows for bilingual lexicography under unequal resource conditions.

Keywords Bilingual lexicography; Yoruba; Galician; resource asymmetry; dictionary article design; AI-assisted lexicography; cross-pair methodological dialogue

1 Introduction

1.1 Resource asymmetry in bilingual lexicography

Bilingual lexicography is rarely carried out under equal conditions. The resources available for different languages vary considerably, ranging from extensive corpora, digital infrastructures, and long-standing lexicographic traditions to environments where such resources remain scarce or fragmented. These differences are not merely technical. They directly affect the way dictionary articles are designed, influencing decisions about sense distinction, equivalence treatment, example selection, and the representation of grammatical information. In this respect, the challenges faced by lexicographers are often determined as much by the availability of evidence as by the linguistic characteristics of the languages themselves. Bilingual dictionaries cannot therefore be understood as neutral repositories of lexical correspondences, but rather as structured lexicographic texts governed by specific organisational principles and theoretical assumptions (Wiegand, 1983). When these principles are applied under uneven conditions of resource availability, lexicographic design becomes a site of negotiation between linguistic description, user needs, and practical constraints.

These challenges become particularly evident in language pairs involving under-resourced languages. In such contexts, lexicographers often must work with incomplete or uneven evidence, relying on small corpora, mediated resources, or expert knowledge to compensate for gaps in the available data (Magueresse et al., 2020). This affects core lexicographic decisions, from the identification and separation of senses to the selection of translation equivalents and usage examples. As Wiegand (2002) points out, equivalence in bilingual lexicography is rarely

a simple one-to-one relationship between lexical units; rather, it depends on linguistic, cultural, and contextual factors. When reliable evidence is limited, establishing and justifying such equivalences becomes considerably more difficult.

At the same time, recent developments in digital lexicography have shown that the absence of large-scale resources does not necessarily prevent the creation of useful and linguistically sound dictionary articles. Increasingly, lexicographers combine different types of evidence, including digital corpora, online resources, translation tools, and expert judgement, to support the design of bilingual dictionary articles (De Schryver, 2023). Within this perspective, lexicography is understood not simply as the description of language, but as a user-oriented activity aimed at providing relevant information for specific communicative situations (Tarp, 2008). For under-resourced language pairs, this makes methodological flexibility and transparent design decisions particularly important. To address this methodological challenge, the paper proposes a cross-pair methodological dialogue in which Galician–German serves as an intermediate lexicographic ecosystem between major European language pairs and the more constrained Yoruba–German context.

1.2 Why Yoruba–German?

The Yoruba–German language pair constitutes an extreme but analytically productive case of such asymmetry. While German benefits from extensive lexicographic traditions, standardised resources, and advanced natural language processing tools, Yoruba presents a markedly different profile, with limited and fragmented corpora (Adewole et al., 2017), inconsistencies in orthographic representation—particularly in tonal marking— (Orife et al., 2020), and a comparatively underdeveloped digital lexicographic ecosystem (Magueresse et al., 2020). These conditions directly affect the feasibility of corpus-driven lexicography and require alternative approaches to sense description, example selection, and grammatical representation.

From a lexicographic perspective, the difficulty is not merely that Yoruba has fewer digital resources than German, but that the available resources are often uneven in their coverage and limited in their ability to support fine-grained dictionary article design. Existing Yoruba online dictionaries and glossaries tend to provide lexical correspondences, but frequently offer limited contextualisation, few usage examples, and inconsistent treatment of diacritics. This is particularly relevant because tonal and vowel diacritics are not peripheral features in Yoruba: they are central to lexical identification, disambiguation, and reliable searchability (Orife et al., 2020). As a result, the design of Yoruba–German dictionary articles requires workflows capable of compensating explicitly for incomplete evidence and of documenting uncertainty rather than concealing it.

Moreover, the relative scarcity of Yoruba–German lexicographic resources means that lexicographers frequently have to rely on indirect pathways involving English as a pivot language. While this practice facilitates access to lexical information, it may also introduce additional layers of semantic approximation and increase the risk of equivalence distortions. Consequently, the challenge is not merely one of resource availability, but of ensuring that lexicographic decisions remain transparent and linguistically justified despite limited empirical evidence.

1.3 Why German?

German is selected as the target language not only because of its well-established lexicographic infrastructure, but also because it functions as a stable pivot language that enables contrastive modelling. German is particularly relevant for the present study because its lexicographic tradition has developed well-established solutions for representing meaning, grammatical behaviour, and usage patterns within dictionary articles. These solutions provide useful methodological insights that can be examined and adapted in more resource-constrained lexicographic environments (Wiegand, 2002).

At the same time, German is not treated here as a neutral target language. Its own lexicographic tradition, especially the attention paid to article structure, grammatical information, compounds, valency, and phraseological patterns, makes it a demanding but productive target language for modelling bilingual dictionary articles. In other words, German does not merely receive equivalents from Yoruba or Galician; it also imposes structural requirements on how equivalence is represented. This is especially visible when a source-language expression can correspond in German to different formal solutions, such as a genitive construction, a prepositional phrase, an adjective, a compound, or an infinitival construction.

1.4 Why Galician–German?

The relevance of Galician–German for the present study lies in its intermediate position between highly resourced European language pairs and markedly under-resourced combinations such as Yoruba–German. Although it cannot be considered a fully developed bilingual lexicographic pair comparable to major European language combinations, neither does it face the same degree of resource scarcity. Rather than being supported by a single consolidated bilingual dictionary tradition, the Galician–German ecosystem consists of a heterogeneous network of monolingual reference works, multilingual tools, terminological resources, contrastive platforms, and mediated resources accessed through Spanish or German.

More importantly, the value of Galician–German as a methodological reference point stems from the fact that Galician participates in a growing ecosystem of linguistic infrastructures, including semantic resources, terminological databases, multilingual annotation frameworks, and language technology initiatives. Recent developments in contrastive lexicography, semantic annotation, AI-assisted lexicographic modelling, and sustainable approaches to dictionary making (Valcárcel Riveiro, 2023; Arias-Arias et al., 2024; Calañas Continente & Domínguez Vázquez, 2023) provide valuable examples of how a minoritised language can strengthen its lexicographic ecosystem and generate methodologies that may be adapted to other resource-constrained contexts.

While these infrastructures do not constitute a ready-made solution for Galician–German lexicography, they illustrate how limited-resource environments can develop transferable methodologies and scalable lexicographic workflows. For this reason, Galician–German is particularly useful for the present study, not as a model to be replicated directly, but as a source of design strategies that can be examined, adapted, and transferred to the more constrained Yoruba–German context.

1.5 Aim and contribution

The present study treats Galician–German not as a linguistic analogue of Yoruba–German, but as a methodological reference point for identifying lexicographic design principles that can be transferred to more constrained resource environments. To achieve this, the paper establishes a methodological dialogue between the Galician–German and Yoruba–German pairs, focusing on the transfer of lexicographic strategies rather than linguistic data.

Rather than describing existing dictionaries or comparing the two language pairs directly, the paper adopts a qualitative, design-oriented approach aimed at exploring how strategies developed within the Galician–German lexicographic ecosystem can inform the construction of Yoruba–German dictionary articles. To this end, selected Galician–German resources are analysed in order to identify concrete microstructural solutions—such as sense structuring, example selection, and the representation of grammatical and constructional information—which are subsequently adapted to the specific challenges posed by Yoruba–German lexicography. The outcome is a set of prototype dictionary articles that explicitly document both resource limitations and the design decisions adopted to address them.

Beyond the specific case examined here, the study contributes to broader discussions on dictionary modelling, AI-assisted lexicographic workflows, and sustainable approaches to bilingual lexicography under conditions of resource asymmetry. By focusing on the transfer of lexicographic strategies rather than linguistic data, it proposes a methodological framework that may be applicable to other under-resourced language pairs facing comparable infrastructural and technological constraints.

To justify this approach, the following section examines resource asymmetry as a lexicographic problem, discusses its implications for dictionary article design, and establishes the theoretical basis for the cross-pair methodological dialogue proposed in this study.

2 Methodological and theoretical basis

The analysis is grounded in dictionary criticism, understood as the systematic description and evaluation of dictionary articles according to explicit lexicographic criteria. Dictionary criticism is relevant here because the study does not merely describe existing resources; it examines how their article structures support, limit or inspire bilingual dictionary design. Swanepoel (2017) defines dictionary criticism as a part of dictionary research concerned with the design features of the dictionaries being evaluated. Wiegand (1993) also emphasises that dictionary criticism includes both positive and negative evaluations, while Nielsen (2009) stresses that dictionary reviews should contribute to the improvement of dictionaries and new editions. In this sense, dictionary criticism provides a suitable methodological basis for using the Galician–German analysis as input for Yoruba–German prototype design.

This study follows a descriptive and evaluative approach. The descriptive component identifies what kinds of information are present in the selected dictionary articles and resources. The evaluative component assesses how far these information types support article design for bilingual lexicography. This follows Ripfel's view, discussed in Swanepoel (2017), that an academic dictionary review should describe design features, evaluate them, justify the evaluation with reference to theory and method, and arrive at an overall assessment. It also follows Gouws (2017), who argues that dictionary criticism should assess whether lexicographic products fulfil their role as authoritative tools for users.

The criteria used in the analysis were adapted from dictionary-critical work on bilingual dictionaries and dictionary evaluation. Hütsch (2017) distinguishes lexicographical aspects such as macrostructure, microstructure, mediostructure, frame structure and typographical presentation, while Svensén (2009) proposes a practical "desktop method" involving familiarisation with the dictionary, examination of different article types, assessment of inner and outer access structures, comparison with similar dictionaries and evaluation of the dictionary's ability to fulfil its intended functions. Since the present study focuses on article design rather than on whole-dictionary evaluation, these broader criteria were narrowed to four article-level dimensions: sense structuring, example treatment, grammatical and constructional information, and culture-specificity/register/non-equivalence.

The study also draws on Gouws and Prinsloo's (2005) account of dictionary structure. Their distinction between the macrostructural element, the microstructural articles and the systematic organisation of article-internal data is useful for evaluating how information is distributed within dictionary articles. For bilingual lexicography, their discussion of integrated and non-integrated microstructures is especially relevant: examples and contextual information are more useful when they are placed close to the sense or translation equivalent they illustrate. This principle directly informs the Yoruba–German prototypes, where examples are attached to the relevant sense, construction or expression.

Because no universal set of evaluation criteria can serve all dictionary reviews equally well, the criteria used here are purpose-specific. Akasu (2013) argues that reviewers should develop criteria that correspond to the objectives of the review and the needs of its readers. In the present study, the objective is to identify article-design strategies that can be transferred from the Galician–German analysis to Yoruba–German prototype articles. The evaluation therefore focuses on the quality and usefulness of article components, rather than on the overall quality of the resources as complete dictionaries.

The analytical dimensions were operationalised as follows:

TABLE 1 Analytical dimensions and coding criteria for the dictionary-critical analysis

Dimension	What was assessed	Scale/coding
Sense structuring	Whether meanings are separated, ordered, labelled and linked to equivalents	0 = absent; 1 = implicit or unclear; 2 = explicit but uneven; 3 = explicit and usable
Example treatment	Presence, relevance and usefulness of examples	Typicality, informativeness and intelligibility rated 0–3; translation adequacy added for bilingual examples
Grammatical/constructional information	POS, gender, inflection, valency, subcategorisation, phraseology and constructions	0 = absent; 1 = basic; 2 = useful; 3 = detailed or contrastive
Culture/register/ non-equivalence	Treatment of cultural, historical, domain-specific, register-sensitive or partially equivalent meanings	0 = absent; 1 = implicit; 2 = labelled or partly explained; 3 = clearly labelled or explained

For example treatment, the scale was adapted from GDEX-inspired criteria (Kilgarriff et al., 2008; Kosem et al., 2019). A score of 0 indicates that no running example is provided. A score of 1 indicates that examples are weak, unclear, poorly linked to the lemma or potentially

misleading. A score of 2 indicates usable but uneven examples. A score of 3 indicates examples that are clear, typical, informative and well connected to the relevant sense or construction. For bilingual examples, translation adequacy was assessed separately: 1 indicates inaccurate or misaligned translation, 2 indicates broadly adequate but awkward or partial translation, and 3 indicates adequate and idiomatic translation.

The ratings are qualitative summaries of the analysed articles. They are used to make the dictionary-critical comparison transparent, not to produce a statistical ranking of the resources. The aim is to identify which kinds of article information each resource contributes and how these contributions can support Yoruba–German article design. The evaluation was conducted jointly by the two authors, each contributing expertise in their respective language pairs. Luke Omoyemi Akinremi contributed expertise in Yoruba and German, while Elena Martín-Cancela contributed expertise in Galician and German bilingual lexicography. The scores were discussed until consensus was reached; therefore, no inter-rater agreement statistics were calculated. Ranges such as 1–2 or 2–3 show that the resources performed differently across the selected lemmas, not that the authors disagreed.

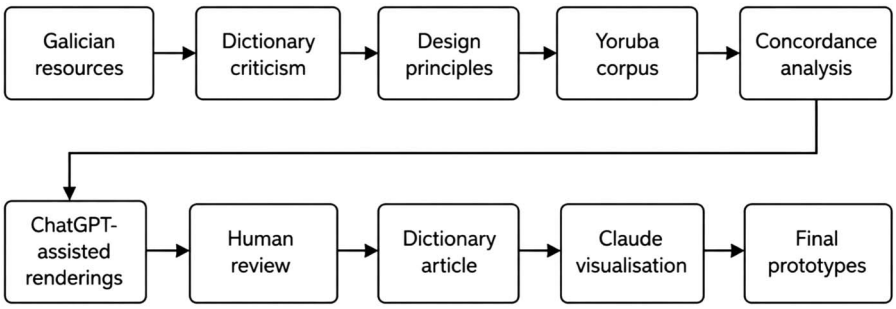


FIGURE 1 Workflow of the Yoruba–German prototype development

Figure 1 summarises the main stages of the study, from the Galician–German dictionary criticism to the final Yoruba–German prototype articles. The workflow distinguishes between human analysis, corpus tools, LLM-assisted drafting and LLM-assisted visualisation. Human work included lemma selection, interpretation of corpus evidence, dictionary-critical evaluation, sense structuring, final equivalent selection and validation of examples. Sketch Engine was used for corpus access and concordance analysis. ChatGPT (GPT-5.5 Thinking) was used after manual corpus analysis to generate candidate German renderings and bridge glosses. Claude AI Sonnet 4.6 was used to convert finalised article content into an HTML-based visual prototype. All AI-assisted outputs were checked and revised by the authors. The prompts were developed iteratively during the lexicographic workflow and adapted to the specific lemma and corpus evidence. They were used mainly to request candidate German renderings and bridge glosses for manually selected Yoruba examples, and the conversion of finalised dictionary article content into an HTML-based prototype without changing the lexicographic content.

3 Galician–German resource analysis: findings for article design

The Galician–German analysis applies the dictionary-critical criteria described above to four resources: Glosbe (n.d.), Galnet (Gómez Guinovart & Solla Portela, 2018), González González's (n.d.) Dicionario da Real Academia Galega (DRAG), and PortLex (Domínguez Vázquez et al., 2018), aiming to identify how these resources organise article-level information and what they

contribute to bilingual article design. The analysis focuses on the selected articles rather than on the complete resources as whole dictionaries.

The study selected six Galician lemmas from the CORGA database's normalised frequency list (Centro Ramón Piñeiro para a Investigación en Humanidades, n.d.) based on distinct linguistic traits, choosing *facer*, *tempo*, *novo*, *peseta*, *aprendizaxe*, and *actividade* to analyse varied grammatical, polysemous, and constructional behaviors. The results show that the resources contribute different types of article information. Glosbe is useful for identifying candidate German equivalents and bilingual examples, but these often require verification. In *facer*, for example, equivalents such as *machen*, *bauen* and *tun* appear without a clearly developed sense hierarchy. In *tempo*, equivalents such as *Zeit*, *Wetter*, *Weile* and *Zeitspanne* point to different semantic areas, but the user receives limited guidance on when each equivalent is appropriate. This confirms the need to link equivalents to source-language senses or constructions.

Table 2 summarises the resource contributions across the four analytical dimensions.

TABLE 2 Resource contribution by analytical dimension

Resource	Sense structuring	Examples	Grammar / construction	Culture / register / non-equivalence	Main contribution
Glosbe	1-2	2	1	1-2	Candidate equivalents and bilingual examples
Galnet	3	0	1	2-3	Semantic benchmarking
DRAG	3	3	2	3	Practical source-language article structure
PortLex	2	2-3	3	1	Nominal valency and contrastive constructions

Scores: 0 = absent; 1 = weak or basic; 2 = usable but uneven; 3 = strong. The scores summarise the analysed sample. PortLex was assessed only for the nominal lemmas *aprendizaxe* and *actividade*, since these were the relevant items available in that resource.

Galnet contributes strongly to semantic benchmarking. Its WordNet-style synsets make semantic distinctions visible, especially where bilingual equivalent lists remain flat or implicit. For *tempo*, Galnet separates meanings relating to time, weather, period, season and specialised domains. For *peseta*, it helps explain why both currency-related and vulgar anatomical meanings appear in the wider resource environment. However, Galnet does not provide running examples or practical article-level guidance for learners. Its main value lies in semantic disambiguation.

DRAG is central to the analysis because bilingual equivalents need to be assigned only after the source-language structure has been understood. As a normative monolingual dictionary, DRAG provides sense division, examples, grammatical labels, register information, domain labels and phraseology. This is particularly important for *tempo*, *novo* and *peseta*, where source-language distinctions must be clarified before German equivalents can be selected. For *peseta*, DRAG distinguishes the historical currency unit, the coin or banknote, and the popular anatomical sense. This kind of source-language structuring is essential for avoiding misleading equivalent ordering in a bilingual article.

PortLex contributes a different kind of evidence. Its strength lies in nominal valency and contrastive constructional information. In *aprendizaxe* and *actividade*, PortLex shows how Galician nominal structures correspond to German patterns involving genitives, prepositional complements, compounds or other constructions. This supports the argument that bilingual dictionary articles should provide constructional guidance where equivalents alone are insufficient.

Example treatment was also assessed separately because examples play a central role in article usability. In bilingual dictionaries, examples help users connect senses, equivalents and usage contexts. Gouws and Prinsloo's (2005) account of integrated microstructure is relevant here: examples are most useful when they are placed close to the sense or equivalent they illustrate. This principle also informed the Yoruba–German prototype articles.

TABLE 3 Example treatment across resources

Resource	Typicality	Informativeness	Intelligibility	Translation adequacy	Interpretation
Glosbe	2	2	2	1-2	Useful bilingual examples, but uneven and requiring verification
Galnet	0	0	0	n/a	No running examples
DRAG	3	3	3	n/a	Strong monolingual, sense-specific examples
PortLex	2-3	3	2-3	3	Strong constructional examples in specialised articles

Scores: 0 = absent; 1 = weak; 2 = usable but uneven; 3 = strong. Translation adequacy applies only to bilingual or multilingual examples.

DRAG provides the strongest model for compact, sense-specific examples. Its examples are generally clear, short and directly linked to the relevant meaning. Glosbe provides bilingual examples, but their quality varies. Some examples are useful, while others are awkward, weakly aligned or insufficiently connected to the intended sense. Galnet provides no running examples. PortLex provides specialised constructional examples which are particularly useful for identifying how nouns behave in multilingual valency patterns.

The analysis of culture-specificity, register and non-equivalence confirms the need for labels, notes and careful ordering. *Peseta* is the clearest case. In Glosbe, vulgar German equivalents appear prominently, while the historical currency sense is less clearly foregrounded. Galnet helps identify the existence of different semantic areas, and DRAG provides the practical distinction between the historical currency sense and the popular anatomical sense. This confirms that culture-specific and register-sensitive meanings require explicit source-language checking before bilingual equivalents are presented to users.

Other cases show that non-equivalence is not limited to cultural words. *Aprendizaxe* requires a distinction between learning as a process and apprenticeship or learning period. *Actividade*

requires attention to domain-sensitive German alternatives such as *Aktivität*, *Tätigkeit*, *Handlung* and *Aufgabe*. In these cases, equivalence depends on sense, domain and construction.

The Galician–German analysis therefore leads to a layered model of article design. The model consists of seven steps: source-language sense structuring, semantic benchmarking, equivalent assignment, example selection, grammatical and constructional guidance, cultural or register annotation, and human verification. These steps translate the dictionary-critical findings into practical design principles for the Yoruba–German prototypes.

The following principles were derived from the analysis:

1. establish source-language senses before assigning German equivalents;
2. link equivalents to specific senses, constructions or expressions;
3. place examples close to the sense or construction they illustrate;
4. include grammatical and constructional information where equivalents alone are insufficient;
5. use labels or notes for cultural, historical, register-sensitive and domain-specific meanings;
6. use cross-references for related lexical items that should not be treated as senses of the headword;
7. verify automatically aggregated or AI-assisted material through human lexicographic judgement.

These principles are applied in the next section to the Yoruba–German prototype articles for *ṣe*, *ayé* and *akọ́nimọ̀ọ́gbá*.

4 Yoruba–German prototype articles

The findings from the Galician–German analysis were applied to three Yoruba–German prototype articles: *ṣe*, *ayé* and *akọ́nimọ̀ọ́gbá*. These lemmas were selected because they represent different article-design challenges. The aim was to test how the layered model described above can be used to organise senses, equivalents, examples, constructions, cultural information and cross-references in an online Yoruba–German dictionary article.

The prototypes were developed from corpus evidence extracted from the Menyo-20k corpus (Adelani et al., 2021) in Sketch Engine. Concordance lines were examined manually in order to identify recurrent meanings, constructions and phraseological patterns. Examples were selected for clarity and lexicographic usefulness. Where corpus lines were too long or context-dependent, they were shortened or lightly adapted while preserving the relevant lexical construction. The examples are therefore corpus-informed and human-edited for dictionary presentation.

Two AI tools were used during the prototype-development stage: ChatGPT, GPT-5.5 Thinking, and Claude AI Sonnet 4.6. These tools were used as auxiliary drafting and visualisation aids. After the Yoruba corpus lines had been manually analysed and grouped, ChatGPT was used to generate possible German renderings of selected examples and bridge glosses. These candidate renderings were then evaluated manually according to semantic fit, constructional compatibility and idiomaticity in German. Claude AI Sonnet 4.6 was used to assist in converting the finalised article content into an HTML-based visual prototype. In both cases, the AI output was

subject to human control. The final German equivalents and example translations therefore remain the result of human lexicographic judgement, supported by AI-assisted drafting and visualisation. This controlled, human–in–the–loop use of LLMs is preferred as studies have shown that ChatGPT–style models can support lexicographic tasks but routinely require expert validation due to errors, overgeneralisations and hallucinated examples (Lew, 2023; Balenović & Proroković, 2025).

The three articles operationalise the model in different ways:

TABLE 4 Lexicographic problems and article-design solutions in the Yoruba–German prototypes

Lemma	Lexicographic problem	Article-design solution
ṣe	Broad polysemous verb with recurrent light-verb constructions	Core equivalent plus constructional subarticles
ayé	Semantic breadth and cultural non-equivalence	Three main senses, one expression, cultural note and cross-references
akọ́nimọ̀ọ́gbá	Domain-specific, morphologically complex noun	Core equivalent, sport label, word-formation note and selected constructions

The article for *ṣe* shows how construction-sensitive equivalent assignment can be represented in a bilingual dictionary. The verb has a broad meaning corresponding roughly to German *tun*, *machen*, *ausführen* or *durchführen*, but its German equivalent often depends on the noun or construction with which it occurs. The article therefore gives a core equivalent first and then lists selected constructional patterns, such as *ṣe àmúlò* "*anwenden, umsetzen*" and *ṣe àfihàn* "*zeigen, darstellen, zum Ausdruck bringen*". The negative identificational construction *kì í ṣe* is treated separately as a special note, since it corresponds to German expressions such as *ist nicht* or *es handelt sich nicht um*, rather than to an action-verb use of *ṣe*.

The article for *ayé* demonstrates the distinction between headword senses, expressions and related lemmas. The final article distinguishes three main sense areas: *Leben*, *Welt* or *menschliche Welt*, and culturally marked *irdisches Leben* or *diesseitige Welt*. The expression *mú ayé dẹrùn* is included as a phraseological expression, translated as *das Leben erleichtern*. Related forms such as *àgbáyé* and *ilẹ̀-ayé* are placed under Siehe auch, since they are better treated as separate lexical items.

The article for *akọ́nimọ̀ọ́gbá* represents a different problem: the treatment of a morphologically complex term in a specialised domain. The article gives the core German equivalents *Trainer*, *Trainerin* and *Coach*, and marks the domain as sport. It also includes a word-formation section explaining the structure *a-* + *kọ́-ni* + *mọ̀ọ́-gbá*. Selected constructions, such as *akọ́nimọ̀ọ́gbá* "*Cheftrainer, Cheftrainerin*" and *yan ẹnìkan gégẹ́ bí akọ́nimọ̀ọ́gbá* "*jdn. zum Trainer ernennen*", show how the lemma functions in recurrent sports-related contexts.

Together, the three articles show how the layered Galician–German findings can be adapted to Yoruba–German lexicography. The resulting model separates senses from constructions and expressions, places equivalents in context, includes explanatory notes where direct equivalence is insufficient, and uses cross-references to connect related lexical items without overloading the main article. The HTML implementation further demonstrates how these information types can be visually organised in an online dictionary environment.

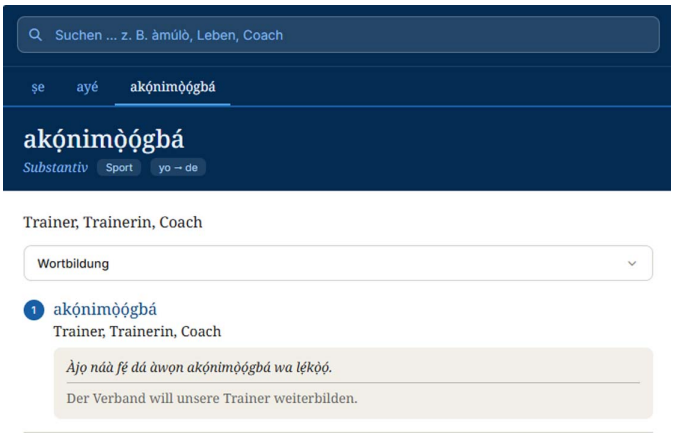


FIGURE 2 HTML prototype[1] interface showing the *akọ́nimọ̀ọ́gbá* article and the general online-dictionary layout

5 Discussion and conclusion

This study has examined how Yoruba–German dictionary articles can be designed under conditions of resource asymmetry by placing the under-resourced pair in dialogue with Galician–German. The analysis shows that even a comparatively better-resourced pair such as Galician–German depends on a mixed set of resources. Glosbe, Galnet, DRAG and PortLex each provide different types of lexicographic evidence, and their value lies in their combined use. The Galician–German analysis produced a layered model for article design. This layered procedure is especially useful for Yoruba–German because there is no comprehensive existing resource from which complete article structures can be directly derived. The Yoruba–German prototypes therefore operationalise the Galician–German findings by turning them into practical decisions about sense structure, equivalents, examples, constructions, notes and cross-references.

The study also demonstrates a controlled role for AI in lexicographic work. ChatGPT and Claude AI Sonnet 4.6 supported drafting, German rendering, and visualisation, while the central lexicographic decisions remained under human control. Lemma selection, interpretation of Yoruba corpus evidence, treatment of senses and constructions, final equivalent selection and article structure were determined through manual analysis. This workflow positions AI as a practical support tool within a human-led lexicographic process.

The scope of the study is deliberately limited. The Yoruba–German output consists of three prototype articles, and the prototype has not yet been tested with users. The German equivalents and example translations would also benefit from further review by German-speaking lexicographers and Yoruba specialists. In addition, the corpus base can be expanded in future work, especially through comparison with larger Yoruba corpora such as Yankari (Akpobi, 2024), while maintaining careful attention to orthography and diacritics. Future stages should therefore include a larger lemma sample, systematic user testing, and further refinement of the HTML prototype.

The main contribution of the study is methodological. It proposes a dictionary-critical and resource-aware model for bilingual dictionary article design for language pairs with limited direct resources, showing how useful design strategies can be developed through comparison with a better-documented language pair.

[1] The prototype is archived at https://doi.org/10.5281/zenodo.21497949.

References

Adelani, D. I., Ruiter, D., Alabi, J. O., Adebonojo, D., Ayeni, A., Adeyemi, M., Awokoya, A. & España-Bonet, C. (2021). The effect of domain and diacritics in Yorùbá–English neural machine translation. In K. Duh & F. Guzmán (Eds.), *Proceedings of Machine Translation Summit XVIII: Research Track* (pp. 61–75). European Association for Machine Translation. https://aclanthology.org/2021.mtsummit-research.6/

Adewole, L. O., Ojokoh, B. A. & Adebiyi, A. A. (2017). A review of Yoruba natural language processing. *International Journal of Advanced Computer Science and Applications*, 8(8), 304–311.

Akasu, K. (2013). Methods in dictionary criticism. In H. Jackson (Ed.), *The Bloomsbury Companion to Lexicography* (pp. 48–61). Bloomsbury.

Akpobi, M. (2025). Yankari: Monolingual Yoruba Dataset. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)* (pp. 1–6). Association for Computational Linguistics.

Arias-Arias, I., Domínguez Vázquez, M. J. & Valcárcel Riveiro, C. (2024). Der Effizienz- und Intelligenzbegriff in der Lexikographie und künstlichen Intelligenz: Kann ChatGPT die lexikographische Textsorte nachbilden? *Lexikos*, 34(1), 51–76. https://doi.org/10.5788/34-1-1879

Balenović, K. & Proroković, J. (2025). The lexicographic potential of artificial intelligence. *Studia lexicographica*. https://doi.org/10.33604/sl.19.36.3

Calañas Continente, J. A. & Domínguez Vázquez, M. J. (2023). About sustainable and socially responsible lexicography: State of the art. *Quaderns de Filologia: Estudis Lingüístics*, 28, 9–19. https://doi.org/10.7203/qf.0.27659

Centro Ramón Piñeiro para a Investigación en Humanidades. (n.d.). *Corpus de Referencia do Galego Actual (CORGA)* (Version 4.1) [Data set]. Retrieved May 20, 2026, from http://corpus.cirp.gal/corga/

De Schryver, G.-M. (2023). Generative AI and lexicography: The current state of the art using ChatGPT. *International Journal of Lexicography*, 36(4), 355–387. https://doi.org/10.1093/ijl/ecad021

Domínguez Vázquez, M. J. (Dir.), Valcárcel Riveiro, C., Mirazo Balsa, M., Sanmarco Bande, M. T., Simões, A. & Vale, M. J. (2018). *Portlex: Diccionario multilingüe de la valencia del nombre*. Universidade de Santiago de Compostela. Retrieved May 20, 2026, from http://portlex.usc.gal/portlex/

Glosbe: Galician-German Dictionary. (n.d.). Retrieved May 20, 2026, from https://glosbe.com/gl/de/

Gómez Guinovart, X. & Solla Portela, M. A. (2018). Building the Galician wordnet: methods and applications. *Language Resources and Evaluation*, 52(1), 317–339. https://doi.org/10.1007/s10579-017-9408-5

González González, M. (Dir.). (n.d.). *Dicionario da Real Academia Galega*. Real Academia Galega. Retrieved May 20, 2026, from https://academia.gal/dicionario

Gouws, R. H. (2017). Dictionary criticism and dictionary critics. In M. Bielińska & S. J. Schierholz (Eds.), *Wörterbuchkritik/Dictionary Criticism* (pp. 33–50). De Gruyter. https://doi.org/10.1515/9783110547108-003

Gouws, R. H. & Prinsloo, D. J. (2005). *Principles and practice of South African lexicography*. SUN Press.

Hütsch, A. (2017). Wörterbuchkritik zu allgemeinen zweisprachigen Wörterbüchern mit Deutsch, Englisch und Französisch: Eine Fallstudie. In M. Bielińska & S. J. Schierholz (Eds.), *Wörterbuchkritik/Dictionary Criticism* (pp. 303–322). De Gruyter. https://doi.org/10.1515/9783110547108-013

Kilgarriff, A., Husák, M., McAdam, K., Rundell, M. & Rychlý, P. (2008). GDEX: Automatically Finding Good Dictionary Examples in a Corpus. In E. Bernal & J. DeCesaris (Eds.), *Proceedings of the XIII EURALEX International Congress* (pp. 425–432). Institut Universitari de Linguistica Aplicada, Universitat Pompeu Fabra.

Kosem, I., Koppel, K., Kuhn, T., Michelfeit, J. & Tiberius, C. (2019). Identification and automatic extraction of good dictionary examples: the case(s) of GDEX. *International Journal of Lexicography*, 32(2), 119–137. https://doi.org/10.1093/ijl/ecy014

Magueresse, A., Carles, V. & Heetderks, E. (2020). *Low-resource languages: A review of past work and future challenges*. arXiv. https://arxiv.org/abs/2006.07264

Nielsen, S. (2009). Reviewing printed and electronic dictionaries: A theoretical and practical framework. In S. Nielsen & S. Tarp (Eds.), *Lexicography in the 21st Century: In Honour of Henning Bergenholtz* (pp. 23–41). John Benjamins.

Orife, I., Adelani, D. I., Fasubaa, T., Williamson, V., Oyewusi, W. F., Wahab, O. & Tubosun, K. (2020). *Improving Yorùbá diacritic restoration*. arXiv. https://arxiv.org/abs/2003.10564

Svensén, B. (2009). *A handbook of lexicography: The theory and practice of dictionary-making*. Cambridge University Press.

Swanepoel, P. (2017). Methods in dictionary criticism. In M. Bielińska & S. J. Schierholz (Eds.), *Wörterbuchkritik/Dictionary Criticism* (pp. 81–112). De Gruyter. https://doi.org/10.1515/9783110547108-005

Tarp, S. (2008). *Lexicography in the borderland between knowledge and non-knowledge: General lexicographical theory with particular focus on learner's lexicography*. Max Niemeyer.

Valcárcel Riveiro, C. (2023). Reutilizar la estructura del diccionario PortLex para crear un diccionario contrastivo de lenguas romances: una propuesta. *Quaderns de Filologia: Estudis Lingüístics*, 28, 125–139. https://doi.org/10.7203/qf.0.26560

Wiegand, H. E. (1983). Was ist eigentlich ein Lemma? Ein Beitrag zur Theorie der lexikographischen Sprachbeschreibung. *Studia Germanistica Gandensia*, 22, 401–474.

Wiegand, H. E. (1993). Wörterbuchkritik. Dictionary criticism. *Lexicographica*, 9, 1–7.

Wiegand, H. E. (2002). Äquivalenz, Äquivalentdifferenzierung und Äquivalentpräsentation in zweisprachigen Wörterbüchern. In H. E. Wiegand (Ed.), *Wörterbuchforschung* (Vol. 1, pp. 411–462). De Gruyter.

Contact information

Luke Omoyemi Akinremi
Universidade de Santiago de Compostela
Luke.Akinremi@usc.es

Elena Martín-Cancela
Universidade da Coruña
elena.martin@udc.es

Cormac Breathnach & Pádraig Ó Mianáin

REVERSING THE NEW ENGLISH–IRISH DICTIONARY DATABASE TO CREATE ENTRY FRAMEWORKS FOR IRISH–ENGLISH AND MONOLINGUAL IRISH DICTIONARIES

Abstract The first major monolingual Irish dictionary, *An Foclóir Nua Gaeilge* [The New Irish Dictionary] was launched in December 2025, when an initial tranche of 40,000 senses was published on focloir.ie. The dictionary will be added to incrementally in 2026 and 2027 to bring it to its content target of 80,000 senses. Along with the monolingual dictionary, a new bilingual Irish-English dictionary is also being produced and will be fully published in 2027. The entry frameworks for the new project were created by reversing the *New English-Irish Dictionary* (NEID) (2013-2017) database. This paper explores the decision to reverse the NEID database, the challenges involved in specifying the task, the resulting Irish-language entry frameworks and their use, the unintended consequences and related challenges, and the solutions found to those challenges.

Keywords Monolingual dictionary; bilingual dictionary; reversal; Irish-language lexicography

1 Introduction

The first major monolingual dictionary of Irish, *An Foclóir Nua Gaeilge* [The New Irish Dictionary] was launched in December 2025, when an initial tranche of 20,000 headwords and 40,000 senses was published on focloir.ie. The dictionary will be added to incrementally in 2026 and 2027 to bring it to its content target of c.30,000 headwords and 80,000 senses. Along with the monolingual dictionary, a new bilingual Irish-English dictionary, the first since 1977, is also being produced as part of the same project, with the publication of the entire bilingual dictionary scheduled to coincide with the completion of the monolingual dictionary in 2027.

This paper explores the options considered in order to create Irish-language entry frameworks for this dictionary project and the reasoning behind the decision to go with the option of reversing the *New English-Irish Dictionary* (NEID) database. The reversal process itself is then discussed, as well as the main challenges involved in specifying the task, the resulting Irish-language entry frameworks and their use, the benefits of the reversal, and some unintended consequences and related challenges.

2 Background

Foras na Gaeilge is the intergovernmental body in Ireland with responsibility for the development of terminology and lexicography in Irish. Its first lexicography project was the *New English-Irish Dictionary* (NEID), published 2013-2017, an online bilingual dictionary containing 48,000 headwords and some 150,000 senses. This was followed by the publication of the *Concise English-Irish Dictionary* (CEID) (2020), an abridged printed version of the online dictionary, which contains 30,000 headwords and approximately 80,000 senses. The last major English-Irish dictionary previous to that had been published in 1959.

When the CEID was published in late 2020, attention turned to planning a new project, either a monolingual dictionary or a new bilingual Irish-English dictionary. No major monolingual

dictionary of modern Irish had ever been published. *An Foclóir Beag* [The Small Dictionary] (Ó Dónaill & Ua Maoileoin) was published in 1991 but only contained 13,000 entries, many of which were simple cross-references. As modern lexicography in Irish had been revitalised by the NEID project, there was an increasingly vocal demand for a modern monolingual dictionary that would serve the needs of the Irish-language community and of the education system, where Irish is a compulsory subject at primary and secondary levels.

On the other hand, by this point, the most recently published major Irish-English dictionary was *Foclóir Gaeilge-Béarla* [Irish-English Dictionary] (Ó Dónaill) in 1977. A digital version of this dictionary had been published by Foras na Gaeilge on the teanglann.ie website in late 2013, and despite being almost 50 years old, the number of users accessing it had grown to almost two million per annum. The urgent need for a modern Irish-English Dictionary couldn't be ignored either.

3 Two dictionaries for the price of one

Ultimately it was decided that instead of focusing on one or the other of the two demands, Foras na Gaeilge would undertake the production of two new dictionaries in tandem: a monolingual Irish dictionary and an Irish-English dictionary, both with c.30,000 headwords and c.80,000 senses. The project began in September 2022 and is due to end in August 2027. The first tranche of 20,000 monolingual entries was published in December 2025 and is available at focloir.ie. The completion of the monolingual dictionary will coincide with the publication of the entire bilingual dictionary.

The rationale for undertaking two dictionaries concurrently was that the two dictionaries would be largely identical in terms of headwords, senses, usage examples, labelling, etc, and that consequently the same entry frameworks could be used for both dictionaries. Essentially, every sense unit would contain a definition field for the monolingual dictionary and translation fields for the bilingual dictionary, and the two different dictionaries could then be separately derived by stylesheets from the one database at the time of publishing.

The approach taken in the previous NEID project is an important point of context for discussing the current project. As discussed by Convery et al. (2010) and by Ó Mianáin & Convery (2014), a corpus-based approach had been taken in the NEID project in order to produce the English-language frameworks. The result was a database of 95,000 frameworks with detailed information on word collocations, word frequencies, real-life example sentences, grammatical information, idioms and phrases, and more (Atkins et al., 2010). This database was called DANTE (Database of ANalysed Texts of English) and is now available open source at dante-dictionary.com (Jakubíček et al., unpublished manuscript).

Beginning a new Irish dictionary project presented an interesting and novel challenge. Given the experience with the NEID project, three different options were considered to produce frameworks for the new Irish dictionary project.

3.1 Option 1: an Irish-language DANTE

Given the value of the DANTE frameworks as a basis for NEID, a corpus-based approach was seen as the ideal approach for the new Irish-language project; a comprehensive, systematic analysis of a suitably large Irish-language corpus would provide entry frameworks rich in detail about current usage of Irish. However, this approach was deemed unfeasible for two main reasons:

(a) The inadequacy of the available Irish-language corpora. The DANTE database was produced by analysing a 1.7-billion-word corpus of English. In 2021, the largest available corpus of Irish contained just over 100 million words. It was also felt that the balance of genres in the corpus was problematic, with legislative texts over-represented while spoken or casual Irish was under-represented.

(b) The lack of experienced lexicographers available. The twenty-strong team that produced the DANTE database each had on average 20 years of experience in lexicography. That team spent almost four years developing the database. By contrast, in 2021, Foras na Gaeilge had two full-time lexicographers, and between retirements and people moving on to other employment when their contracts with the NEID project came to an end, much of the Irish-language lexicography expertise developed between 2008 and 2020 was no longer available to the organisation. It was felt that even if a sufficient corpus of Irish were available, the task of training new staff to produce an Irish-language DANTE would prove too onerous and time-consuming, and all that before producing or publishing a single entry in the new dictionaries.

3.2 *Option 2: edit and develop the existing Irish-English dictionary of 1977*

Another option considered was to take the last bilingual Irish-English dictionary, *Foclóir Gaeilge-Béarla* [Irish-English Dictionary] (1977), and edit and develop it in order to produce a modern dictionary reflecting current usage of Irish. On the face of it, there were a number of advantages to this approach, for instance Foras na Gaeilge owns the copyright to the 1977 dictionary and an XML database of that dictionary had already been produced for the digital version of it published online at teanglann.ie in 2013. This XML database is stored in the same dictionary writing software as the NEID database (IDM's Dictionary Production System (DPS)).

That said, the data itself isn't very clean and would have required a considerable clean-up and conversion job in order to make it manageable in a familiar format for the current project.

More pertinently, the dictionary in question was already nearly 50 years old and was reflective of the time in which it was written. By today's standards, it is prescriptive and conservative, and quite old-fashioned in terms of style, presentation and language. It contains thousands of headwords and senses which wouldn't be included in a dictionary in this day and age due to infrequency of use. And of course, there are tens of thousands of new headwords and senses that have become commonplace in Irish in the intervening period that aren't in the 1977 dictionary.

Finally, the Ó Dónaill dictionary from 1977 is regarded as the 'bible' of standard Irish. It is an iconic work and its popularity among the Irish-language community has endured despite the passage of time. It was therefore felt that it could be a difficult and somewhat uncomfortable task for lexicographers to chop and change its content to conform to a modern style guide. Legitimate questions such as "should new senses be written as senses from the 2020s or should they be written in the style of the original publication?" would prove difficult to grapple with. It was also felt that editorial decisions could be interpreted by some as a criticism or invalidation of the original work, and the final product would be seen as a new version of an old dictionary rather than a completely new dictionary in and of itself. It was decided to let the sleeping giant lie.

3.3 Option 3: reversing the New English-Irish Dictionary

The final option considered was to take the entries of the *New English-Irish Dictionary* (NEID), published online 2013-2017, and reverse them. Simply put, every source language-target language pairing in the database would be inverted, with headwords, phrases, examples, etc, all swapping places with their translations.

One of the advantages of this approach was that it could be done in IDM's DPS. The staff were very familiar with the software, and it would also maintain a level of consistency across databases in terms of the names of fields, elements and attributes, and would allow for cross-referencing between sibling sense units in both databases.

Secondly, although specifying the job was difficult and painstaking on our side, and executing the job relied heavily on the technical expertise on IDM's side, preliminary investigations indicated that the various iterations involved in reversing the NEID database wouldn't take long compared with the years it would take to produce an Irish-language DANTE.

4 The chosen option: reversing NEID

Having considered the pros and cons of all three options, it was decided to proceed by reversing the NEID database to create the initial entry framework database for the new Irish dictionaries project. The expediency and ensuing cost-effectiveness of this option was the main reason behind the decision, plus the fact that those leading the new project would be familiar with the layout and composition of the resultant entry frameworks.

5 The results of the reversal process

The basic premise was simple: for each sense unit, take the Irish translation and make it a headword, and take the English headword and make it the corresponding translation. The Irish-language part of speech label in NEID was retained as the part of speech label in the new database. Domain labels were retained. A new 'definition' field was later added for the monolingual dictionary. Part of speech information for English-language content was removed.

5.1 Simple entries

The simple, terminological, single-sense entries in NEID were the easiest to reverse. There are almost 30,000 single-sense entries in NEID and the most simple of them – those with one headword, one sense and one translation – were quite straightforward to reverse. In the single-sense example from NEID below, the headword 'judicial review' and the Irish-language translation are reversed, giving us the entry framework for 'athbhreithniú breithiúnach':

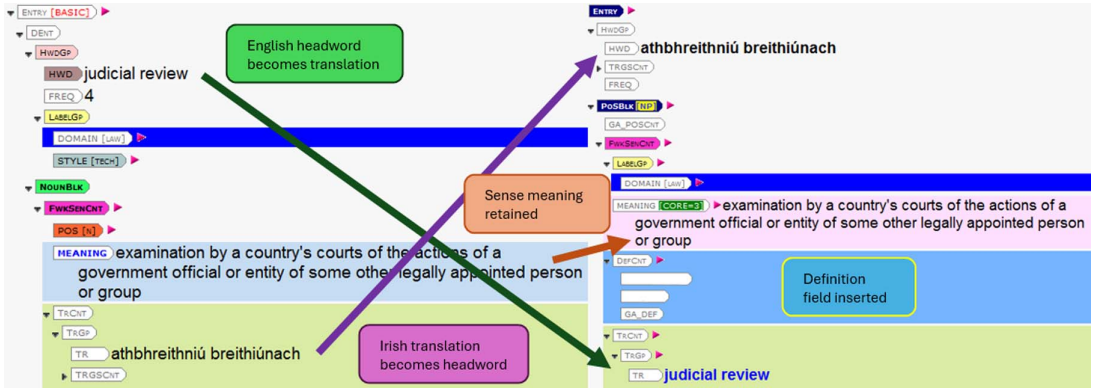


FIGURE 1 Simple reversal of entry 'judicial review' to create entry framework 'athbhreithniú breithiúnach'

The next level of complexity came in the case of NEID entries that had more than one translation in their sense unit. For example, the entry 'alkaline battery' in NEID has two translations: 'cadhnra alcaileach' and 'ceallra alcaileach'. The instruction in this case was to create a new entry framework for each of the translations, and to include 'alkaline battery' as the translation in both:

TABLE 1 Example of 'alkaline battery' in NEID and resultant two reversed frameworks

alkaline battery NOUN cadhnra alcaileach ceallra alcaileach	**cadhnra alcaileach** *NP* alkaline battery	**ceallra alcaileach** *NP* alkaline battery
'alkaline battery' in NEID	The resultant reversed 'cadhnra alcaileach' entry	The resultant reversed 'ceallra alcaileach' entry

This approach was taken regardless of the number of translations in a given NEID sense unit, for example, 'fleet-footed' below produced six reversed frameworks:

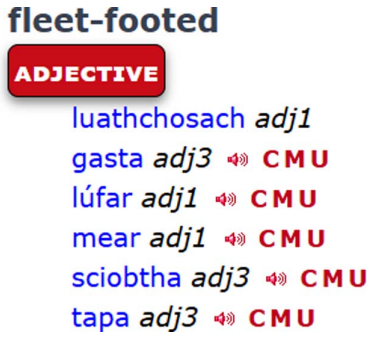


FIGURE 2 NEID entry 'fleet-footed'

5.2 Reversing multi-sense entries

A similar approach was specified in the case of multi-sense entries in NEID. For example, in the case of the entry 'agitator' in NEID, the translation from each sense became a unique entry framework in its own right:

TABLE 2 Example of 'agitator' in NEID and resultant four reversed frameworks

agitator NOUN 1 *POL* activist gríosóir *masc3* CMU suaiteoir *masc3* CMU 2 machine corraitheoir *masc3* CMU suathaire *masc4* CMU	**gríosóir** *MASC3* 1. *POL activist* agitator	**corraitheoir** *MASC3* *CONSTR machine* agitator
	suaiteoir *MASC3* 1. *POL activist* agitator	**suathaire** *MASC4* *machine* agitator
'agitator' in NEID, showing two different senses, each with two translations	The reversed 'gríosóir' and 'suaiteoir' entries	The reversed 'corraitheoir' and 'suathaire' entries

5.3 The same translation in multiple entries in NEID

The first significant structural challenge arose in the case of Irish words that appear as a translation in multiple entries in NEID, for example, the word 'iontach' in Irish which is used as a translation in 131 sense units in 119 entries in NEID. Rather than creating multiple entry frameworks with the same title in the new database, the specification was that any time a translation appeared more than once in NEID, **a new sense unit** in the corresponding reversed entry framework should be created. One of the first iterations of the reversal of the entry 'iontach' looked like this:

iontach

ADJ1

1. impressive
2. *excellent, wonderful* marvellous, marvelous
3. *very good, impressive* amazing
4. *very surprising* amazing
5. astonishing
6. *INF very good* awesome
7. *awesome, wonderful* breathtaking
8. *intelligent, skilful* brilliant
9. *interesting, exciting* colourful, colorful
10. delightful
11. *INF, DAT wonderful* divine
12. *very good* excellent
13. *wonderful* fabulous
14. *extremely good* fantastic
15. fascinating
16. formidable
17. *strange, surprising* funny
18. *wonderful, enchanting* magical
19. magnificent
20. *of rise, progress* meteoric
21. *extremely good* phenomenal
22. remarkable
23. *beautiful, impressive* splendid
24. stupendous
25. *INF, DAT* super
26. superb
27. *very great, intense* terrific
28. *INF very good, fantastic* terrific
29. *INF very good, fantastic* tremendous
30. *INF great, excellent* wicked

FIGURE 3 The first 30 reversed sense units for 'iontach'; translations in blue, sense disambiguators in italics

5.4 Invalid candidate entry frameworks

Another issue arising from the reversal of NEID was the creation of thousands of invalid multi-word entry frameworks that often *contained* the correct lemma but weren't valid headwords or

lemmas in their own right, but should be redirected as potential phrases or usage examples under their correct host frameworks. For example, all of the seven highlighted translations from the entry 'abandon' in NEID in the image below became candidate entry frameworks in the reversal, but only one of them ('fág') is a legitimate lemma:

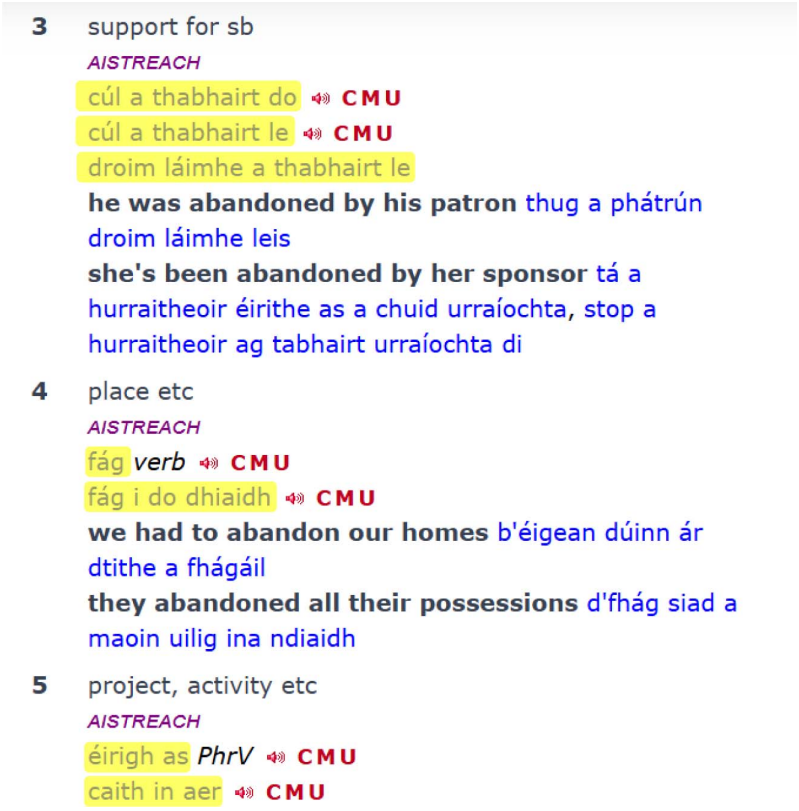


FIGURE 4 Three senses from the entry 'abandon' in NEID, base translations highlighted in yellow

As there was no means to automatically assign a particular word in a multi-word expression as the appropriate host lemma, the only, if somewhat painful, solution was to manually mark any invalid multi-word entry frameworks to be redirected to the correct lemma. All the multi-word entry frameworks were collated in a spreadsheet and checked line by line by editors, and either confirmed or rejected as valid frameworks. In cases where a framework was deemed invalid, the correct lemma was recorded in the next column. The final spreadsheet of invalid frameworks was used to redirect all of the content in each one as multi-word expressions under the correct lemma. This redirected material was marked in its new framework in order to distinguish it from genuine multi-word expressions emanating from the reversal of multi-word expressions in NEID.

Table 3 below gives an example of the work in the case of the highlighted translations from the three senses in 'abandon' given above. Note the legitimate framework 'fág' isn't redirected.

TABLE 3 Example of manual work redirecting invalid multi-word expressions to correct lemma frameworks

Candidate Entry Framework	Literal translation	Redirected to framework for
cúl a thabhairt do	to turn your back on	cúl
cúl a thabhairt le	to turn your back on	cúl
droim láimhe a thabhairt le	to turn the back of your hand towards	droim láimhe
fág	to leave	
fág i do dhiaidh	to leave behind you	diaidh
éirigh as	to give up	éirigh
caith in aer	to throw in the air	aer

The initial reversal of NEID produced a database of about 70,000 potential entry frameworks for the new project. Of those, some 20,000 invalid or non-lemma frameworks were redirected to their appropriate lemma, leaving a database of about 50,000 genuine entry candidates. Ultimately a lot of these invalid frameworks were not only invalid as headwords but also as multi-word expressions, and would eventually be downgraded to usage examples during the editing process.

5.5 Missing Irish-language frameworks

When the initial candidate frameworks for the new project were created by reversing the entries in NEID, some basic Irish-language lemmas didn't appear at all. For example, the common words 'aréir' and 'anuraidh' in Irish mean 'last night' and 'last year' respectively, but as there are no headwords or multi-word expressions for 'last night' or 'last year' in NEID (they appear in usage examples), the words 'aréir' and 'anuraidh' don't appear as a base translation in any entry. These two words and hundreds of others, including basic grammar words in Irish that don't have a direct equivalent in English, weren't created as candidate frameworks in the reversal because they were never there as base translations to be reversed in the first place. This highlighted the fact that the reversal process alone could not be relied upon to generate a valid headword list for an Irish-as-a-source-language dictionary.

To address this issue, a two-pronged approach was taken. First, the list of reversed candidate frameworks was compared with the list of headwords from the aforementioned *Foclóir Gaeilge-Béarla* [Irish-English Dictionary] (1977) and from *An Foclóir Beag* [The Small Dictionary] (1991) and any headwords not among the reversal results were noted. Then a list of lemmas was generated from the Corpus of Irish for Lexicography, a 100m-word corpus developed to support the new dictionary project in conjunction with the Gaois Research Group in Dublin City University (Ó Meachair et al., 2021), and the most frequent 20,000 lemmas were also recorded. IDM were then able to create basic template entry frameworks for these missing lemmas.

As a final extra measure to identify potential gaps in the Irish-language provision, editors monitor the database for gaps as they work through their batches of entries and record any additional headwords or senses they feel should be covered.

By mid-2026, almost four hundred new entries have been drafted from scratch and it is estimated that another three or four hundred more will be drafted by the end of the project.

6 Benefits of the reversal

6.1 Expediency and cost

The primary benefit of the reversal approach, and ultimately the reason it was chosen, was the cost – both financially and in terms of human resources. The reversal project produced a viable entry framework database within a year, with only two lexicographers working on it on a part-time basis. By comparison in terms of resources and time, the DANTE phase of NEID took almost four years to complete with up to twenty people working on it at various stages (Convery et al., 2010), and the alternative option of building a modern dictionary on the foundations of the Irish-English dictionary from 1977 would have required similar resources, time and funding. Given the staff resources available, the urgency to build on the renewed profile and momentum of contemporary lexicography in Irish as a result of NEID and CEID, and particularly given the fact that this foundation-laying work was at pre-project stage for which funding at the scale of DANTE or similar would simply not be made available, the reversal option was the catalyst for securing the funding and making the new Irish dictionary project possible. It is important to note, however, that despite the many benefits of reversing the NEID database, an Irish-language DANTE would still be considered the best possible option. The richness and depth of material produced by the DANTE project couldn't be replicated by reversing the NEID dictionary entry database.

6.2 Familiarity with the data structure

It was also beneficial to the Foras na Gaeilge team that the database generated in the reversal was a database with which they were familiar in terms of structure and layout. This greatly reduced the need for training and upskilling, and it made it much easier to work out and specify customisations to meet the emerging requirements of each of the two dictionaries being produced.

6.3 Modernity and relevance

As the NEID database is the basis for a contemporary English-Irish dictionary which is reflective of how both English and Irish are used in Ireland today, the reversal project created a similarly contemporary database that would serve as the basis for both the monolingual Irish dictionary and the bilingual Irish-English dictionary. And by the same token, both databases generally reflect and prioritise content that is relevant to modern Ireland in terms of culture and customs, current affairs, history and traditions, etc, as well as a focus on key terminology in primary and secondary education.

As discussed below, the reversal process wasn't perfect and the resulting database of entry frameworks created a number of challenges for the project. However, given the staffing, budgeting and time constraints at play, the reversal of NEID has proven to be a prudent and efficient approach for producing a database of entry frameworks for an ambitious new project that might otherwise not have got off the ground.

7 Challenges arising from the reversal

Outside of the pre-editing issues discussed above in relation to the composition of the reversed database, such as the 20,000 entry frameworks that needed to be manually redirected and the hundreds of missing entry frameworks that needed to be drafted from scratch, the following

were the main challenges that arose during the editing of the new dictionaries as a result of the nature of the reversed database.

7.1 Excessive English-language influence on Irish-language entries

At Euralex in 2008, Krek et al. discussed reversing the English-Slovenian dictionary in order to produce a framework for the Slovenian-English dictionary. They wrote about "the funny mirror of language from the lexicographer's perspective" and discussed how "the conversion process revealed a distorted image of the source language" (Krek et al., 2008, p. 541). This is consistent with our own experience in using the reversed English-Irish dictionary as a basis for an Irish-as-source-language dictionary. Without question, the biggest issue since editing began on the new dictionaries has been how sense boundaries and definitions in English are an ever-looming presence while Irish-language entries and senses are being identified, delineated and defined.

A number of particular circumstances exacerbated this issue for the new Irish dictionary project. Firstly, only one of the four full-time editors recruited for the project had any previous experience in lexicography.

Secondly, as is the case with all minority languages, the Irish language has been absorbing an increasing influx of words, phrases and usages from the dominant language, English, and there is a wide range of views on what is an acceptable borrowing and on how and when, if ever, borrowings become acceptable in the Irish-language canon.

Thirdly, as NEID had been the first English-Irish dictionary in fifty years, thousands of translation gaps had to be filled with translations without precedents in previous dictionaries; this in some cases resulted in approximate translations or in best-guess translations in NEID that may work within the sense parameters of an English-Irish entry but would not stand scrutiny as entries in a monolingual Irish dictionary (and subsequently as bilingual Irish-English entries). The scope for 'Englishisms' in NEID was increased as the translators and editors prioritised 'this is how I would say it' in the Irish content over more formal or dated Irish-language equivalents that might be found in older dictionaries. Ironically, an 'Englishism' appearing in NEID could, in itself, be taken as a validation that could decide the fate of a given headword or sense in the new dictionaries.

This issue manifested itself in a number of ways, for example,

(a) Entries or senses that would traditionally be deemed erroneous or inexact in Irish. In the case of the English-Irish entry for 'fleet-footed' in NEID (above), only the first translation is a fully reversable match for 'fleet-footed'. The other five translations are close matches and are suitable in the context of the English-Irish entry, but when reversed, would be unlikely to prompt 'fleet-footed' as an English-language equivalent (four of them mean 'quick' and the other 'nimble').

(b) Over-splitting of Irish-language senses on the basis of English-language near-synonyms. When the entry framework for 'crógacht' (= courage) was created in the reversed database, it contained senses from the NEID entries for 'gallantry', 'bravery', 'valour', 'daring' and 'courageousness'. These words may mean slightly different things in English, but the Irish word 'crógacht' had to be taken on its own merits and the final published entry contains only two senses; one for courage in the face of physical or mortal danger, and another for courage in adversity or in difficult circumstances. Similarly, the entry framework for the headword 'iontach' shown above gives an indication how editors have to be careful about

applying English-language sense boundaries to Irish-language words. In the case of 'iontach', we would end up with an Irish dictionary that (incorrectly) defines 'wonderful', 'stupendous', 'phenomenal', 'remarkable' and 'terrific' rather than the Irish word 'iontach'. The entry hasn't yet been completed but we expect the final entry will contain only two senses and a couple of multi-word expressions.

(c) Register labelling from English being erroneously applied to Irish content. A translation peculiarity in NEID was that a lot of informal to vulgar English words and phrases had no Irish equivalents in a similar register, and ended up being translated using 'vanilla' Irish words or phrases as equivalents, for example the word 'shit' in English is seen by most as a vulgar word but the Irish equivalent in NEID, 'cac', doesn't raise the same flags. When the NEID database was reversed, however, the register label transferred from the English content to the matching Irish content, and necessitated a specific task to verify which Irish-language entries or senses in the new dictionaries are genuinely informal, vulgar, dated, etc, in current Irish usage.

7.2 Invalid usage examples

In the course of inverting all headwords and their translations, all multi-word expressions and their translations, all usage examples and their translations in NEID were also inverted and, where more than one base translation was being inverted to form new reversed sense units, copied into each related entry framework. This in principle would be helpful to the editors as each sense unit would contain ready-made relevant usage examples, but it also resulted in a large number of usage examples that didn't actually contain the new Irish-language lemma or multi-word expression under which they were included in the reversed database.

For example, in the first sense of 'awkward' in NEID, one base translation is given: 'míchompordach'. However, in the translations of the usage examples, a number of other translation options are also given. This was done to enrich the NEID entry and give users more translation options. The flip side of this approach, when it came to the reversed database, was that the 'míchompordach' framework in the reversed database contained four usage examples that didn't contain the lemma 'míchompordach' at all. These usage examples are highlighted in red in the table below.

TABLE 4 Example of 'awkward' in NEID and reversed framework 'míchompordach'

embarrassed míchompordach *adj1* CMU **he felt awkward amongst them** ní raibh sé ar a chompord ina measc, bhí cearthaí air ina measc **she felt slightly awkward about it** mhothaigh sí sórt míchompordach faoi, ní raibh sí ar a compord faoi **there was an awkward silence** bhí tost míchompordach ann, bhí tost léanmhar ann	2. *embarrassed* awkward • ní raibh sé ar a chompord ina measc / bhí cearthaí air ina measc he felt awkward amongst them • mhothaigh sí sórt míchompordach faoi / ní raibh sí ar a compord faoi she felt slightly awkward about it • bhí tost míchompordach ann / bhí tost léanmhar ann there was an awkward silence
The first sense of the entry 'awkward' in NEID	The corresponding reversed sense in the entry framework 'míchompordach'

In a perfect world, all usage examples would be redirected to the entry framework pertaining to their main lemma, but similar to the invalid headwords mentioned above, there was no way of automatically redirecting them, and in any case, unlike the invalid headwords that at least

were full sense units in their own right, these invalid usage examples were just stand-alone examples that would be impossible to redirect to a valid host sense unit in any framework. It was therefore decided to leave all examples where they were and instruct editors to ignore the invalid ones. All examples were hidden in the stylesheet by default during the reversal process so that invalid examples wouldn't inadvertently appear in the published dictionary entries, and editors have to change the 'Hide' attribute if they choose to display a particular example.

8 Conclusion

The decision to reverse the *New English-Irish Dictionary* database in order to create entry frameworks for a monolingual Irish dictionary and an Irish-English dictionary has proven to be a pragmatic and effective strategy. The approach had its limitations but it proved to be an efficient way to generate a large-scale database of entry frameworks in spite of tight financial and time constraints. The challenges encountered in the project, especially the prevalence and influence of the English language in Irish-language frameworks, as well as the creation of tens of thousands invalid headwords, demonstrate the importance of human editorial intervention. Overall, the approach laid an important and viable foundation for two major dictionary resources for a minoritised language. The approach demonstrated one potential use case for existing data that is accurately tagged and of high quality, especially for lesser funded languages.

References

Atkins, B. S., Kilgarriff, A. & Rundell, M. (2010). Database of Analysed Texts of English (DANTE): the NEID database project. In A. Dykstra & T. Schoonheim (Eds.), *Proceedings of the XIV EURALEX International Congress* (pp. 549-556).

Convery, C., Ó Mianáin, P., Ó Raghallaigh, M., Atkins, S., Kilgarriff, A. & Rundell, M. (2010). The DANTE Database (Database of ANalysed Texts of English). In A. Dykstra & T. Schoonheim (Eds.), *Proceedings of the XIV EURALEX International Congress* (pp. 293-295).

Jakubíček, M., Rundell, M., Breathnach, C. & Ó Mianáin, P. (unpublished manuscript). DANTE resurrected: Large open source lexical database for English.

Krek, S., Šorli, M. & Kocjančič, P. (2008). The Funny Mirror of Language: The Process of Reversing the English-Slovenian Dictionary to Build the Framework for Compiling the New Slovenian-English Dictionary. In E. Bernal & J. DeCesaris (Eds.), *Proceedings of the 13th EURALEX International Congress* (pp. 535-542).

Ó Meachair, M. J., Ó Raghallaigh, B., Bhreathnach, Ú., Ó Cleircín, G. & Scannell, K. (2021). Tiomsú Corpais don Taighde Foclóireachta: Corpas Foclóireachta na Gaeilge (CFG2020). *TEANGA, the Journal of the Irish Association for Applied Linguistics*, 28, 278–305.

Ó Mianáin, P. & Convery, C. (2014). From DANTE to Dictionary: The New English-Irish Dictionary. In A. Abel, C. Vettori & N. Ralli (Eds.), *Proceedings of the 16th EURALEX International Congress: The User in Focus* (pp. 807-817).

Dictionaries:

An Gúm (1959). *English-Irish Dictionary*. (T. de Bhaldraithe, Ed.) Baile Átha Cliath. An Gúm.

An Gúm (1977). *Foclóir Gaeilge-Béarla [Irish-English Dictionary]*. (N. Ó Dónaill, Ed.) Baile Átha Cliath. An Gúm.

An Gúm (1991). *An Foclóir Beag [The Small Dictionary]*. (N. Ó Dónaill & P. Ua Maoileoin, Eds.) Baile Átha Cliath. An Gúm.

Ó Mianáin, P. (2020). *Concise English-Irish Dictionary*. Baile Átha Cliath. Foras na Gaeilge.

Websites:

Foras na Gaeilge & Lexical Computing. (2026). Retrieved from DANTE Lexical Database: https://dantedictionary.com/#/dante

Foras na Gaeilge. (2026a). Retrieved from New English-Irish Dictionary: www.focloir.ie

Foras na Gaeilge. (2026b). Retrieved from Dictionary and Language Library: www.teanglann.ie/en/

Contact information

Cormac Breathnach
Foras na Gaeilge
cbreathnach@forasnagaeilge.ie

Pádraig Ó Mianáin
Foras na Gaeilge
pomianain@forasnagaeilge.ie

Valeria Caruso & Geraint Paul Rees

FROM LEXICAL THEORY TO AI TRANSLATION

Procomplement Verbs and ChatGPT

Abstract Procomplement verbs (PCVs) are a highly frequent yet understudied class of semantically idiosyncratic verb–clitic constructions found across Romance languages. Despite their relevance for everyday communication, they remain only marginally represented in grammars and dictionaries. Building on recent theoretical work on Italian PCVs and drawing on Generative Lexicon theory, this study investigates the relationship between the syntactic–semantic properties of PCVs and their treatment by large language models. A dataset of 160 highly frequent Italian PCVs, primarily extracted from GRADIT and comprising 256 usage examples, was submitted to ChatGPT 5.2 for translation from Italian into English. The analysis focuses on the model's ability to account for the semantic, pragmatic, and phraseological complexity of these constructions. The results show that translation quality is influenced less by the specific subclass of PCVs than by transversal properties such as connotation and the pragmatic functions encoded by individual utterances. While ChatGPT generally succeeds in translating strongly conventionalised and highly idiomatic expressions, it frequently encounters difficulties with polysemous PCVs, context-sensitive pragmatic meanings, and morphologically complex verb–clitic combinations. The findings highlight the value of a fine-grained syntactic–semantic classification of PCVs for bilingual lexicography and suggest potential applications for the improvement of AI-assisted translation systems.

Keywords Procomplement verbs; phraseology; generative lexicon; bilingual AI-assisted lexicography; AI-assisted translation; ChatGPT; LLMs

1 Introduction

This paper aims to evaluate the performance of a large language model (LLM)-driven tool when translating a class of linguistic units that poses significant challenges for human translators and bilingual lexicography. In doing so, it hopes to shed light on ways to harness machine translation for lexicographic applications. More specifically, it aims to provide a clear impression of the ability of ChatGPT to translate Italian procomplement verbs (PCVs) into English and further inform theory about this verb class. Typologies of procomplement verbs from previous studies form an analytical framework for an assessment of ChatGPT 5.2's translation performance. The results reveal that, although the tool is capable of producing generally accurate and idiomatic translations, its output is not yet fully reliable. In particular, the analysis highlights recurring difficulties at the intersection of specific PCV subtypes and different dimensions of non-referential meaning, including the speech-act value conveyed by individual utterances and stylistic aspects of connotative meaning.

2 Procomplement Verbs (PCVs)

2.1 Features

Procomplement verbs represent a special class of semantically idiosyncratic verbs in Romance languages that are conjugated with complement clitic pronouns (De Mauro, 1999–2000, p. xxxiv), as illustrated by the following dictionary examples where PCVs (bolded) could reasonably be translated as 'to have it in for' (i.e. to harm or criticise someone):

Que veut dire cela? à qui ***en avez-vous?*** (Molière, in Le Robert)
Lit.: What does that want to say? At whom do you have it?
Qualcuno mi vuol male o addirittura il mondo intero ***ce l'ha*** *con me.* (Moravia, in GDLI)
Lit.: Someone wants evil to me, or even the whole world has it with me.
La vecina ***la ha cogido*** *con nosotros y nos hace la vida imposible.* (usage example drawn from the RAE)
Lit.: The neighbour has taken it with us and makes life impossible for us.

Although PCVs are high-frequency forms in everyday spoken Catalan, French, Italian, and Spanish, they have been largely disregarded in the literature and are only marginally covered by dictionaries. Building on Russi (2008) and generative studies such as Espinal (2009), Caruso (2025) identified two types of reference instantiated by clitics (anaphoric and deictic). Drawing on Pustejovsky's generative lexicon theory (Pustejovsky, 1995), she further proposed that different mechanisms govern the interaction between the semantics of verbs and those of their arguments lexicalized by clitics (or cliticised arguments). The interplay between reference and verb–argument structure allows for the identification of different subclasses of PCVs, providing a coherent syntactic–semantic description that is valuable for human-oriented lexicography—especially pedagogical lexicography—and potentially useful for improving AI-based resources.

It should be emphasised that the lemmatisation of PCVs represents one of the most significant achievements of Tullio De Mauro's GRADIT in the treatment of multiword expressions. First published in 1999, this ground-breaking dictionary was the first—and remains the only—work in Italian lexicography to include multiword expressions as lemmas. This innovation not only led to a substantial increase in the number of lexical items recorded for Italian, but also made a more accurate and reliable description of the Italian lexicon possible. In particular, it recognised PCVs as a special class of verbal lexemes which, although highly frequent in everyday interaction, had not been addressed elsewhere. However, unfortunately, De Mauro's treatment of this verb class is limited to the inventory of 159 PCVs provided in the dictionary and a brief explanation in the introduction:

> «In the lemma list of the dictionary [. . .] 159 pro-complement [verbs] are included, they have a meaning of their own, which cannot be derived from the main verb, or which is very opaque: *avercela, cavarsela, fregarsene* [Eng. to have it in for sb, to cope/manage, not to care about]». (De Mauro, 1999-2000, p. xxxiv)

Russi (2008) was the first scholar to explicitly acknowledge this gap in the literature in a comprehensive monograph on Italian clitics. Much of the volume is devoted to the behaviour of the most frequent verbs belonging to this class, which are examined within the context of a broader discussion on the continuum between lexicalisation and grammaticalization processes of Italian pronouns over time. Espinal's (2009) work is another important contribution to the study of PCVs across Romance languages. It focuses on the similarities between PCVs and the complex predicates found in polysynthetic languages, situating the discussion within the broader generativist debate on clitic processing. Espinal places particular emphasis on the most idiosyncratic PCVs, such as the Catalan *ballar-la* ('suffer', from *ballar*, 'dance') and highlights the lack of compositional mechanisms available for the interpretation of such idiomatic multiword expressions.

Caruso (2025) represents a further step forward in the study of this verb class and one to which the present analysis is indebted. Building on this work, the current study aims to systematically assess the contribution that large language models (LLMs) can make to the lexicographic description of PCVs, focusing in particular on the quality of Italian to English PCV translations provided by ChatGPT 5.2.

2.2 Classification

In Caruso (2025) *procomplement verbs* are attributed to two primary types of reference encoded by the lexicalized clitic. On the one hand, there are verbs where the clitic refers to or anticipates an element present in the surrounding text: *non so se nella valigia$_1$ ci$_1$ sta tutto* ('I don't know if everything fits in the suitcase'). In these cases, the clitic retains an *anaphoric* value. On the other hand, verbs such as *prenderle* [le botte] ('take a beating'), *arrivarci* [alla soluzione] ('get it/understand'), or *attaccarla* [una malattia] ('catch it', i.e. a disease) involve clitics that lexicalize an implicit argument, which is only made explicit in marked constructions such as left-dislocation: *le botte le hanno prese tutte nella caserma di Bolzaneto* ('they all got beaten up in the Bolzaneto barracks'). In the current analytic framework, the first type is referred to as *anaphoric procomplement verbs*, while the latter are called *deictic procomplement verbs*.

The distinction between the two modes of referentiality has syntactic ramifications. Deictic procomplement verbs have an unexpressed argument and are one-place predicates. Anaphoric procomplement verbs, by contrast, complete their meaning with an argument that is recoverable from the co-text or the communicative context. Thus, they take two arguments—at times, with obligatory expression of the second argument (*in salotto$_1$ ci$_1$ andrebbe un divano*, 'A sofa would be just right in the living room').

A further pertinent remark concerns the metalinguistic label introduced by De Mauro in GRADIT. The term *procomplement verb* contrasts with *pronominal verb*, which refers to verbs combining with other clitics in a strictly compositional way and grammaticalizing different types of diathesis, as in *abbracciarsi* ('hug each other'), *vestirsi* ('get dressed'), and *mangiarsi un panino* ('eat oneself a sandwich'). By contrast, procomplement verbs constitute a heterogeneous class of verbal multi-word expressions in which clitic complement pronouns are lexically integrated into the verb form, contributing semantically and/or syntactically in different ways. On this basis, PCVs can be grouped into several subtypes distributed along a continuum ranging from relatively transparent constructions, in which the clitic retains its argumental value, to fully lexicalized and semantically opaque idioms (e.g. *farcela*, 'make it').

Beyond referential anchoring, the semantics and syntax of procomplement verbs can be explained by general principles of argument omissibility and by standard mechanisms of argument–predicate interaction, as described in lexicalist approaches (e.g., Pustejovsky, 1995; Ježek, 2016). These mechanisms support a model that distinguishes between co-text-based reference (anaphoric PCVs) and indexical or ostensive reference (deictic PCVs), giving rise to various verb subtypes.

Among deictic procomplement verbs, several subtypes can be identified: PCVs in the central group exhibit a *default argument*, since the clitic replaces an implicit but highly predictable and culturally salient nominal argument which is congruent with the verb's semantics and its collocational/distributional profile. For instance, in *azzeccarla* ('guess right'), the clitic *la* conventionally evokes *la risposta* ('the answer'), agreeing in gender, number, and syntactic role with the cliticised noun phrase. A second group consists of *PCVs with a coerced argument type*, where the clitic presupposes an abstract or eventive complement that reshapes the verb's core meaning. In *pagarla*, for example, the clitic does not refer to a concrete object but rather encodes the event of "paying the consequences", metaphorically extending the literal meaning of a financial transaction. A further subtype is found in *predicative deictic PCVs*, where the clitic stands for an elided predicate rather than a nominal argument. In *provarci* ('try'), *ci* corresponds to an unexpressed infinitival clause (e.g., *provare a fare qualcosa*), thus functioning as a marker of predicate ellipsis. Some deictic PCVs also share striking resemblances to *light verb constructions*, for instance where the verb is semantically bleached, and the elliptic noun encodes the

actual predication. In *prendersela* ('take offense'), *la* encodes an inferred nominal event like 'an offense' or 'a reason for anger'. In contrast, the fully lexicalized PCVs, or *idiomatic PCVs*, are completely opaque and no longer decomposable. Verbs like *battersela* ('flee'), *spuntarla* ('succeed'), and *sguazzarsela* ('live it up') exhibit total semantic opacity, with clitics that no longer exhibit referential function.

Alongside the deictic PCVs, the second macro-category consists of *anaphoric procomplement verbs*, where the clitic relates to an element present in the co-text or discourse context which is necessary to complete the verb's semantics and syntax. In *ne$_i$ andrà del suo prestigio$_i$* ('his/her reputation will be at stake'), *ne* marks an obligatory argument tied to the predicate. Yet, among anaphoric PCVs there are also instances where predicates may remain implicit, thus offering instances of *null-object instantiations*. Examples include verbs like *fregarsene* ('not to care'), *entrarci* ('be relevant'), and *farsene* ('make use of'), in which *ne* (and occasionally *ci*) is fully grammaticalized and recoverable via inference: *che se ne fa di tutta quella roba?* can also be found as *che se ne fa$_\emptyset$* ('what does he need all that stuff for?').

A final, distinct class is made up of *connotative-pragmatic procomplement verbs*, where the clitic contributes affective nuance rather than argumental content (Masini, 2015). These verbs actually express a "benefactive" meaning (Ježek, 2004) similar to the middle diathesis in ancient Greek. Russi (2008) argues that they encode a form of subjective involvement on the part of the speaker or another participant. For example, in an utterance such as *lei dovrebbe smetterla di bere*, a fitting translation would be 'you really should stop drinking' or even 'you need to stop drinking' rather than a literal 'you have to stop drinking'. Therefore, in verbs such as *smetterla* and *finirla*, the clitic *la* is considered as a marker of *affectedness*: its presence signals that the speaker is personally affected by, or emotionally involved in the situation, whereas its omission yields a more neutral and detached interpretation. This interpretation is also expressed by Masini (2015) who, for her part, analyses the clitic cluster *sene* as the marker of a specific emphatic function [+EMPHASIS]. In this construction, *sene* does not contribute referential or locative meaning, but rather signals the involvement of the subject (or another participant) in the event, highlighting a heightened degree of subjective participation: *(me ne) sono andata a spasso* might be rendered as 'I took myself off for a walk', rather than simply 'I went for a walk'. Masini further relates this function to the Italian transitive middle construction (*mangiarsi, bersi*), where the clitic *si* similarly marks subject involvement. Unlike the *transitive middle*, however, the *sene* construction is largely restricted to intransitive predicates and may therefore be regarded as a type of *intransitive middle construction*. Synchronically, it displays a certain degree of autonomy: it is less lexicalized than many other PCVs, exhibits analogical productivity, and is distinguished by its primarily pragmatic rather than argument-structural contribution.

Overall, this classification outlines a richly articulated continuum of procomplement constructions in Italian, where clitics function either as argument saturators or as pragmatic-emotive markers. Their degree of syntactic and semantic integration determines the placement of each verb type within the broader continuum of lexicalization and grammaticalization processes (Russi, 2008).

2.3 Connotation

Apart from the subtle differences discussed so far, PCVs are, for the most part, connoted and informal lexical items used to convey expressive nuances and pragmatic values. For this reason, they tend to be underrepresented in written discourse and in more formal or controlled communicative contexts. This has undoubtedly contributed to the very limited attention they have received in both grammars and dictionaries, resulting in a significant gap not only in Romance

linguistics, but also in the broader understanding of the phraseological continuum. Indeed, comparable pronominal phrasemes can be found in a variety of other languages. Consider, for instance, the English expressions *do it with someone*, *push it/push your luck*, and *like it or lump it*, where the pronoun *it* contributes, to varying degrees, to the semantics of the multiword expression. Masullo & Bertora (2014, p. 203) make a similar point, observing that "Asimismo, abundan en inglés las locuciones con *it* como objeto expletivo" ('expressions containing *it* as an expletive object also abound in English').

By exploiting the semiotic potential of pronouns, PCVs reveal interesting mechanisms of compositional semantics within the domain of complex lexical units. Through the incorporation of implicit arguments, they give rise to synthetic lexicalisations characterised by a high degree of semantic transparency, a feature that arguably enhances their descriptive and expressive potential. The role of connotation as a phenomenon arising from figurativeness has been extensively discussed by Dobrovol'skij & Piirainen (2005), who regard idioms as additional means of lexicalising concepts in languages. In their view, idioms are connoted lexical items that speakers choose in order to express their ideas more effectively, since the literal meaning associated with an idiom contributes to a more evocative form of communication. This observation may help explain the widespread use of PCVs in everyday language, where they are often preferred as a means of conveying connoted meanings such as aggression, personal involvement in the event described, and even obscene or sexual undertones. As predominantly informal lexical items used in spontaneous communication, PCVs constitute a valuable expressive resource, as they allow speakers to encode evaluative and affective nuances directly within the verbal expression itself, thereby enriching the expressive repertoire available in everyday interaction.

3 The Experiment

3.1 Data

To assess the suitability of current large language models for lexicographic tasks, an experiment was conducted using data extracted from the *GRADIT*. The aim was to evaluate the performance of ChatGPT 5.2 in the translation of usage examples. It was hoped that the candidate translation would provide a valuable starting point for identifying appropriate cross-linguistic correspondences and enriching bilingual dictionary entries.

The analysis was based on a dataset of 160 Italian PCVs. The original inventory extracted from GRADIT was refined by excluding obsolete verbs and regional forms no longer widely used (e.g., *sbroccolarsela*). In addition, polysemous PCVs listed as subsenses in GRADIT were treated as separate lexical units and added to the list of lemmas. Five noteworthy lexemes that were not included in the dictionary (Russi, 2008), but are highly frequent in contemporary standard Italian (Berruto, 1987), such as *metterci* and *saperne*, were also added. Usage examples for these additional lemmas were retrieved from the ItTenTen20 corpus.

3.2 Translation Procedure

All examples were submitted to ChatGPT 5.2 with an explicit prompt not to alter, shorten, or reformulate the source text, and to provide both the original sentence and its English translation as output. The lexical items were presented in batches of fifty entries, organized in tabular form and marked with numerical identifiers. In several cases, multiple examples were available for the same PCV, making it possible to observe intra-lexical variation in the model's

translation behaviour. The final dataset comprised 256 example sentences, corresponding to an average of 1.6 examples per lexical unit.

3.3 Analysis Procedure

Following Caruso (2025), each PCV was assigned to one of three major classes: deictic, anaphoric, or connotative-pragmatic. The two authors of the paper, a native speaker of Italian and a native speaker of English, served as expert annotators in evaluating the translations produced by ChatGPT. Given the subtle semantic and pragmatic nuances conveyed by many of the example sentences, as well as their often limited contextual information, the annotation process was conducted by consensus rather than through the calculation of inter-annotator agreement. In many cases, it proved difficult to determine whether the English translation fully preserved the semantic and pragmatic values of the source sentence or, conversely, introduced meanings absent in the original, as illustrated by the usage examples discussed below ([1]-[14]).

Consider, for instance, the examples in [1] and [2], where ChatGPT's translation is given. In [1], the English expression *to lose one's life/health* does not adequately convey the affective meaning encoded by the Italian *rimetterci la vita, la salute*, which is more effectively rendered by expressions such as *to pay for something with one's life/health*. Similarly, in [2], the translation *You're wrong if you think you can move me* has a predominantly descriptive value and fails to convey the reproachful attitude lexically encoded by the Italian expression *la sbagli se . . . .*

1. rimetterci la vita, la salute → *to lose one's life, one's health* [stylistic error];
2. la sbagli se pensi di commuovermi → *You're wrong if you think you can move me* [pragmatic error].

Similar decisions concerning semantic values extending beyond the referential meaning of Italian PCVs had to be taken jointly by both annotators to ensure that the transfer from the source language to the target language adequately reflected the specific semantic and pragmatic properties of these lexical items. Given the features of PCVs discussed above, and in particular their ability to convey specific speech acts such as orders, reproaches, and threats, translation accuracy was assessed using four parameters: denotation, style, affect, and pragmatics.

Misinterpretations of the referential value of the source sentence were classified as denotative errors. Examples [3]–[6] illustrate this type of error; the expected translations are provided in square brackets.

3. non ci va, è inutile provare → *It won't work, there's no point in trying* [It **doesn't fit (in)**, there's no point in trying]
4. Hai la macchina? Sì, ce l'ho. → *Do you have **a** car? Yes, I do* [Do you have **the** car? Yes, I do]
5. ti aspetto a cena, guarda che ci conto → *I'll wait for you for dinner, I'm counting on it* [I'm expecting you for dinner, I'm counting **on you**]
6. Carlo e Giacomo se le sono date per lei → *Carlo and Giacomo **beat each other up** over her* [**fought, got into a fight, came to blows**]

Stylistic errors arise when translations fail to preserve the sociolinguistic and interactional cues encoded by the source expression (Coseriu, 1969; Cruse, 1986, pp. 283-285). A useful diagnostic consists in asking: *"Who would use this expression?"* and *"In what context or situation would it be*

appropriate?". In examples [78], the English translations preserve the propositional content of the Italian utterances but neutralize their stylistic and interpersonal meaning. Consequently, they fail to convey the familiarity with the interlocutor and the informal setting implied by the original expressions.

7. le ho buscate da mio padre → *I got beaten by my father* [My father gave me a beating]

8. quello che dici non ci azzecca! → *What you're saying has nothing to do with it!* [What you are saying doesn't make sense]

The affective dimension encompasses utterances that encode a heightened degree of speaker involvement (Cruse, 1986, pp. 275-278; Ježek, 2016, p.43). While this feature is most salient in the connotative-pragmatic PCVs discussed above, it is also observable in other PCV types, as illustrated in example [9-11]. A useful diagnostic is whether the following question can be answered positively: *"Does this utterance express a heightened degree of speaker involvement in the event or toward one of its participants?"*.

9. Connotative-pragmatic PCV:
   Perché non se ne restano a casa loro e cercano dicomunicare? → *Why don't they stay at home and try to communicate?* [Why don't they just stay at home where they belong and try and communicate?]

10. Idiomatic PCV:
    Ho come l'impressione che Virtual DJ sia diventato un software obsoleto e non se lo fumi più nessuno" → *I have the feeling that Virtual DJ has become obsolete software and that no one takes it seriously anymore* [I get the impression that Virtual DJ has become obsolete and nobody touches it anymore]

11. Deictic PCV:
    ci hai azzeccato a scegliere la tua professione → *You made the right choice in picking your profession* [You chose the right professions, You were spot on . . . , You made a good call . . .]

As regards pragmatics, translations can be considered successful insofar as they preserve the illocutionary force of the source utterance. The relevant question is therefore whether the target sentence can be interpreted as the same speech act as the original (e.g. a request, a compliment, or a threat). In examples [12-14], however, this requirement is not fully met. Although the referential content of the Italian utterances is preserved with little difficulty, the translations reduce them to straightforward descriptive or locutionary acts. As a result, they fail to convey the illocutionary force of the compliment in [12], the directive force of the order in [14], and the threatening force associated with the interrogative utterance in [13].

12. Anaphoric PCV
    Anna così magra ci guadagna → *Anna looks better being that thin* [Being that slim suits Anna]

13. Idiomatic PCV
    ce l'hai con me? → *Have you got something against me?* ["Are you mad at me?" or "Do you have it in for me?"]

14. Connotative-pragmatic PCV
    adesso è ora di finirla → *Now it's time to stop* [That's enough now/enough already]

4 Results and Discussion

The results reveal both the strengths and weaknesses of AI in the translation of PCVs. The overall translation accuracy rate of 82% (see Table 1) is relatively high. However, it means that approximately one in five examples were translated incorrectly—a significant problem for real-world lexicographic applications.

TABLE 1 Comparative overview of ChatGPT's translation accuracy by PCV type, illustrating the frequency of successful translations and error categories across the dataset (N = 1,024). The total number of observations derives from the evaluation of 256 usage examples across four assessment parameters

Type of error	Anaphoric PVCs	Deictic PVCs	Idiomatic PVCs	Pragmatic-affective PVCs	Total
none	238	280	191	129	838
stylistic	16	23	22	12	73
affective	11	20	18	10	59
denotative	8	12	5	2	27
pragmatic	7	5	12	3	27
Total	**280**	**340**	**248**	**156**	**1024**
none	85%	82%	77%	83%	82%
stylistic	6%	7%	9%	8%	7%
affective	4%	6%	7%	6%	6%
denotative	3%	4%	2%	1%	3%
pragmatic	3%	1%	5%	2%	3%

The main difficulties do not appear to depend primarily on the specific subclasses of PCVs identified above. Rather, they stem from a set of transversal properties across the different classes identified so far, especially the connotative dimension and the correct interpretation of performative values, or speech-act functions. To some extent, highly idiomatic [15] and strongly conventionalized [17] PCVs tend to receive successful translations, and the model is often able to preserve their connotative or emphatic meaning, see especially [16]. Greater challenges arise, however, when the translation requires the recognition and rendering of pragmatically loaded meanings, such as the benefactive values associated with connoted-pragmatic PCVs (Ježek, 2004; Masini, 2015; Russi, 2008, pp. 54–57), whose interpretation is highly context-sensitive. In such cases, translations may be lexically accurate yet pragmatically inadequate [18]. More generally, ChatGPT often struggles to reproduce the illocutionary and discourse-oriented functions encoded by PCVs, particularly when a high degree of idiomaticity is coupled with stylistic, affective, and pragmatic values (see Table 1). In such cases, translations tend to preserve the propositional content while failing to convey the intended pragmatic force [19].

15. dopo lunghe discussioni l'ho spuntata.
    Lit.: After long discussions, I have blunted it.
    After long discussions, I got my way

16. chi se ne frega?
    Lit.: Who rubs himself about it?
    Who gives a damn?

17. È ora di piantarla con queste sciocchezze.
    Lit.: It is time to plant it with these foolish things.
    It's time to stop this nonsense.

18. Me ne vado al cinema.
    Lit.: I go myself away to the cinema.
    #I'm going to the cinema.
    =I'm taking myself to the cinema.

19. guarda come se la ride!
    Lit.: Look how he/she laughs it to himself/herself!
    ?Look at him, laughing it off!
    =Look at him, having a good laugh!

Similarly, when idiomaticity interacts with stylistic variation, affective meaning, or specialized pragmatic functions, ChatGPT's output may become problematic. In [20], the model fails to recover the intended speech-act value of the utterance: the original Italian sentence conveys a reproach directed at the addressee's reluctance to carry out his/her duties, whereas the translation incorrectly frames the complaint in terms of the effort required to complete the task.

20. ha finito il lavoro, ma ce n'è voluto! → *He finished the job, but it took a lot!* [He finished the job, but it took **him** long enough!]

In addition, the system shows an uneven ability to handle subtle semantic contrasts induced by clitic variation. In some cases, fine-grained distinctions are successfully captured, as with *ridersela* vs. *ridersene* ('laugh off' vs. 'scoff at'). However, literal translations are also produced with morphologically complex PCVs involving multiple clitics, such as *ridacchiarsela*, which is rendered as 'giggling to themselves', incorrectly interpreting the clitic cluster as a reflexive adjunct rather than as a lexicalized unit.

Weaknesses emerge in the translation of the more idiomatic senses of polysemous PCVs. This is the case, for instance, of *farla* referring either to defecation [21] or to playing a trick on someone [22], which ChatGPT renders through inappropriate collocations. This failure to account for metaphorical and idiomatic meanings is reminiscent of findings reported in previous studies on the use of ChatGPT in lexicography (cf. Jakubíček & Rundell, 2023). By contrast, ChatGPT performs well when the different senses of a polysemous PCV correspond to strongly conventionalised expressions in English, as in *andarsela a cercare*, which is appropriately rendered as *to bring it on oneself, to go looking for trouble*, or *to ask for it*.

21. l'ha fatta nel vater → *He did it in the toilet*

22. farla a un avversario → *to pull a trick on an opponent* [play a trick on/get the better of/to pull x off against an opponent]

Interestingly, ChatGPT often fails to capture the different degrees of intensification encoded by the Italian PCVs expressing INDIFFERENCE. The semantic domain is structured along a gradient of expressive intensity, ranging from the relatively neutral *importarsene* and the more affectively marked *stropicciarsene* to the progressively more vulgar *fottersene*, *sbattersene*, and *strafottersene*, examples from [23] to [29]. For this set of verbs, the tool provides appropriately graded translations only when the contrastive degrees of intensification are made explicit within the examples themselves, as in [27].

23. non me ne importa niente, la cosa mi è indifferente → *I don't care at all, it makes no difference to me*

24. me ne stropiccio di lui! → *I don't give a damn about him!*
25. non me ne fotte niente!→ *I don't give a damn about it!*
26. me ne fotto di lui e delle sue prediche! → *I don't give a damn about him and his lectures!*
27. sbattersene l'anima, le palle, i coglioni → *to not give a damn at all/to not give a shit/to not give a fuck*
28. me ne strafotto del giudizio degli altri → *I really don't give a damn about other people's judgment*
29. strafottersene dei genitori → *to really not give a damn about one's parents*.

Overall, the findings confirm that, while ChatGPT 5.2 can produce reasonably idiomatic translations of PCVs under favourable conditions, it still struggles with their polysemy, pragmatic load and, in some cases, their morphological complexity. These results support the relevance of a fine-grained syntactic–semantic classification of PCVs not only for pedagogical lexicography, but also as a resource for improving AI-driven translation systems.

A worthwhile extension of this protocol would be to test other LLMs, particularly those more extensively trained on Italian, in order to identify any significant differences with respect to the translations examined so far. Furthermore, it would be particularly interesting to evaluate ChatGPT again by providing it with the translation guidelines derived from the present study, thereby assessing its ability to improve its performance through targeted prompting and instruction-based adaptation. Despite these avenues for further research, the translations of the example sentences collected so far and refined within the present discussion constitute a solid starting point for the development of a high-quality bilingual lexicographic resource covering the highly idiosyncratic class of PCVs.

References

Berruto, G. (1987). *Sociolinguistica dell'italiano contemporaneo*. Roma: La Nuova Italia Scientifica.

Caruso, V. (2025). *Tra lessico e grammatica: I verbi procomplementari*. UniorPress.

Coseriu, E. (1969). *Einführung in die strukturelle Linguistik*. Narr.

Cruse, D. A. (1986). *Lexical semantics*. Cambridge University Press.

De Mauro, T. (1999–2000). Introduzione. In *Grande dizionario italiano dell'uso* (pp. VII–XLVII). UTET.

Dobrovol'skij, D., & Piirainen, E. (2005). *Figurative language: Cross-cultural and cross-linguistic perspectives*. Elsevier.

Espinal, M. T. (2009). Clitic incorporation and abstract semantic objects in idiomatic constructions. *Linguistics*, 47(6), 1221–1271.

Jakubíček, M., & Rundell, M. (2023). The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography? In M. Medveď, M. Měchura, I. Kosem, J. Kallas, C. Tiberius, & M. Jakubíček (Eds.), *Electronic lexicography in the 21st century (eLex 2023): Invisible Lexicography. Proceedings of the eLex 2023 Conference* (pp. 518–533). Lexical Computing CZ.

Ježek, E. (2004). Interazioni tra aspetto e diatesi nei verbi pronominali italiani. *Studi di grammatica italiana*, 22, 239–281.

Ježek, E. (2016). *The lexicon: An introduction*. Oxford University Press.

Masini, F. (2015). Idiomatic verb-clitic constructions: Lexicalization and productivity. In *Morphology and Semantics – MMM9 Online Proceedings* (pp. 88–104).

Masullo, P. J., & Bertora, H. R. (2014). Objetos acusativos expletivos en el español rioplatense. In *VI Congreso Internacional de Letras: Transformaciones culturales. Debates de la teoría, la crítica y la lingüística* (pp. 195–205). Facultad de Filosofía y Letras, Universidad de Buenos Aires.

OpenAI. (2026). *ChatGPT* (GPT-5.2) [Large language model]. Retrieved February 31, 2026, from https://chatgpt.com

Pustejovsky, J. (1995). *The generative lexicon*. MIT Press.

Russi, C. (2008). *Italian clitics: An empirical study*. Mouton de Gruyter.

Sun, Y., Sheng, D., Zhou, Z., & Wu, Y. (2024). AI hallucination: Towards a comprehensive classification of distorted information in artificial intelligence-generated content. *Humanities and Social Sciences Communications*, 11(1), 1–14.

Dictionaries

Battaglia, S. (Ed.). (1961–2002). *Grande dizionario della lingua italiana*. UTET.

De Mauro, T. (Ed.). (1999–2000). *Grande dizionario italiano dell'uso*. UTET. (GRADIT)

Real Academia Española. (2021). *Diccionario de la lengua española* (23rd ed.). Retrieved May 31, 2026, from https://dle.rae.es

Robert, P., & Rey, A. (2001). *Grand Robert de la langue française*. Dictionnaires Le Robert. Retrieved May 31, 2026, from https://www.lerobert.com

Contact information

Valeria Caruso
Università di Napoli L'Orientale
vcaruso@unior.it

Geraint Paul Rees
Universitat Pompeu Fabra
geraintpaul.rees@upf.edu

Acknowledgements

We gratefully acknowledge the support of the University of Naples "L'Orientale" PRA Project 2025–2027: *I verbi procomplementari italiani, spagnoli e catalani: un'analisi teorica e comparativa per la traduzione e la comunicazione interculturale* – Italian, Spanish and Catalan Procomplement Verbs: A Theoretical and Comparative Analysis for Translation and Intercultural Communication.

Zoe Gavriilidou

SPANISH LOANWORDS IN MODERN GREEK

Linguistic Analysis and Lexicographic Treatment in the Dictionary of Standard Modern Greek

Abstract This study aims to provide a systematic linguistic and lexicographic analysis of Spanish-derived loanwords in Modern Greek, addressing the limited attention given to Spanish as a donor language. Methodologically, it draws on a dataset of 130 items, combining 105 entries from the Dictionary of Standard Modern Greek (DSMG) with 25 additional items collected from contemporary digital discourse. The analysis adopts a typological framework distinguishing loanwords, loanblends, and loanshifts, alongside grammatical, morphophonological, and semantic-domain analyses.

The results indicate a clear predominance of loanwords, with loan translations and semantic extensions occurring less frequently. Nouns constitute the dominant grammatical category, while morphophonological integration is largely systematic, aligning with Greek inflectional patterns. Semantically, borrowings cluster in culturally salient domains such as food, music, and material culture, whereas core vocabulary remains largely unaffected. The findings also highlight the predominantly indirect transmission of Spanish borrowings through intermediary languages and the accelerating role of global media in recent diffusion.

The study contributes by documenting Spanish as a secondary but meaningful donor language in Modern Greek and by identifying gaps in the DSMG, advocating for more systematic etymological attribution and a usage-based, dynamic approach to lexicographic practice.

Keywords Language contact; lexical borrowing; Spanish loanwords; Modern Greeκ; Greek Lexicography

1 Introduction

Linguistic borrowing constitutes one of the most dynamic phenomena of language contact, as it is directly associated with the formation of neologisms, language change, and the evolution of linguistic systems. A considerable body of research has approached the phenomenon from diverse theoretical and methodological perspectives (Humbley, 1974; Thomason & Kaufman, 1988; Anastasiadis, 1994; Matras, 1998; Thomason, 2001; Johanson, 2002; Matras, 2007; Haspelmath, 2008; Haspelmath & Tadmor, 2009), addressing a wide range of issues. These include the typology of borrowings (Haugen, 1950; Anastasiadis, 1994; Matras & Sakel, 2007), the motivations, pathways, and underlying causes of borrowing—such as cultural influence, the reflection of historical events in the lexicon, the representation of stereotypes, and naming needs (Haspelmath, 2008)—the degree and intensity of language contact that facilitate borrowing (Thomason & Kaufman, 1988), the lexical categories and parts of speech that are more susceptible to borrowing (Myers-Scotton, 2002; Matras, 2007), as well as the linguistic elements and structures that resist or readily accommodate borrowing, as captured in borrowability scales (Matras, 1998). Further lines of inquiry concern synchronic and diachronic approaches to data analysis and the processes and stages through which borrowings are adapted and integrated into the recipient language (Anastasiadis, 1994).

Despite the extensive international literature on the topic, borrowing in Modern Greek has attracted comparatively limited attention within the Greek linguistic community. Notable exceptions include the foundational works of Anastasiadis (1986; 1994), which focus on

neology and borrowing from French and English; Charalambakis (1991) and Petrounias (1995) on calques in Greek; Papanastasiou (2003) on borrowing; the dissertations of Vasmanoli (2007) and Zafeiri (2015) on reborrowings and calques; Fliatouras & Kokkini (2017) on semantic borrowing in Greek terminology; the studies of Gavriilidou (2018) and Gavriilidou (2019) on borrowings from Russian; Mitits & Gavriilidou (2020) on loanblends in the speech of Greek heritage speakers; and Androulidakis (2019) on borrowings from German. At the same time, several studies have highlighted limitations in contemporary Greek dictionaries regarding the incorporation and lexicographic treatment of borrowings (Anastasiadis, 1994; Gavriilidou, 2018; Androulidakis, 2019).

In light of these gaps, the present study aims:

(a) to document and systematically examine the characteristics of borrowings from Spanish into Greek, which have not been thoroughly investigated to date, given that Spanish is not traditionally regarded as a major donor language, unlike French, English, Italian, or Turkish, in the Greek context;

(b) to analyze the lexicographic treatment of Spanish borrowings in the *Dictionary of Standard Modern Greek* (DSMG) of the Triantafyllidis Foundation and to propose the inclusion of additional loanwords that appear to have entered the Greek lexicon in recent years.

The study begins by outlining the methodology and clarifying the theoretical concepts used. It proceeds to examine the nature of contact between Greek and Spanish and the channels of borrowing activated across different historical periods. This is followed by a detailed analysis of the characteristics of the borrowings, including their types, morphophonological adaptation, and semantic classifications. The study concludes with a critical discussion of their lexicographic treatment in the DSMG.

The research addresses the lack of systematic documentation and lexicographic evaluation of Spanish borrowings in Modern Greek. Its contribution lies in shedding light on a relatively understudied aspect of the history of Modern Greek by focusing on Spanish as a source language whose role in lexical transfer has not been systematically explored. Its originality derives from the dual linguistic and lexicographic perspective adopted. The study contributes to the broader discussion of linguistic borrowing in Modern Greek by introducing data from a less prominent donor language and supports the advancement of Modern Greek lexicography by emphasizing the need for a more comprehensive and dynamic representation of borrowed elements.

2 Methodology and adopted terminology

The present study is based on a dataset of 105 lexical items compiled through advanced search procedures in the DSMG. The dataset includes entries whose etymological field in the DSMS contains the marker “Sp.”, the adjective “Spanish,” or the proper noun “Spain.” Borrowings from Spanish are defined as lexical items that entered Greek either directly from Spanish or indirectly through intermediary languages, including French, English, and Italian. The list was further supplemented with 25 additional Spanish-derived items not recorded in the DSMG, identified through systematic web searches and selected on the basis of recurring attestation in contemporary Greek usage.

Within the framework of this study, Spanish is defined as the source or donor language (L2), whereas Greek is treated as the recipient language (L1). The term borrowing is used in

accordance with Thomason & Kaufman (1988, p. 21), referring to "the incorporation of foreign elements into the speakers' native language."

With respect to the classification of borrowed forms, the analysis adopts the typology proposed by Einar Haugen (1950), in combination with the framework developed by Anastasiadis (1994). On the basis of this combined typological model, three principal categories are distinguished: loanwords, loanblends, and loanshifts.

Loanwords refer to lexical items incorporated into the vocabulary of L1 from L2 in the context of language contact. This category includes internationalisms, defined as items with a shared etymological origin and similar meaning across multiple languages (e.g., telephone), as well as foreignisms, which denote culturally specific elements associated with the donor language community (e.g., σομπρέρο 'sombrero').

Loanblends, referred to as pseudo-loans in Anastasiadis (1994), constitute hybrid formations combining native and borrowed elements. Although they formally resemble lexical items of L2, they are in fact creations of L1 that draw on morphological or lexical resources from L2 (e.g., φρουταρία[1] 'fruit shop', σαπουνέ 'soap-like').

Loanshifts are further divided into semantic borrowings (meaning extensions) and loan translations (calques). In semantic borrowing, the transferred element concerns meaning, which is integrated into the semantic structure of an existing L1 lexical item (e.g., φαλαγγίτης 'phalangite'). In loan translations, the result is a mono- or multi-word unit in L1 that reproduces the structure of a compound lexical unit from L2 (e.g., γαλαζοαίματος 'blue-blood', ψυχροπολεμικός 'of Cold War'). Anastasiadis (1994) classifies both subtypes as indirect or partial borrowings.

From a typological perspective, loanwords fall under material borrowing, whereas loanblends and loanshifts are categorized as structural borrowing (Matras & Sakel, 2007).

3 Language contact and paths of borrowing

Greek and Spanish have never experienced prolonged and intensive contact comparable to that between Greek and Italian or Turkish. Spain did not exercise political or military control over Greek-speaking territories; consequently, no stable conditions of bilingualism emerged that could have facilitated large-scale direct borrowing from Spanish into Greek. Nevertheless, specific historical circumstances and channels of communication introduced a limited number of Spanish-derived lexical items into Greek.

A key historical juncture occurred after 1492, when Sephardic Jews expelled from Spain settled in cities of the Ottoman Empire, particularly in Thessaloniki. Through Judeo-Spanish (Ladino), several lexical items entered Greek, including παρέα 'group of friends', derived from Spanish pareja 'pair,' which developed the meaning 'company, group of friends' in Greek.

[1] This is a typical loanblend found in the speech of Greek Heritage speakers in US and Australia. Loanblends like this are used to fill vocabulary gaps of heritage speakers who find it easier to use stems from the majority language, in which they are generally more proficient, and affixes from their HL, when they produce speech in the heritage language. In situations of contact between English and Greek, since English does not mark grammatical gender, an obligatory feature in Greek, the above combination becomes an efficient vocabulary compensation strategy for overcoming lexical gaps in Greek by assigning grammatical gender to English words through the addition of a Greek affix.

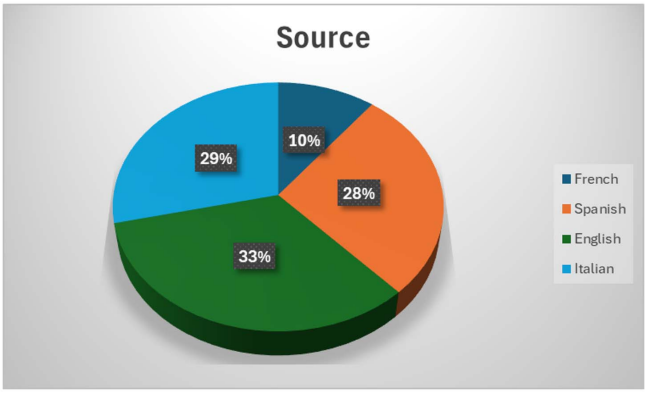


FIGURE 1 Distribution of Spanish loanwords by source/intermediary language

At the same time, Venetian rule (13th–18th centuries) and the strong cultural influence of Italy in the Greek-speaking world established Italian as a primary conduit of borrowing. Spanish-derived elements entered Greek through Italian (e.g., αβίζο 'warning', αγαντάρω 'to secure/to endure', καμαρίλα 'camarilla', καντηλίτσα 'sailor's knot', ντομάτα 'tomato', βανίλια 'vanilla') or Venetian (e.g., γαλέρα 'galley', τσιγάρο 'cigarette', φανφαρόνος 'a show off').

During the eighteenth and nineteenth centuries, French functioned as an intermediary channel for culturally marked borrowings from Spanish into Greek. Terms associated with dance, music, and lifestyle—such as μπολερό 'bolero', ταγκό 'tango', αλπακά 'alpaca', ιδαλγός 'hidalgo', and σοκολάτα 'chocolate'—were transmitted into Greek primarily via French, which operated as an international language of culture and prestige.

In the nineteenth and twentieth centuries, and more markedly in the twenty-first century, Spanish lexical items entered Greek through more direct routes, either via cultural contact with Spain and Latin America or indirectly through English, which now serves as the principal global medium of diffusion. In recent decades, television (Spanish-language soap operas), the internet, and social media have functioned as major dissemination channels. Lexical items related to gastronomy, music, and Latin American culture have entered Greek either directly from different varieties of Spanish (e.g., σάλσα 'salsa', τάκο 'taco', νάτσος 'nachos', μπουρίτο 'burrito', γκουακαμόλε 'guacamole') or through English (e.g., τεκίλα 'tequila', εμπάργκο 'embargo', τζάγκουαρ/ιαγούαρος 'jaguar', κόνδορας 'condor', φλαμίνγκο 'flamingo'), which Greek speakers encounter frequently online. The rapid circulation of memes, songs, and videos on platforms such as YouTube, TikTok, and Instagram has further accelerated the global diffusion of such terms.

As shown in Figure 1, the majority of Spanish borrowings recorded in the DSMG are indirect or mediated (72.4%), whereas direct borrowings from Spanish account for 28%.

In conclusion, Spanish-derived borrowings in Modern Greek are relatively few and dispersed, and their presence is markedly more limited compared to borrowings from Italian, French, English, or Turkish. Their introduction has occurred primarily through intermediary languages, including Italian/Venetian, French, and English. In the contemporary period, however, the expanding influence of the internet and social media has substantially enhanced the penetration of Spanish lexical items into Greek usage, particularly in the domains of popular culture, music, and gastronomy.

4 Linguistic analysis

4.1 Types of borrowings

The majority of items in the dataset derived from the DSMG consist of loanwords (93 out of 105), including a considerable number of internationalisms and foreignisms (e.g., αρένα 'arena', εμπάργκο 'embargo', κακάο 'cacao', κανίβαλος 'cannibal', τεκίλα 'tequila', λάμα 'llama', σομπρέρο 'sombrero', σανγκρία 'sangria', σαβάνα 'savanna'). Loan translations follow in frequency (6 out of 105), for example γαλαζοαίματος 'blue-blood'<Loan translation from Fr. sang-bleu or Eng. blue blood (from Spanish), πολιτοφυλακή 'militia'< Loan translation from Ital. guardia civica or Span. guardia civil, κοκοφοίνικας 'coconut palm'< Loan translation from Eng. coco palm < Span. Coco, ψυχροπολεμικός, 'Cold War-related'< Loan translation from Eng. cold-war- (<Loan translation from Span. guerra fria). These forms entered Greek through intermediary languages such as French, English, or Italian. The dataset also includes five reborrowings (γαλέρα 'galley'< Ven. galera < Span. Galera < Med. Greek γαλαία 'small pirate ship', καμαρίλα 'camarilla'< Ital. camarilla < Span. Camarilla, diminutive of camara < Ancient Greek. Καμάρα 'arch', κόφα 'basket': Medieval. κόφα < Ital. coffa < Span. cofa < Arab quffa < Ancient Greek. Κόφινος 'basket', παλάβρα 'word'< Span. palabra< Lat. parabola < Hellenistic παραβολή, σιλό 'storage tower'< Fr. silo < Span. siro < Lat. sirus 'wheat container' < Ancient Greek Σιρός 'wheat container') and one instance of semantic borrowing (φαλαγκίτης 'phalangite').

Nouns constitute the dominant grammatical category: 80 of the 105 items are nouns, followed by 22 adjectives, one verb (αγαντάρω 'to secure/to endure'), one adverb (διάνα 'bullseye'), and one interjection (ολέ 'ole'). This distribution is consistent with previous research (Myers-Scotton, 2002; Matras, 2007), which identifies nouns as the most borrowable category due to their referential function and relative structural independence. By contrast, verbs and adjectives require deeper integration into morphological and syntactic systems, which constrains their transfer. Adverbs, given their primarily modifying function, are less frequently borrowed and tend to enter the language under specific stylistic or cultural conditions.

More specifically, Van Hout & Muysken (1994, p. 42) observe that "A very important factor involves one of the primary motivations for lexical borrowing, that is, to extend the referential potential of a language. Since reference is established primarily through nouns, these are the elements borrowed most easily." This position highlights the functional role of nouns in linguistic communication: they serve as the principal carriers of reference to persons, objects, places, or abstract concepts. Accordingly, when a language comes into contact with another, speakers tend to adopt nouns in order to designate new realities or to fill lexical gaps.

The limited presence of borrowed verbs and adjectives can be explained by the greater degree of integration these categories require. Verbs, for instance, carry extensive morphological marking and participate in complex syntactic structures, which constrains their transfer across languages (see Myers-Scotton, 2002). Nouns, by contrast, exhibit greater morphological and syntactic autonomy, which facilitates their adoption and functional integration. Adverbs, similarly, are rarely primary targets of lexical borrowing. This pattern reflects their functional properties: unlike nouns, adverbs do not introduce new referential entities into discourse but operate mainly as modifiers or intensifiers. As a result, the need for their borrowing arises less frequently and is typically associated with stylistic choices or with the introduction of items carrying specific cultural connotations.

4.2 Phonological and morphological adaptation

The integration of a borrowed element into the linguistic system of the recipient language is crucial for its stabilization and smooth use. However, the degree of adaptation is not uniform; it varies depending on factors such as the time of borrowing, speakers' familiarity with the source language, and the sociolinguistic attitudes of L1 speakers toward the donor language. (Haspelmath, 2008). The more demanding the integration into the morphological and syntactic system, the lower the likelihood of direct borrowing of form.

Anastasiadis (1994, p. 81) defines the primary criterion for distinguishing between adapted and non-adapted borrowings as the extent to which they diverge from the phonological, morphological, and semantic subsystems of the recipient language. This degree of deviation provides a basis for assessing how fully a borrowing has been assimilated, and thus serves as an indicator of its linguistic acceptance and, ultimately, its integration into the L1 lexicon.

It is important to emphasize that the process of adaptation does not proceed simultaneously across all linguistic levels. Phonological integration often precedes morphological and semantic adjustment, which may follow at different rates. For this reason, the analysis that follows examines sequentially the three principal dimensions of adaptation: phonological, morphological, and semantic.

The phonological compatibility between Greek and Spanish largely accounts for the absence of complex consonant clusters or non-native word-final consonants in Spanish borrowings.

Likewise, almost all lexical borrowings from Spanish are morphologically adapted in Geek, as Spanish gender-marking endings align with corresponding patterns in Greek with feminine nouns ending in -α (e.g., αρένα 'arena', καλντέρα 'caldéra', κάστα 'caste', παστίλια 'lozenge', ρούμπα 'rumba', σαβάνα 'savanna', σανγκρία 'sangria', σοκολάτα 'chocolate') and animate masculine nouns ending in -os (e.g., βακαλάος/μπακαλιάρος 'cod fish', ιδαλγός 'hidalgo', κανίβαλος 'cannibal' , μούτσος 'cabin boy', ραβίνος 'rabbi', τραμπούκος 'bully') occurring most frequently. This smooth integration is facilitated by structural similarities between the two languages, as Spanish, like Greek, exhibits a binary gender system, with prototypical masculine endings in -o (e.g., niño, gato, perro) and feminine endings in -a (e.g., niña, gata, casa, mesa).

The dataset derived from the DSMG also includes 21 indeclinable nouns, predominantly neuter (e.g., αβίζο 'warning', εμπάργκο 'embargo', ζενίθ 'zenith', λάμα 'llama', σιλό 'storage tower', ταλκ 'talc', τανγκό 'tango', ταρτάν 'running track surface', τσα-τσα 'cha-cha-cha', τσίλι 'chilli', φλαμίνγκο 'flamingo'). The assignment of neuter gender constitutes the most common strategy for indeclinable borrowings in Greek, as the neuter functions as the default gender in the language (Anastasiadis, 1994; Tsimpli & Hulk, 2013; Gavriilidou, 2026).

4.3 Semantic categories

For the analysis of the semantic domains to which Spanish-derived borrowings in Greek belong, the study adopts the methodology applied in Gavriilidou (2019) in the investigation of Russian borrowings in Modern Greek. More specifically, it draws on the inventory of semantic fields developed by Haspelmath & Tadmor (2009) for cross-linguistic analysis of borrowing patterns.

TABLE 1 Distribution of Spanish borrowings into Greek across the semantic fields proposed by Haspelmath & Tadmor (2009)

Semantic Categories	Borrowings from Spanish to Greek
The physical world	αληγείς 'alguacil', καλδέρα 'caldera', πάμπα 'pampa', σαβάνα 'savanna'
Kinship	ινφάντης 'infante'
Animals	Βακαλάος/μπακαλιάρος 'cod', ιαγουάρος/τζάγκουαρ 'jaguar', κόνδορας 'condor', λάμα 'llama', τζάγκουαρ 'jaguar', τσιντσιλά 'chinchilla', φλαμίνγκο 'flamingo', μπαρακούντα 'barracuda', αρμαντίλο 'armadillo'
The body	αλμπίνος 'albino', δάγκειος 'dengue'
Food and drink	κακάο 'cacao', καριόκα 'carioca', μανταρίνι 'mandarin', ντομάτα 'tomato', παστίλια 'lozenge', πατάτα 'potato', σαγκρία 'sangria', σοκολάτα 'chocolate', ταπιόκα 'tapioca', τεκίλα 'tequila', τσίλι 'chili', φιδές 'vermicelli', αβοκάντο[a] 'avocado', τάκο 'taco', τορτίγια 'tortilla', μπουρίτο 'burrito', κεσαδίγια ή κεσαντίγια 'quesadilla', εντσιλάδα 'enchilada', γκουακαμόλε 'guacamole', νάτσο 'nacho', παέγια 'paella', τσούρος 'churro', σέρι 'sherry', φαχίτα 'fajita', ταμάλα 'tamale', γκασπάτσο 'gazpacho'
Clothing and grooming	αλπακάς 'alpaca', ζιλέ 'gilet', μερινός 'merino', μπολερό 'bolero', πόντσο 'poncho', σομπρέρο 'sombrero', κρεμεζής 'crimson'
The house	αρένα 'arena', σιλό 'silo', χασιέντα 'hacienda', πάτιο 'patio'
Agriculture and vegetation	κοκοφοίνικας 'coconut palm', μαριχουάνα 'marijuana'
Basic actions and technology	αγαντάρω 'to endure', γαλέρα 'galley', εμπάργκο 'embargo', κάργκο 'cargo', καντηλίτσα 'sailor's knot', καστανιέτα 'castanet', κόφα 'basket', λάσο 'lasso', περγαντί 'pergandhi', πιρόγα 'pirogue', πιστόλα 'pistol', πεσέτα 'peseta', πλατίνα 'platinum', πούρο 'cigar', ρεάλι 'real', ταλκ 'talc', ταμπάκος 'tobacco', ταρτάν 'tartan', τσιγάρο 'cigarette'
Motion	κλότσος 'kick', μάμπο 'mambo', μπολερό 'bolero', πελότα 'pelota', ρούμπα 'rumba', ταγκό 'tango', τσα-τσα 'cha-cha', φλαμέγκο 'flamenco', σάλσα 'salsa', μακαρένα 'macarena', μαριάτσι 'mariachi', πινιάτα 'piñata'
Possession	
Spatial relations	
Quantity	Ζενίθ 'zenith'
Time	
Sense perception	
Speech and language	παλάβρα 'word', μπούρδα 'non-sense'
Social and political relations	αβίζο 'notice', γαλαζοαίματος 'blue-blood', δον 'don', ντον 'don', ιδαλγός 'hidalgo', καμαρίλα 'camarilla', κανίβαλος 'cannibal', κάστα 'caste', μαλαγάνας 'cunning person', μούτσος 'cabin boy', νέγρος 'Black person', παρέα 'company', πολιτοφυλακή 'militia', τραμπούκος 'thug', τσικό 'chico', φαλαγγίτης 'phalangist', φανφαρόνος 'boastful person', χούντα 'junta', μουλάτος μουλάτα 'mixed-race'
Law	προνουνστιαμέντο 'pronunciamiento', γκερίγια 'guerrilla'
Religion and belief	Ραβίνος 'rabbi'

TABLE 1 (*CONTINUED*)

Semantic Categories	Borrowings from Spanish to Greek
Modern world	μουντιάλ 'World cup'
Cognition	διάνα 'bullseye', λουμπάρδα 'bombard', μπομπάρδα 'bombard', ψυχροπολεμικός 'Cold War-related'
Warfare and hunting	
Miscellaneous function words	λαμόγια 'crooks', τόγκα 'toga', ολέ 'olé'

[a]The table includes, in italics, the 105 loanwords recorded in the DSMG, as well as 25 additional loanwords not documented in the dictionary.

Within this framework, borrowability is understood as pertaining to meanings or concepts rather than to individual lexical items. The focus is therefore on the frequency with which a given concept is expressed through borrowing across different languages. To operationalize this approach, the authors employ a fixed list of approximately 1,460 concepts, which are grouped into the following twenty-four semantic domains:

The distribution of Spanish-derived borrowings in Modern Greek, as presented in Table 1, clearly indicates that the majority of items cluster in semantic domains associated with material and cultural life, particularly in the domain of food and drink, which contains a notably high number of borrowings. This finding aligns with the general observations of Haspelmath & Tadmor (2009), according to which culturally specific and specialized vocabularies exhibit higher borrowability than core lexical domains (e.g., kinship, body, time).

At the same time, the presence of borrowings in categories such as motion (primarily dance and musical genres), social and political relations, and actions and technology highlights the role of Spanish as a vehicle of cultural models, artistic practices, and sociopolitical terminology. In contrast, borrowings are scarce in domains associated with core vocabulary (kinship, body, time, quantity), confirming the relative resistance of these conceptual areas to lexical borrowing.

5 Lexicographic practice

The lexicographic treatment of Spanish loanwords in the DSMG highlights several methodological and practical issues related to etymological documentation. In a number of cases, there is uncertainty regarding the origin of a word. Characteristic examples include the loanword δον 'don', where the dictionary points to both Italian and Spanish origin; καριόκα 'carioca', which is accompanied by an etymological reference to the Spanish/Portuguese word carioca; and the loanword κάστα 'caste', which is indicated as having entered Greek via Italian and French. The presence of such ambiguities suggests that the dictionary adopts a cautious approach, recording multiple borrowing pathways where the etymological evidence is not fully conclusive.

At the same time, however, a lack of systematicity is observed in the indication of Spanish origin, which is often omitted in favor of the intermediary language. In several entries, such as βανίλια 'vanilla', αλιγάτορας 'alligator', ροντέο 'rodeo', ράντσο 'ranch', and κάργκο 'cargo', the etymology stops at Italian, French, or English, without indicating that the ultimate source lies in Spanish. This practice creates a distorted picture of the historical origin of the vocabulary.

Furthermore, the DSMG does not include a considerable number of Spanish-derived items that are well established in both spoken and written Modern Greek. Lexical items such as αβοκάντο 'avocado', τάκο 'tacco', τορτίγια 'tortilla', μπουρίτο 'burrito', κεσαδίγια/κεσαντίγια 'quesadilla', εντσιλάδα 'enchilada', γκουακαμόλε 'guacamole', νάτσο 'nachos', παέγια 'paella', τσούρος 'churros', χασιέντα 'hasienda', γκερίγια 'guerilla', σάλσα 'salsa', μαριάτσι 'mariachi', σέρι 'sherry', πινιάτα 'piñata', μπαρακούντα 'barracuda', φαχίτα 'fajita', ολέ 'olé', μουλάτος/μουλάτα 'mulatto', ταμάλα 'tamala', γκασπάτσο 'gazpacho', αρμαντίλο 'armadillo', πάτιο 'patio' are absent, despite their clear integration into contemporary usage[2].

Finally, cases such as φιέστα 'fiesta'—etymologized as originating from Italian, although it has entered most languages via Spanish—and αρμάδα 'armada', attributed to Venetian despite plausible Spanish origin, raise further questions regarding etymological consistency. A more systematic and comprehensive approach could enhance the status of the dictionary as a reliable reference tool and more accurately reflect the extent of Spanish influence on the Modern Greek lexicon.

6 Conclusions

The analysis of Spanish-derived borrowings in Modern Greek highlights their limited number and their concentration in culturally specific and specialized semantic domains, while also demonstrating the distinctive contribution of Spanish to the development of the Modern Greek lexicon through indirect channels of transmission. The findings show that Spanish borrowings in Greek conform to cross-linguistic hierarchies of borrowability, with a clear predominance of nouns and of domains associated with gastronomy, music, clothing, and sociopolitical relations. The examination of their lexicographic treatment in the DSMG reveals both strengths—such as the recognition of alternative etymological pathways—and weaknesses, including lack of systematicity, omissions, and, in some cases, questionable etymological attributions.

Overall, the study underscores the need for a more comprehensive and consistent lexicographic representation of Spanish borrowings, one that more accurately reflects the role of Spanish as a donor language. The systematic inclusion of the 25 items currently absent from the DSMG, along with a clearer distinction between direct and indirect borrowings, would substantially improve the existing lexical inventory and provide users with a more accurate account of the historical development of the Greek lexicon. Given the relatively small dataset, however, broader generalizations about Greek–Spanish language contact should be made with caution.

References

Anastasiadis, A. (1986). Neology in Standard Modern Greek. *Scientific Yearbook of the Faculty of Philosophy,* 65. Aristotle University of Thessaloniki. [In Greek]

Anastasiadis, A. (1994). *Neological borrowing in Modern Greek: Direct borrowings from French and Anglo-American English: A morphophonological analysis*. Aristotle University of Thessaloniki. [In Greek]

Androulidakis, G. (2019). Lexical, translation, and semantic borrowings in Modern Greek from German: Representative examples from a pilot study based on contemporary monolingual dictionaries.

[2] For the purposes of this study, a lexical item was considered integrated into contemporary Greek usage when it occurred repeatedly across independent Greek-language sources, appeared with Greek orthographic adaptation where applicable, and was used without metalinguistic explanation as an ordinary lexical item.

In T. Markopoulos, C. Vlachos & A. Archakis (Eds.), *Proceedings of the 14th International Conference on Greek Linguistics* (pp. 81–90). University of Patras. [In Greek]

Charalampakis, Ch. (1991). Loan translations in Modern Greek from European languages. In A.-F. Christidis (Ed.), *Studies for the Greek language: Proceedings of the 11th Annual Meeting of the Department of Linguistics* (pp. 81–102). [In Greek]

Fliatouras, A. & Kokkini, A. (2017). Semantic borrowings in Modern Greek terminology. In K. Valeontis (Ed.), *Greek language and terminology* (pp. 102–112). Hellenic Society for Terminology. [In Greek]

Gavriilidou, Z. (2019). Lexical borrowing from Russian into Modern Greek. In N. Topintzi, N. Lavidas & M. Moumtzi (Eds.), *Selected Papers on Theoretical and Applied Linguistics from the 23rd ISTAL* (pp. 76–87). https://ejournals.lib.auth.gr/thal/article/view/7322

Gavriilidou, Z. (2018). Russian borrowings in Greek and their presence in two Greek dictionaries. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (Eds.), *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts* (pp. 297–308). Ljubljana University Press. https://euralex.org/publications/russian-borrowings-in-greek-and-their-presence-in-two-greek-dictionaries/

Gavriilidou, Z. (2026). Testing Prototypicality and Default Gender in Language Acquisition of Heritage and L1 Greek Preschoolers. In D. Katsarou & I. Dosi (Eds.), *Optimizing Bilingual Development in Typical and Atypical Children: Language Acquisition, Use, and Interventions* (pp. 133-162). IGI Global Scientific Publishing. https://doi.org/10.4018/979-8-3373-1752-6.ch005.

Haspelmath, M. (2008). Loanword typology: Steps toward a systematic cross-linguistic study of lexical borrowability. In T. Stolz, D. Bakker & R. Salas Palomo. (Eds.), *Aspects of language contact* (pp. 43–62). Mouton de Gruyter.

Haspelmath, M. & Tadmor, U. (2009). *Loanwords in the world's languages*. Mouton de Gruyter.

Haugen, E. (1950). The analysis of linguistic borrowing. *Language*, 26, 210–231.

Humbley, J. (1974). Toward a typology of lexical borrowing. *Cahiers de Lexicologie*, 25(2), 46–70.

Institute of Modern Greek Studies (Manolis Triantafyllidis Foundation) (1998). *Dictionary of Standard Modern Greek*. Aristotle University of Thessaloniki. https://www.greek-language.gr/greekLang/modern_greek/tools/lexica/triantafyllides

Johanson, L. (2002). *Structural factors in Turkic language contact*. Curzon.

Matras, Y. (1998). Utterance modifiers and universals of grammatical borrowing. *Linguistics*, 36, 281–331.

Matras, Y. (2007). The borrowability of structural categories. In Y. Matras & J. Sakel (Eds.), *Grammatical borrowing in cross-linguistic perspective* (pp. 31–75). Mouton de Gruyter.

Matras, Y. & Sakel, J. (Eds.). (2007). *Grammatical borrowing in cross-linguistic perspective*. Mouton de Gruyter.

Mitits, L. & Gavriilidou, Z. (2020). Loanblends in the speech of Greek heritage speakers: A corpus-based lexicological approach. In Z. Gavriilidou, M. Mitsiaki, A. Fliatouras, L. Mitits & S. Kiosses (Eds.), *Proceedings of the XIX EURALEX Congress: Lexicography for Inclusion*, 1 (pp. 351–360). https://euralex2020.gr/wp-content/uploads/2020/11/EURALEX2020_ProceedingsBook-p351-360.pdf

Myers-Scotton, C. (2002). *Language contact: Bilingual encounters and grammatical outcomes*. Oxford University Press.

Papanastasiou, G. (2003). *Borrowing in Modern Greek*. Aristotle University of Thessaloniki. https://www.komvos.edu.gr/glwssa/odigos/thema_a7/thema_pdf.pdf [In Greek]

Petrounias, E. (1995). Loan translations and the etymologies of Modern Greek. *Greek Linguistics*, 95, 791–801.

Tsimpli, I. & Hulk, A. (2013). Grammatical gender and the notion of default: Insights from language acquisition. *Lingua*, 137, 128–144.

Thomason, S. (2001). *Language contact*. Georgetown University Press.

Thomason, S. & Kaufman, T. (1988). *Language Contact, Creolization, and Genetic Linguistics*. University of California Press.

Van Hout, R. & Muysken, P. (1994). Modeling lexical borrowability. *Language Variation and Change*, 6(1), 39–62. https://doi.org/10.1017/S0954394500001575

Vasmanoli, E. (2007). *Loanwords returning to Modern Greek* [Doctoral dissertation, National and Kapodistrian University of Athens.] [In Greek]

Zafeiri, S. (2015). *Loan translations in Modern Greek* [Doctoral dissertation, National and Kapodistrian University of Athens] [In Greek]

Contact information

Zoe Gavriilidou
Democritus University of Thrace
zoegab@otenet.gr

Ivana Lovrić Jović & Lana Hudeček

SYNONYMS THAT NEVER MET

On the Compilation of the Dictionary of the Dubrovnik Idiom

Abstract This article outlines the Dictionary of the Dubrovnik Idiom project at the Institute for the Croatian Language. Given its unconventional conceptual framework, this dictionary adopts a diachronic approach, incorporating headwords from diverse historical periods (spanning from the 16th to the 21st century) and across a broad spectrum of linguistic registers: from the vernacular to high-prestige literature. From the perspective of a fixed point on the diachronic axis (e.g. present-day usage), some synonyms may, depending on their temporal distance, be classified as obsolete or archaic. However, when arranged along both the synchronic and diachronic dimensions, synonym sets can be further differentiated according to functional-stylistic and register-based distribution. The paper presents a typology of the synonymy encountered during the lexicographic processing. Particular attention is paid to cases in which the dictionary records synonyms that never met, that is, synonyms for which corpus evidence shows no temporal overlap.

Keywords Born digital dictionary; Dubrovnik idiom; historical lexicography; synonymy

1 Introduction

The Dubrovnik idiom[1] occupies a special place in the Croatian language, as the literary works of Dubrovnik writers of the 16th and 17th centuries largely laid the foundation for the standardization of Croatian. Today, the Dubrovnik vernacular is less and less used, and can be said to be in decline.

The lexical material of the Dubrovnik vernacular is relatively well represented in the diachronic *Academic Dictionary of the Croatian and Serbian Language* (ADCSL) and in several dictionaries devoted specifically to the Dubrovnik idiom (Mikalja, 1649/51; Bojanić & Trivunac, 2002;[2] annex dictionaries in Lovrić Jović, 2014, 2015; Lovrić Jović & Jozić, 2016, etc.). However, this concerns only a small number of lexicographically processed items, which are difficult to access. In the ADCSL, for example, the Dubrovnik lexicon is embedded within a 23-volume dictionary. Although the dictionary exists in digital form, it is not possible to retrieve this lexicon through simple searches, and there is no unique label indicating Dubrovnik provenance. Besides, these lexicographically processed items may be of limited reliability.

The need to integrate, within a single lexicographic resource, both the lexicon of literary texts that contributed to the standardization of Croatian and the lexicon of the vernacular – documented in genres appropriate to vernacular usage – arises from the fact that no reliable dictionary of this idiom is currently available, while the vernacular itself is undergoing rapid

[1] We use the term idiom as a superordinate term for the concepts of vernacular and language. The term idiom incorporates both the local Dubrovnik vernacular, both in the form of the local speech and in the form of the stylized literary language presented mainly in low-register literature (comedies, congratulatory folk songs, satirical periodicals, etc.), and the literary language written primarily in high-register literature. For a deeper theoretical explanation of this terminology, see Lovrić Jović, 2006, pp. 173–175; 2014, pp. 6-9.

[2] The last corpus-based dictionary related to the Dubrovnik idiom was produced in the 17th century by Jakov Mikalja (1649/51; reprint: Mikalja, 2011). It is considered a precursor to the modern type of dictionary. The only reliable dictionary of the Dubrovnik vernacular (one of whose authors is a linguist) is not corpus-based, but it is still usable for scholarly purposes: Bojanić, M. & Trivunac, R. (2002).

attrition. Recognizing both the scholarly and the broader societal need for a comprehensive dictionary of the Dubrovnik idiom, the Institute for the Croatian Language initiated the dictionary project in 2024 under the leadership of Ivana Lovrić Jović, within which the *Dictionary of the Dubrovnik Idiom* (hereafter: DDI) is being developed (see 2. Dictionary of the Dubrovnik Idiom (in progress)).

Work on the DDI has provided deeper insights into numerous linguistic phenomena and relationships, including phonological, grammatical, morphological, and lexical ones. On the basis of the corpus data, it has also been established that the Dubrovnik dialect contains synonyms that were not used contemporaneously. This particular type of synonymy is discussed in the second part of this article. The first part presents the current state of work on the DDI.

The aim of this article is to draw attention to this specific type of synonymy which, given that it typically involves synonymous pairs or series of synonyms of different origins, also provides evidence that may be of interest to historians. Although this article does not address interpretations extending beyond the linguistic domain, it demonstrates that such synonymous pairs and series are treated with particular care in the DDI, which documents their usage both implicitly, through illustrative examples, and explicitly, in the notes accompanying the dictionary entries (see Table 1).

2 Dictionary of the Dubrovnik Idiom (in progress)

The DDI is being developed as part of a project of the same name funded by the European Union under the NextGenerationEU programme. The project runs from 1 January 2024 to 31 December 2027 under the acronym RDG (*Rječnik dubrovačkoga govora*, 'Dictionary of the Dubrovnik Idiom').[3] It was preceded by an internal institute project phase during which most of the corpus was compiled. The dictionary will be freely accessible online, and the publication of the first 730 dictionary entries is planned for the end of 2027.

The DDI is conceived as a born-digital, corpus-based diachronic dictionary covering the period from the 16th to the late 20th century. It aims to provide a comprehensive account of the lexicon of the city of Dubrovnik over five centuries and across a range of written genres, both the lexicon of high literature and that of the organic vernacular recorded in specific genres. Such a concept has doubled the lexicographic challenges, while at the same time there is no readily identifiable similar dictionary to serve as a model.

It is intended for several user groups: historical linguists and dialectologists; playwrights and actors; and the general public interested in the Dubrovnik vernacular, particularly the inhabitants of Dubrovnik (Lovrić Jović & Kramarić, 2024), especially those concerned about the possible disappearance of the Dubrovnik vernacular. Not only the vocabular, but even the phonological, morphological, syntactical and prosodic features that rendered it authentic are nearly extinct. The geopolitical situation in the Dubrovnik area has led to demographic changes that are not favourable to the Dubrovnik idiom, so it is now heard very rarely in Dubrovnik. The DDI therefore provides the opportunity to document the idiom while it is still spoken to some extent. (Lovrić Jović & Kramarić, 2024, p. 802).

It is, of course, unrealistic to expect that this gap can be filled within a four-year project involving only a limited number of collaborators (five). This is a complex lexicographic undertaking, a significant portion of which is devoted to establishing sustainable foundations (the corpus, guidelines, and technological infrastructure). Following this phase, the project is intended

[3] More information about the dictionary can be found on the project website: https://rdg.jezik.hr/

to evolve as an expanding dictionary, either within a new externally funded project or as a permanent project of the Institute for the Croatian Language. A more extensive DDI, based on a corpus and adhering to rigorous lexicographic principles, can realistically be expected in approximately ten years.

It is important to note that a key innovation of the DDI lies in the fact that no dialectal e-dictionary currently exists in Croatia. Given that dialect lexicography is largely left to amateur efforts – at best involving contributors from the language sciences (not necessarily lexicographers, lexicologists, or philologists in the strict sense) – collaboration with IT specialists is often difficult, not least due to financial constraints.

Printed dictionaries are frequently compiled by speakers who emigrated decades ago (which can be advantageous, as they tend to preserve older linguistic forms more faithfully) and are published with the support of local cultural associations or independently. Among these, high-quality glossaries occasionally appear, for example Veselić (2020), Radišić (1999), Panža (2015).

The DDI, therefore, cannot be modelled on any existing dialectal e-dictionary, particularly as it is conceptually pioneering. However, various online resources do provide wordlists of vernacular speech compiled by speakers themselves. These typically involve basic processing, where meanings are defined through standard-language synonyms or brief descriptions (when synonym-based definition is not possible).

E-glossaries of older Dubrovnik literature are somewhat more reliable. These are usually produced under the auspices of cultural institutions, so even if the author is not identified, it can be assumed that they are based on existing annex dictionaries from editions of canonical literary works. (See: https://muzej-marindrzic.eu/marin-drzic-dundo-maroje/marin-drzic-dundo-maroje-rjecnik/).

2.1 Corpus

The DDI is a corpus-based dictionary. For the purposes of the project, a newly built historical corpus was compiled, covering several centuries. It integrates a wide range of genres and styles, including administrative and legal documents, personal letters, folk songs, memoirs, and literary works – from Držić's comedies to 21st-century satirical journalism and traditional cookbooks.

To ensure a comprehensive diachronic overview, the project also draws on classic dictionaries (Mikalja, ADCLS). The team uses the corpus tool Sketch Engine to digitize and process manuscripts and printed texts into a consistent, searchable format. This **comprehensive** corpus highlights the evolution of the Dubrovnik lexicon, from high-register literature to colloquial vernacular. It offers an all-encompassing and eminently appropriate showcase of the Dubrovnik idiom for this purpose.

Since the vernacular, as an authentic living expression, cannot be directly accessed for earlier periods, information on it can only be obtained indirectly from those parts of written texts in which the Dubrovnik vernacular is most clearly reflected (dramatic works, occasional poetry, transcribed handwritten testimonies), as well as from genres that effectively convey spoken language, such as epistolary prose. For more recent periods, the lexicon will be excerpted from audio recordings of speakers of the Dubrovnik vernacular, among the most valuable of whom are emigrants to other continents, who best preserve the speech of the second half of the 20th century.

The DDI will also include a detailed list of corpus components, and each example in a dictionary entry will be accompanied by a precise indication of its source.

2.2 Wordlist

The wordlist for the current phase of the project contains 730 headwords. The compiled corpus contains material for several thousand lexicographic entries, according to the current editorial principles, which exclude from the dictionary words that are identical in both the Dubrovnik dialect and the standard language.

The main principle for selecting dictionary headwords, based on the lexicon provided by the corpus, is their distinctiveness in relation to the standard language. This difference must be lexical or semantic; phonological and morphological differences alone are not sufficient. If the dictionary were to include headwords that differ only in phonological features – such as local *pako* for standard *pakao* ('hell'), local zo for standard *zao* ('evil'), or local *viđet* for standard *vidjeti* ('to see') – processing the entire Dubrovnik lexicon would take several decades. In such a dictionary, as headwords, all words from the standard language whose forms reflect some phonological feature of the Dubrovnik idiom would have to be included. This would encompass all standard-language lexemes ending in *-nt*, *-st*, *-zd*, all infinitive forms, and all lexemes whose form allows for: realized ijekavian jotation, various sound changes in all parts of the word (reduction of initial consonant clusters, *lj* > *j* reduction, final -m > -n (Adriatic feature), contraction of *-ao/-eo* > *-o*, secondary *h, hv* > *f*, etc.).

If a word is attested in both the Dubrovnik vernacular and the standard language but with different meanings, the Dubrovnik variant is treated as not belonging to the standard language and is included in the DDI as a headword. In such cases, only its specific local meaning is defined, since the standard meaning is assumed (speakers of the Dubrovnik vernacular are proficient in standard Croatian, and a typical Dubrovnik sentence also includes words that belong to the standard language). The same principle applies to words attested only in the pre-standard period.

An important feature of the Dubrovnik idiom, also reflected in the wordlist, is the strong historical influence of the Italian language.[4] This influence affected the local vernacular, resulting in the incorporation of many Italianisms, which will therefore feature predominantly in this dictionary. This does not mean that a typical Dubrovnik sentence – regardless of time, style, or genre – contains more Italianisms than Slavic words; rather, since this is a contrastive dictionary, Italianisms are more salient. It is well known that distinctive Dubrovnik vocabulary is more easily identified among loanwords than among items of Slavic origin (Rešetar, 1952, p. 99).

2.3 Structure of the Dictionary Entry

The structure of the dictionary entry in the DDI is discussed in detail in Lovrić Jović & Kramarić (2024). Here we will list only the basic components of a dictionary entry: headword,

[4] As for the Italian dialect that influenced the Dubrovnik idiom, according to Muljačić (1962, p. 341), the Venetian language initially had the strongest influence, but was later replaced by Tuscan, which Dubrovnik inhabitants learned from books or while in Italy. The local population was less familiar with Tuscan. The use of Venetian declined among merchants and sailors, while Italian replaced Venetian as the cultural language learned at Italian universities and used by merchants.

grammatical block, variant (phonological or derivational, see section 3.1), meaning (including the stylistic/professional label, grammatical restrictions, definition of meaning, examples, internal links connecting synonyms and antonyms), etymology, notes, and subentry. Etymology, notes, and also a subentry may also be linked to the lemma when they apply to all the meanings of the word. Meanings are defined through standard-language synonyms or brief descriptions (when synonym-based definition is not possible).

3 Synonymy in the DDI

The lexicographic concept, described in detail in Lovrić Jović & Kramarić (2024), has introduced additional challenges, while also offering new insights into the Dubrovnik vernacular. One such insight, discussed in this paper, concerns synonymy, particularly synonyms that do not co-occur in actual usage.

Synonyms in the DDI are presented according to the following basic principles:[5]

1. Synonyms are listed under the relevant sense of a word within a dictionary entry ("Colloquially, we speak of the synonym or antonym (etc.) for a 'word', but it is usually more accurate to speak of sense relations. Since different senses of a single word enter into relations with different sets of other words." (Murphy, 2016, p. 440).

2. Synonyms are cross-referenced bidirectionally if both are included in the DDI and share the same definition of meaning; differences in usage are indicated either by stylistic labels or recorded in a note.

3. Synonyms that are not currently included in the DDI – that is, those not processed in the first phase of dictionary development are entered in a separate section for so-called "dead synonyms" (synonyms that are not cross-referenced because they have not yet been processed as headwords). They will be processed in later phases of the project (the principle is that all synonyms should be searchable in the dictionary – at least in the long term).

4. Standard-language (contact) synonyms are included in the definition: whenever possible, meanings are defined using a standard-language synonym; for example, the headword *imbrolj* ('intrigue, scheme') has the synonyms *intrigarija* and *kabala*, while the semantically equivalent standard-language word *spletka* is used as the definition of meaning.

5. Phonological variants are not listed as synonyms and the decision whether a derivational variant is treated as a synonym or as a variant depends on the lexicographer's judgement of whether an average user would perceive the lexeme and its variant as one or two separate lexical items.

3.1 Synonyms and Lexeme Variants

The structure of the DDI includes two sections in each dictionary entry that are relevant to the discussion in this paper: the section listing synonyms and the section presenting variants of the headword lexeme. The decision to include both sections in the dictionary has required a clear

[5] In dictionaries compiled or being compiled at the Institute for the Croatian Language, synonyms are regularly included. This is the case in the *School Dictionary of the Croatian Language* and in the *Croatian Web Dictionary – Mrežnik* (demo version A–NJ). In *Mrežnik*, other lexical units that enter into semantic relations with the headword are also included (antonyms, hyponyms, and co-hyponyms—both of which form a closed and relatively small set). A meronymic relation is also implicitly recorded.

delimitation of which lexemes are assigned to each. Variants are generally considered phonological forms (often arising from different adaptations of loanwords into Croatian), and less frequently derivational variants.

Words listed as synonyms have their own dictionary entries and are fully processed. Words listed as variants also have their own dictionary entries, from which they are cross-referenced to the most well-attested variant using the cross-reference marker *v.* 'see'. In the dictionary entry of the most established variant, the other variants are listed in the "variants" section. For the treatment of synonyms, the principle underlying this arrangement is quite clear: synonyms are distinct words. However, one may ask why variants, which are forms of the same word, have their own dictionary entries. The reason is that the intended users of the dictionary may sometimes need considerable linguistic knowledge to recognise with certainty that a word they have encountered is a variant of another word. Since the DDI is not intended solely for linguists, we opted for a user-friendly approach in which variants also have their own dictionary entries, referring users to the headword under which the complete treatment is provided.

For example, the word *rožbina* ('relatives') has the variants *rodžbina* and *roďbina*. The dictionary entry for the headword *rožbina* includes the following components: the headword, grammatical information, variants, definition, examples with their sources, synonyms, and a note (see Table 1; only one example from this dictionary entry is presented in the table). The dictionary entry for a variant includes only grammatical information, a cross-reference *v.* ('see'), and an example with its source (see Table 2).

TABLE 1 Dictionary entry of the headword *rožbina*[a] ('relatives')

Headword	**rožbina**
Grammatical block	im. ž. (G jd. rožbine, G mn. rožbina)
Variants	inačice: rodžbina, roďbina
Definition (standard language synonym)	**rodbina**
Example + example source	*Zahvale Bogu, nahodim se da imam dobara koliko mi je dosta za moć izabrat jednu ženu koja će imat sve po meni i koja u onemu što mi bude podložna i poslušna neće moć prikarat me ni nje prćijom, ni nje rožbinom.* Dubrovačke frančezarije, 18. st.
Synonyms	SINONIMI: parentela :1, parentijera :1, parentat :1
Note (translated)	The elements of the group of synonymous dictionary entries (*parentat*, *parentijera*, parentela, and *rodžbina/rožbina* ('relatives, kinship, family')) are diachronically distributed as follows: the Italianism *parentat* (Italian *parentato*) is attested in texts and in dictionaries that record the vernacular of the twentieth and twenty-first centuries. The Italianism *parentijera* is attested in the corpus in dramatic texts from the sixteenth and seventeenth centuries. The Italianism parentela (Italian *parentela*) is found in the eighteenth century, while the two phonological variants of the Slavic word, *rodžbina* and *rožbina*, are attested in the high literary language of the sixteenth century (*rodžbina*) and in the dramatic language of the eighteenth century (*rožbina*).

[a]Among the variants of the headword *rožbina*, no major differences should be sought, since the writing system in the period to which these attestations belong was inconsistent; that is, there was no standardized orthography, particularly for the sounds *dž* and *đ*. It should therefore be borne in mind that these attestations were transcribed by editors of modern editions who did not mutually harmonize their principles of transcription

TABLE 2 Dictionary entry of the variant *rođbina* ('relatives')

Headword	**rođbina**
Grammatical block	im. ž. (G jd. rođbine, G mn. rođbina)
Reference *see* (v.)	*v.* rožbina
Example + example source	*Među ostalim molim rođbinu moju da me se spomenu [u] molitvi.* Oporuka 60, Državni arhiv u Dubrovniku, 17. stoljeće

The variant section in DDI often includes derivational variants. With this decision, we partly align ourselves with linguists working in lexical semantics who define synonymy as *"a relation between distinct lexemes with different roots referring to the same extralinguistic entity."* (Which is a refined definition of the basic and simplest form of synonymy as *"a relation between distinct lexemes referring to the same extralinguistic entity"* (Bussmann, 1996, p. 1164)). In this way, synonyms that do not meet this criterion are excluded from their lexical semantics analysis (Petrović, 2005; Pišković, 2024), and a distinction is made between "true" (lexical) synonymy and phonological, morphological, and derivational variation.

However, although phonological synonyms are consistently excluded from the synonym section and placed instead in the variant section, derivational synonyms may appear in both sections. While this may be seen as inconsistency, it is a justified – indeed, deliberate – one. This is explained by the fact that the dictionary is always oriented toward its user, that is, the person for whom it is written. In the case of potential derivational variants, the decision on whether they will be treated in the DDI as synonyms or variants depends on a purely practical consideration: the lexicographer's judgement of whether the user will readily recognize them as variants. If the lexicographer judges that such recognition is not straightforward, the word is included in the synonym section and treated as a separate headword.

For example, the case of *nova* and *novitat* ('news') is theoretically no different from that of *pacijencija* and *pacijenca* ('patience'). In both pairs, the sources of the lexemes are foreign (Italian and Latin, respectively), and the words are similar in form and semantically identical. Nevertheless, in the dictionary *nova* and *novitat* are linked as synonyms, while *pacijenca/ pacijencija* are treated as variants of a single lexeme. Such decisions are guided by the needs and capabilities of the general user, not only those of specialist philologists. For the same reason, *doglas* and *doglasak*, meaning 'news', are treated as separate headwords linked as synonyms, even though they are derivational variants. A philologist user might recognize that *doglasak* is a derivational variant of *doglas*, but for an actor, director, dramaturge, or speaker, they function as two distinct lexical items (which is not the case with phonological variants such as *festanjuo/ festanjul* 'traditional honorary title in Dubrovnik').

3.2 Typology of Synonyms in DDI

Without entering into a discussion of whether true synonyms actually exist in language at all (i.e. words of identical meaning interchangeable in all contexts of use), or whether synonyms, even when denoting the same extralinguistic entity, nevertheless necessarily differ at least minimally in the contexts in which they are used, in their connotative meaning, etc., the material collected for the DDI allows us to distinguish two groups of synonyms:

1. Synonyms that most closely approximate the notion of true/absolute synonyms, that is, they coincide to the greatest extent in all their dimensions – denotative, connotative,

specialized, and social, denoting essentially the same concept from the same perspective of conceptualization (Pišković, 2024, p. 442).

2. Synonyms denoting the same concept from different perspectives of conceptualization (near synonyms (de Bessé, 1997, p. 142), i.e. cognitive synonyms).

The first group includes:

- synonyms consisting of a Dubrovnik lexical item/items and a loanword/loanwords from Italian, coexisting within the same period, not differing in usage, and carrying the same connotative value, e.g. *ardimen(a)t – smionstvo* ('boldness') or *koradž – hrabrenstvo* ('courage').

- some synonyms (excluding partial synonyms) connected in corpus examples by the conjunction *ili* ('or') (Dubrovnik forms of this conjunction also include *iliti* and *aliti*), e.g. *ognjište aliti komin* ('a place where cooking is done', today: 'kitchen'; Lovrić Jović, 2015, p. 118), *popopiti iliti fratar obući* ('to be ordained'; Lovrić Jović, 2015, p. 165).

- synonyms occurring within the same historical period and stylistic register for which, at least according to corpus attestations, no difference in connotative meaning can be observed. Certain lexemes remain in a synonymous relationship over long periods. For example, *djevojka* and *godišnica* (in the meaning 'maidservant') have been present in the Dubrovnik idiom from the fifteenth/sixteenth century to the present day and, unlike the pejoratives *godulja* and *goduljina*, which also denote a maidservant, they carry neutral meaning and occur in texts of various types. Many such synonyms are found among the lexemes considered variants in the DDI, e.g. *pokujstvo* and *pokuće*, attested in eighteenth-century wills.

The second group includes:

- synonyms belonging to the same period but different linguistic styles.
  For example, a series of lexemes from the eighteenth-century Dubrovnik idiom are defined in the dictionary by the standard-language expressions *lopov* ('thief') and *hulja* ('scoundrel'): *lupež*, *lupešković*, *manigodo*, *farfanat*, *birba*, *galijot*. In the same century, the synonyms *čevrljonica*, *jezičnica, govoruša, govorušina, razložnica, labara*, and *labarina* are attested, which are defined in the dictionary by the standard-language expression *brbljavica* ('chatterbox'). All the listed lexical items occur in dramatic texts known as *frančezarije*. These are translations/adaptations of comedies by the seventeenth-century French playwright Jean-Baptiste Poquelin Molière into a stylized eighteenth-century Dubrovnik idiom (Lovrić Jović, 2014, p. 242). This suggests that these synonyms serve to characterize the speech of individual dramatic characters.

- synonyms belonging to different periods, referred to in the title of this paper as *synonyms that never met*.
  Such synonyms are cognitive synonyms for which corpus evidence shows no temporal overlap. The dictionary does not employ an explicit archaism label; instead, temporal distribution is conveyed implicitly through dated usage examples. Further observations are provided in editorial notes forming an integral part of the dictionary entry (see the note in Table 1). This type of synonym will be discussed in more detail in the following section.

3.3 Synonyms that never met

From the perspective of a fixed point on the diachronic axis (e.g. present-day usage), some synonyms may, depending on their temporal distance, be classified as obsolete or archaic.

However, when arranged along both the synchronic and diachronic dimensions, synonym sets are further differentiated according to functional-stylistic and register-based distribution. The DDI does not require any effort on the part of the user to identify such synonyms. The notes accompanying these entries explicitly explain that the synonyms "never met". This is also partly evident from the illustrative examples. As a general principle, one example is provided for each century in which a lexeme is attested. Additional examples are included only when they demonstrate usage in different types of texts (e.g. literary, administrative, or scholarly) or illustrate particular morphological or syntactic features. The gaps themselves in the illustrative examples section of the dictionary entry allow users to infer that the lexeme was used only during a particular historical period. Moreover, the note accompanying a lexeme that never coexisted with its synonym(s) specifies the period during which each of the synonymous lexemes was in use. Since the dictionary is strictly corpus-based, temporal delimitation must take into account corpus representativeness and generational overlap. Nevertheless, clear tendencies emerge. Here we provide several examples.

(1) One illustrative example concerns lexemes denoting 'a small glass for liqueur or brandy' in the Dubrovnik vernacular: *bićerin, bićerić, kupica*, and *žmulić*. Corpus evidence reveals a clear diachronic distribution. The Venetianism *bićerin* (Ven. *biccerin*) is attested only in twentieth-century sources, whereas the synonymous *bićerić*, a Venetian loan formed with a Croatian diminutive suffix, occurs exclusively in seventeenth- and eighteenth-century texts. The Italianism *kupica* (It. *coppa*) is attested in nineteenth- and twentieth-century sources, but not earlier, while the near-synonym *žmulić* ('small glass') is rare but attested over a wide time span, from the sixteenth to the twentieth century.

(2) From the synonymous set of lexemes *baraška, barufa*, and *litij*, meaning 'quarrel' or 'dispute', the Venetian borrowing *barufa* (Venetian *barufa*) appears in the corpus in the eighteenth and twentieth centuries, whereas the synonymous *litij* (Italian *lite, litigio*) occurs rarely, at the beginning of the eighteenth century (in a will from 1713). The synonym *baraška* (Venetian *borasca*, Italian *burrasca*) is attested in its figurative meaning (its primary meaning being 'storm'), namely 'a turbulent social situation, quarrel', in a text reflecting traces of the colloquial idiom of the twentieth century. This means that the word *barufa* was in use simultaneously with its synonyms *baraška* (twentieth century) and *litij* (eighteenth century), but the corpus contains no attestations of the words *litij* and *baraška* being used within the same historical period.

(3) The synonyms *doglasak*, *nova, novitat*, and *vjestica* mean 'news' and occur in clearly defined periods. The lexeme *doglasak* is attested several times in an eighteenth-century dramatic text; the word *nova* occurs in a sixteenth-century dramatic text as well as in an eighteenth-century one (the same text in which the synonymous *doglasak* is attested), whereas the word *novitat* (Lat. *novitas*, gen. sg. *novitatis*; Italian *novità*) is first attested in the eighteenth century and appears with high frequency in various genres up to the very end of the period covered, that is, into the twenty-first century (in texts published in the twenty-first century that represent the idiom of the twentieth century).

(4) The synonyms *skerac, maskara*, and *berta* occur with the meaning 'joke, mockery' and are distributed as follows. The Italianism *skerac* (Italian *scherzo*) is attested only in the twentieth century, whereas the Turkism *maskara* is attested only in the eighteenth century. The Italianism *berta* (Italian *berta*), frequent in the expression *kalat bertu* ('to make a joke'), is found in Dubrovnik dramatic literature up to the nineteenth century. We may therefore say that the lexeme *skerac* never encountered the synonymous lexemes *maskara* and *berta*.

(5) The following pair of synonyms is an example of Italian–Croatian language contact. The Italianism *beleca*, meaning 'beauty', is attested in the twentieth century (and can still be heard in the present-day vernacular) and is recorded in a dictionary documenting the

twentieth-century Dubrovnik idiom (cf. Bojanić & Trivunac, 2001). The lexeme *ljepos(t)*, meaning 'beauty', is of Slavic origin and appears in high literary language in the sixteenth and seventeenth centuries in the works of writers and lexicographers who documented their language (Dinko Ranjina, Ivan Gundulić, and Bartol Kašić). *Beleca* and *ljepos(t)* never encountered one another.

4 Conclusion

The digital, corpus-based *Dictionary of the Dubrovnik Idiom*, currently being compiled at the Institute of the Croatian Language, is conceived as a comprehensive account of the lexicon of the city of Dubrovnik over five centuries and across a range of written genres. Lexicographic concept of the DDI, described in detail in Lovrić Jović & Kramarić (2024), has given rise to additional challenges while also offering new insights into the Dubrovnik vernacular. One such insight, discussed in this paper, concerns synonymy across time and registers, specifically synonyms that do not co-occur in actual usage.

This paper presents a typology of synonyms occurring in the DDI, which are divided into two basic groups: synonyms belonging to the same historical period and least susceptible to extralinguistic influence (those that most closely approximate the notion of true/absolute synonyms), and synonyms denoting the same concept from different perspectives of conceptualization (near synonyms (de Bessé, 1997, p. 142), i.e. cognitive synonyms). Both groups are further divided in subgroups where special attention is paid to synonyms that do not co-occur in actual usage.

It is particularly interesting—not only for linguists but also for historians, scholars of cultural history, and sociologists—that in certain periods a word of Slavic origin is in use, while in others a word of Italian origin is preferred. The paper therefore, in addition to addressing synonymy in the DDI, indirectly indicates that data provided in dictionaries can also be used as an important source of information about the cultural history of speakers of the idiom they document.

During the first phase of the project, collaboration with scholars from other disciplines—primarily historians—was not envisaged. Consequently, the present paper offers only a preliminary insight into an issue that warrants more thorough investigation in future research. Recognizing the importance of documenting information that extends beyond the linguistic domain, the notes accompanying the dictionary entries for synonyms that "have never met" (see an example of such a note in Table 1) include data that we consider potentially relevant for future interdisciplinary research.

References

ADCSL = *Rječnik hrvatskoga ili srpskoga jezika* [Academic Dictionary of the Croatian or Serbian language] (1880–1976) I – XXIII. JAZU.

Bojanić, M. & Trivunac, R. (2002). *Rječnik dubrovačkog govora. Srpski dijalektološki zbornik XLIX*. SANU.

Bussmann, H., Kazzazi, K. & Trauth, G. (1996). *Routledge Dictionary of Language and Linguistics*. Routledge.

Lovrić Jović, I. (2014). *O starome dubrovačkom govoru nazbilj. Jezična analiza dubrovačkih frančezarija*. Institut za hrvatski jezik i jezikoslovlje.

Lovrić Jović, I. (2015). *Ja, Krsto Lučin Dubrovčanin, činim sad ovi testamenat. Jezična analiza dubrovačkih oporuka iz 17. i 18. stoljeća s transkripcijom i aneksnim rječnikom*. Institut za hrvatski jezik i jezikoslovlje.

Lovrić Jović, I. & Jozić, Ž. (2016). *Libro od svetoga Vlaha. Jezična studija s rječnikom*. Institut za hrvatski jezik i jezikoslovlje.

Lovrić Jović, I. & Kramarić, M. (2024). The Dubrovnik Idiom Through Time – Crafting a Diachronic Dictionary. In K. Štrkalj Despot & A. Ostroški Anić & I. Brač (Eds.), *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress* (pp. 791–805). Institut za hrvatski jezik.

Mikalja, J. (2012). *Blago jezika slovinskoga (1649./1651.): transkripcija i leksikografska interpretacija*. Institut za hrvatski jezik i jezikoslovlje.

Muljačič, Ž. (1962). Dalmatski elementi u mletački pisanim dubrovačkim dokumentima 14. stoljeća. Prilog raguzejskoj dijakronoj fonologiji i dalmatsko-mletačkoj konvergenciji. *Rad JAZU*, 327, 237–380.

Murphy, M. L. (2016). Meaning Relations in Dictionaries: Hyponymy, Meronymy, Synonymy, and Contrast. In P. Durkin (Ed.), *The Oxford Handbook of Lexicography* (pp. 439–457). Oxford University Press.

Panža, P. M. (2015). *Rječnik govora Blata na Korčuli*. Institut za hrvatski jezik.

Petrović, B. (2005). *Sinonimija i sinonimičnost u hrvatskom jeziku*. Hrvatska sveučilišna naklada.

Pišković, T. (2024). *Uvod u leksičku semantiku hrvatskoga jezika*. Matica hrvatska.

Radišić, T. (1999). *Ričnik splitskoga govora*. Mala splitska biblioteka.

Rešetar, M. (1952). *Najstarija dubrovačka proza*. SANU.

Veselić, K. (2020). *Blago konavoskijeh riječi*. Matica hrvatska Konavle.

Contact information

Ivana Lovrić Jović
Institute for the Croatian Language
Ilovric@ihjj.hr

Lana Hudeček
Institute for the Croatian Language
Ilovric@ihjj.hr

Acknowledgement

This article was produced within the framework of the project “The Dictionary of the Dubrovnik Idiom”, funded by the European Union – NextGenerationEU. The views and opinions expressed are solely those of the authors, and these views do not necessarily reflect the official positions of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

Nika Stefan, Johan van der Zwaag & Hindrik Sijens

BEYOND THE SINGLE STANDARD

Addressing Variation and User Needs in an Online Frisian Dictionary

Abstract This paper examines the challenges of representing linguistic variation in Frisian lexicography, focusing on the Dutch-Frisian online dictionary. While existing lexicographic resources already accommodate variation to some extent, discrepancies remain between dictionary representation and actual language use. Survey data and usage examples reveal complex patterns in which standard forms coexist with contact-induced variants, and in some cases no fully equivalent Frisian alternatives exist, leading speakers to adopt Dutch forms in Frisian discourse. These findings highlight a tension between normative guidance and descriptive adequacy, particularly in contexts where speakers vary in competence and where written Frisian differs from spoken usage. The current label system, although informative, does not always provide sufficiently systematic or user-oriented support. Building on insights from user-oriented lexicographic theory, this paper proposes a more flexible model in which dictionary information is accessed through multiple modes. These include a normative mode offering clear guidance, a comparative mode making variation explicit, and a learning mode tailored to users with less experience. By structuring variation in a transparent and user-sensitive way, the proposed approach aims to align lexicographic practice more closely with actual language use while supporting both standardization and linguistic diversity.

Keywords Labelling; online dictionaries; minority language; language variation; standardization; description/prescription; lexicographic practice

1 Introduction

Frisian lexicography is characterized by a persistent tension between linguistic variation and normative standardization. On the one hand, Standard Frisian is codified in grammars, orthographic conventions, and dictionaries, and serves as the primary model for written language as well as, to a certain extent, formal and public communication. On the other hand, spoken Frisian exhibits considerable variation, shaped by regional dialects, individual speaker preferences, and sustained contact with Dutch. This divergence between codified norms and actual language use creates challenges for both language users and lexicographers.

Speakers of Frisian differ widely in their competence (Provinsje Fryslân, 2020; Dekker et al., 2025), confidence, and attitudes toward linguistic variation (Stefan, 2022). For many, the standard language provides an important point of reference, particularly in writing and educational contexts. At the same time, the forms encountered in everyday communication often deviate from this standard, incorporating dialectal features and contact-induced forms (Sjölin, 1976; Breuker, 2001; Merkuur, 2021). As a result, speakers may experience uncertainty about what constitutes acceptable or appropriate Frisian, especially in contexts that require more formal or public language use. This uncertainty is further reinforced by differences in the acceptance of linguistic variants: forms that are widely used are not always perceived as acceptable, and vice versa. Previous research in Frisian lexicography has already highlighted the importance of this tension. In particular, the competition between Frisian-origin words (including purist formations) and Dutch loanwords has been identified as a central issue (Duijff & Van der Kuip, 2018). Lexicographic practice has long attempted to address these challenges, for instance in works such as the Frysk Hânwurdboek 'Frisian Concise Dictionary' (Duijff et al., 2008), which already incorporated mechanisms to guide users in their lexical choices. Studies based on sociolinguistic data also show that speakers' linguistic behaviour is far from uniform. Variation

occurs not only in terms of frequency of use, but also in patterns of acceptance, age-related differences, and context-dependent language use (Klinkenberg et al., 2018; Stefan, 2022). More recent survey data complement these findings by providing additional insight into speakers' awareness of variation and their self-reported language behaviour across contexts (Dekker et al., 2025).

Taken together, these findings point to a highly complex linguistic landscape in which different dimensions of variation – frequency, acceptability, and usage – interact but do not necessarily align. This complexity poses a fundamental challenge for lexicography. Dictionaries are expected to provide clear guidance and support speakers in their linguistic choices, yet they must also account for the variability inherent in actual language use. As has been argued in lexicographic theory, this requires a careful balance between descriptive adequacy and prescriptive or proscriptive guidance (Bergenholtz & Gouws, 2010). From a user-oriented perspective, this tension is further complicated by the diversity of dictionary users and use situations. Modern lexicographic theory emphasizes that dictionaries should be designed with specific user needs in mind, taking into account different levels of competence, purposes, and contexts of use (Tarp, 2008; Lew, 2011). In practice, however, many dictionaries still implicitly assume a relatively homogeneous user group and offer a single, unified presentation of lexical information.

The increasing shift toward digital lexicography opens up new possibilities for addressing these challenges (Lew & De Schryver, 2014). Unlike printed dictionaries, digital resources are not constrained by the need to present a fixed and uniform set of entries. Instead, they can offer flexible access to lexical information, adapt to different user needs, and provide multiple layers of guidance (Atkins & Rundell, 2008; Fuertes-Olivera, 2025). Research on electronic lexicography and dictionary use has shown that users interact with online dictionaries in diverse and often non-linear ways, selecting information according to their immediate needs rather than following a predefined structure (Lew, 2015). The existing Dutch-Frisian online dictionary (Online Nederlânsk-Frysk Wurdboek, ONFW) already incorporates some of these possibilities, for instance through a separately selectable preferred-vocabulary list that provides normative guidance. However, this setup remains largely oriented toward a single standard and does not fully capture the complexity of spoken Frisian variation or the diversity of user needs and expectations.

This paper examines how the tension between linguistic variation and normative standardization manifests itself in contemporary Frisian lexicography and explores how it can be addressed in an online lexicographic environment. Building on empirical data on language use, acceptance, and variation, we analyse the limitations of current approaches and propose a more flexible model that accommodates different user profiles and communicative contexts. In particular, we argue for the development of multiple dictionary modes, supported by a refined system of labels, which together allow for a more transparent representation of variation while maintaining a coherent normative framework. The following section therefore examines the current structure and design of the ONFW in more detail, focusing on how variation and normative guidance are presently represented. This serves as a basis for the subsequent discussion of how lexicographic practice can be adapted to better account for linguistic variation and diverse user needs.

2 The current lexicographic setup

The most recent dictionary is the Dutch-Frisian online dictionary (ONFW), which is part of the website Frysker. This is a translation dictionary primarily intended to support language production. In principle, Dutch lemmas and fixed collocations are translated into standard Frisian. The target audience for this dictionary comprises both active users who wish to translate Dutch texts into Frisian and passive users who wish to learn more about the Frisian language.

In addition to Dutch speakers, the dictionary is also widely used by speakers of Frisian themselves, as their competence in Dutch typically exceeds their command of written Frisian (Duijff & Van der Kuip, 2018). To guide users in selecting appropriate equivalents, the dictionary provides various types of additional information. In many cases, labels indicating usage restrictions are attached to translation equivalents. These labels allow users, for instance, to determine whether a form is a loanword, a colloquial variant, or an obsolete expression. In addition, usage explanations may be provided at different levels of the microstructure.

The present discussion focuses on sociolinguistic labels, which are intended to guide users in situations where multiple translation options are available, including forms that fall outside the standard language. The label system can be broadly divided into five categories: language variety, frequency, diachronic status, medium, and origin (see Table 1). Together, these categories provide users with cues about how to interpret and select between competing variants.

TABLE 1 Categories and labels

Category	Labels
Language variety	Clay Frisian, Woodland Frisian, North Quarter dialect, Southwest Quarter dialect (the four main dialects of Frisian)
Frequency	rarely used; on occasion; especially
Diachronic status	obsolete, becoming obsolete; archaism; neologism
Medium	(mainly) colloquial speech; (mainly) written language
Origin	loanword, loan phrase, purism; Anglicism, Germanism, Gallicism

A number of examples illustrate how these labels function in practice (see Table 2). For instance, for Dutch *achter* 'behind', both *efter* and *achter* are offered, differentiated by labels indicating written versus colloquial usage. Similarly, the Dutch *tante* 'aunt' may be translated as *tante*, labelled as a colloquial loanword, or *muoike*, marked as obsolete and associated with written language. These examples show that the dictionary does not simply provide a single standard equivalent, but already attempts to represent variation and guide users in navigating it.

TABLE 2 Examples of the use of labels

Dutch	Gloss	Translation	Label
achter	after, behind	efter achter	(mainly) written language (mainly) colloquial speech
zaterdag	Saturday	sneon saterdei	no label, considered as standard form Wood Frisian
getuige	witness	tsjûge getuge	no label, considered as standard form loanword
smartphone	smartphone	smartphone snoadfoan	no label, considered as standard form purism
tante	aunt	tante muoike	(mainly) colloquial speech; loanword obsolete; (mainly) written language
vergadering	meeting	gearkomste gearsit fergearring fergadering	(mainly) written language (mainly) written language literary (mainly) colloquial speech; loanword

In addition, survey data on the use of linguistic resources further highlight the complexity of the lexicographic context. While earlier survey data indicate that 46% of respondents had access to a (digital) Frisian dictionary (Klinkenberg et al., 2018), more recent data provide insight into actual usage patterns (Dekker et al., 2025). The results indicate that active use of such resources is relatively limited: a large majority of respondents report little, or no use of physical dictionaries (86%) and a substantial proportion report limited use of digital dictionaries (72%). Similarly, most respondents (62%) indicate that they rarely or never use translation tools for Frisian, such as Google Translate or the integrated translation function within Frysker. However, among those who do use Frysker, the integrated translation function appears to play an important role: approximately 80% of Frysker users make use of this functionality. This suggests that dictionary content is often accessed through translation workflows rather than through direct consultation. Moreover, informal feedback from users indicates that resources such as Frysker are generally well known and positively valued, even among speakers who do not use them on a regular basis.

At first glance, these findings might suggest that lexicographic resources play only a limited role in everyday Frisian language practices. However, a more nuanced interpretation emerges when these results are considered in relation to broader patterns of language use. Rather than relying heavily on formal or standardized resources, speakers appear to draw primarily on their existing linguistic competence, often rooted in spoken language. Writing practices, in particular, tend to reflect everyday usage rather than strict adherence to codified norms (Jongbloed-Faber, 2021). Digital resources are nevertheless widely appreciated among certain user groups, including students, journalists, and active Frisian speakers. This apparent discrepancy between reported usage frequency and perceived usefulness indicates that lexicographic tools occupy a more nuanced position: while not consistently used, they remain an important reference point, particularly in situations of linguistic uncertainty. For lexicography, this has important implications. The relatively limited reported use of dictionaries does not necessarily indicate a lack of need but may instead point to a mismatch between existing tools and users' actual practices and expectations. Moreover, it highlights the potential role of lexicographic resources in supporting speakers who experience uncertainty about their language use, particularly in more formal or public contexts.

Despite this relatively elaborate system of labels (Table 1), closer inspection reveals a number of challenges that indicate a need for further development. The label system has evolved gradually over time and does not appear to be based on a consistently articulated framework. As a result, individual labels vary in scope, precision, and usefulness. In some cases, distinctions are not fully transparent or systematically motivated. For instance, labels such as *Anglicism*, *Germanism*, and *Gallicism* overlap with the broader category of *loanword*, which in practice is primarily used to refer to borrowings from Dutch. This creates ambiguity, as it remains unclear whether the label loanword refers specifically to Dutch borrowings, to borrowing in general, or to a particular historical layer of lexical influence. In addition, some labels lack a clear and consistently applied definition. Boundaries between categories are often difficult to establish in practice. For example, the distinction between forms labelled as 'obsolete' and those described as 'becoming obsolete' is not straightforward, and the criteria underlying such classifications are not explicitly defined. Similar issues arise in the category of frequency, where labels such as 'rarely used' and 'on occasion' are employed, despite being only loosely distinguishable in meaning. A further issue concerns the degree of subjectivity involved in the assignment of labels. Categories such as 'obsolete', 'becoming obsolete', or 'frequently used' are not based on systematically collected usage data, but rather reflect lexicographic judgement and experience. While such expert-based assessments are often unavoidable, this means that different editors may arrive at different evaluations of the same form, particularly in a linguistic context characterized by variation across regions, speaker groups, and communicative settings.

The treatment of neologisms presents a further complication. In the case of Frisian, many innovations enter the language via Dutch, which raises questions about how newly introduced forms should be categorized and evaluated. The example of *smartphone* (Table 2) illustrates this tension: while the purist formation *snoadfoan* exists as a theoretically "Frisian" alternative, it appears to have very limited uptake in actual usage.

Taken together, these observations suggest that the current labelling system does not always align with actual language practices or user expectations. Some labels may offer valuable guidance, while others are applied inconsistently or may be of limited relevance to users. This points to a broader issue: the extent to which existing lexicographic categories reflect the needs, competences, and perspectives of different user groups. These limitations do not invalidate the current system, but they do highlight the need for a more systematic and user-oriented approach to the representation of variation. In particular, they raise questions about how labels can be refined, reorganized, or supplemented in ways that better support users in navigating linguistic variation. The following section therefore examines the linguistic reality with which this lexicographic system must engage, focusing on patterns of variation, acceptance, and language use in spoken Frisian.

3 Spoken Frisian and the tension between standard and usage

Unlike Standard Frisian, which is codified in grammars, orthographic conventions, and dictionaries, spoken Frisian is characterized by extensive variation, much of which arises through sustained contact with Dutch (Sjölin, 1976; Breuker et al., 1984; De Haan, 1997). Both L1 and L2 speakers regularly employ Dutch loanwords and contact-induced constructions in everyday speech (Stefan et al., 2026), which lexicographic approaches must therefore take into account. However, Dutch influence is not uniform: some forms are widely used, while others remain limited to smaller groups. This variation is illustrated by quantitative data from L1 speakers who reported the highest level of spoken Frisian competence in the survey (Tables 3-4). For example, alongside the Frisian form *besteegje*, the Dutch-derived *bestede* 'to spend' is used by a majority of respondents (Table 3), whereas in other cases the Frisian form dominates, as with *h(i)elendal* versus *helemaal* 'completely' (Table 4).

TABLE 3 Frequency of use of *besteegje* and *bestede* ('to spend') among highly proficient L1 speakers (n = 479)

Variant	Origin	Percentage	Interpretation
besteegje	Frisian	26%	minority usage
bestede	Dutch	74%	dominant in spoken usage

TABLE 4 Frequency of use of *h(i)elendal* and *helemaal* ('completely') among highly proficient L1 speakers (n = 476)

Variant	Origin	Percentage	Interpretation
h(i)eledal	Frisian	77%	dominant in spoken usage
helemaal	Dutch	23%	minority usage

Age-related patterns further complicate the picture. Lexical choice varies across generations, suggesting ongoing change. In the case of 'to eat sweets' (Dutch *snoepen*), Table 5 shows a clear age gradient: younger speakers prefer the Dutch-derived *snoepe*, while older speakers favour the Frisian *snobje*. This indicates a shift toward increased use of Dutch-derived forms, although more recent data would be needed to confirm its extent.

TABLE 5 Age-related variation in the use of *snobje* and *snoepe* ('to eat sweets') among highly proficient L1 speakers (n = 472)

Age group	snobje (Frisian)	snoepe (Dutch loan)
<30	27%	73%
30-39	31%	69%
40-49	35%	65%
50-59	53%	47%
60-69	72%	28%
≥70	84%	16%
Total	53%	47%

However, frequency of use alone does not determine the status of such forms. Their acceptance among speakers constitutes a second, equally important dimension. A sociolinguistic study in Fryslân (Stefan & Klinkenberg, 2019) shows that neither strongly puristic nor heavily Dutch-influenced texts are perceived as fully acceptable. Instead, speakers orient toward an intermediate norm. Crucially, this evaluation is selective: some loanwords are accepted, while others are rejected. For instance, *iderien* (from Dutch *iedereen*) instead of Frisian *elkenien/eltsenien* 'everyone' is widely accepted, whereas *bijna* (from Dutch *bijna*) instead of *hast* ('almost') is generally disapproved of. This indicates that speakers have fine-grained intuitions about which contact-induced forms are compatible with Frisian.

Language contact in spoken Frisian is therefore gradient: forms are widespread but unevenly distributed and variably accepted. Speakers actively navigate between variants rather than reproducing Dutch influence indiscriminately. As shown in Table 6, awareness of lexical variation is considerably more widespread than the active adjustment of one's own speech.

TABLE 6 Reported adjustment of Frisian language use and perception of 'correct' usage among speakers (n = 3271/3655); based on an analysis of data from a recent sociolinguistic survey in Fryslân (Dekker et al. 2025)

Response category	Adjusting own language use (%)	Recognizing 'correct/incorrect' usage (%)
Never	37%	15%
Rarely	24%	12%
Sometimes	27%	31%
Often	9%	35%
Always	3%	7%

These findings highlight the complexity of spoken Frisian, where frequency, acceptance, and usage do not always align. This creates a fundamental challenge for lexicography: dictionaries must provide guidance that reflects actual usage while maintaining a coherent norm. The variability observed in spoken Frisian therefore highlights the need for a lexicographic approach that accommodates gradience rather than enforcing a strict dichotomy between "correct" and "incorrect." This tension between descriptive adequacy and normative guidance lies at the core of contemporary Frisian lexicography.

4 Example-based challenges in labelling

The complexity of representing variation in a lexicographic framework can be illustrated by means of concrete examples, which reflect broader patterns observed in the survey data discussed above. A first example concerns the Dutch adjective *gezond* 'healthy', which is translated in ONFW as *sûn*, without any additional label. This presentation suggests that *sûn* functions as the standard Frisian equivalent. However, in actual language use, forms such as *gesond* – a loanword from Dutch adapted to Frisian phonology – are also frequently used by speakers. While the dictionary presents only the standard form, this does not fully reflect the range of forms encountered in practice. A similar issue arises in the case of derived forms such as *gezondheid* 'health', represented as *sûnens* in the dictionary. Although *sûnens* can be considered the conventional Frisian equivalent, other forms (e.g. *gesondheid*) occur in actual usage, particularly under the influence of Dutch. At the same time, not all formally conceivable variants are attested in practice. For instance, while Dutch allows derivations in *-heid*, in Frisian one-syllable adjectives typically form abstract nouns with *-ens* (e.g. *sûnens* 'health', *smoargens* 'dirt, filth', *tûkens* 'cleverness'), which means that alternative formations are not always morphologically possible. This illustrates a broader point also reflected in the empirical data: not all possible or historically motivated forms have the same status in actual language use, and their distribution across speakers remains uneven.

A second example concerns more complex lexical items such as *snelheidsbeperking* 'speed limit'. The dictionary provides *faasjebeheining* as a Frisian equivalent, labelled as formal and associated with purist language use. However, this form appears to have little uptake in everyday communication. By contrast, the Dutch-based form *snelheidsbeperking*, often adapted to Frisian pronunciation, is widely used by speakers. Importantly, this case differs from examples such as *sûn* versus *gesond*. In the latter case, a Frisian-origin form exists alongside a contact-induced variant. In the case of *snelheidsbeperking*, however, the problem is more fundamental: there is no widely accepted or fully equivalent Frisian alternative. Forms such as *faasjebeheining* or *faasjebeperking* may exist as alternatives motivated by purist lexical principles, but they are often perceived as unnatural or are not used productively in everyday language. Similar issues arise with other lexical items that are difficult to translate. For instance, the Dutch term *veelpleger* 'repeat offender, recidivist' does not have an established Frisian equivalent and is not consistently represented in the dictionary. In practice, news media such as Omrop Fryslân frequently use the Dutch term "veelpleger" – sometimes marked with quotation marks – within otherwise Frisian texts. A comparable situation can be observed with words such as *kringloopwinkel* 'second-hand shop'. While the dictionary offers alternatives such as *twaddehânskwinkel* and *ynbringwinkel*, these do not fully correspond to the meaning or usage of the Dutch term, and their actual use is unclear. In contrast, *kringloop* itself is used in practice, sometimes even without explicit marking as a borrowing.

These examples illustrate a further dimension of the lexicographic challenge. In contexts of intensive language contact, not all concepts can be translated in a straightforward or functionally equivalent way. In such cases, bilingual speakers may simply adopt the Dutch term in Frisian discourse, resulting in forms that are widely used but difficult to accommodate within

a traditional lexicographic framework based on clear equivalents. This highlights a key limitation of the current approach: the assumption that each lemma should correspond to a well-defined target-language equivalent does not always hold. Instead, lexicographic practice must also account for cases in which equivalence is partial, unstable, or absent.

5 Towards a more flexible and user-oriented lexicographic model

The analysis presented in this paper has shown that contemporary Frisian lexicography operates within a complex and dynamic linguistic environment, in which variation, usage, and normative expectations do not always align. The preceding sections have illustrated how standard forms coexist with contact-induced variants, how acceptability and frequency diverge, and how, in some cases, no fully equivalent Frisian alternative exists. Together, these observations point to a fundamental tension between lexicographic representation and actual language use. A key challenge for future lexicographic practice is therefore to move beyond the current, partly implicit representation of variation toward a more structured and transparent approach. While existing resources already provide multiple variants in some cases, this variation is not always systematically organized or aligned with different user needs. Rather than resolving variation by selecting one "correct" form, dictionaries can provide users with structured access to multiple variants, together with information relevant to their specific needs and contexts of use.

Building on this idea, online lexicographic tools can be designed to support multiple modes of access tailored to different types of users and communicative situations. Table 7 provides an overview of the proposed dictionary modes and their respective functions.

TABLE 7 Overview of proposed dictionary modes

Mode	Purpose	Use situation/user need	Output
Normative mode	clear guidance, no ambiguity	writing in formal or standard contexts; users seeking clarity and a single, reliable option	one recommended form
Comparative mode	insight into variation	exploring alternatives; understanding differences between variants	multiple variants + labels + usage notes
Learning mode	support language acquisition	language learning; users with limited competence or confidence	accessible forms + explanations

As shown in Table 7, these modes do not represent separate dictionaries, but rather different perspectives on the same lexical data. In a normative mode, the dictionary provides a single recommended form, allowing users to write with confidence without having to consider alternatives. In a comparative mode, by contrast, variation becomes visible: users can access multiple forms, accompanied by labels and usage notes that help them interpret differences in distribution, frequency, and acceptability. A learning mode may further simplify this information by focusing on forms that are particularly accessible and useful for less experienced users. It can also provide insight into actual usage by indicating which forms are most commonly used by speakers, even when these differ from the standard form. Importantly, these modes are not mutually exclusive. Rather, they allow different types of users – ranging from confident L1 speakers to L2 learners – to interact with the same lexical resources in ways that

correspond to their needs, competences, and expectations. In this framework, the existing preferred-vocabulary list can be understood as an instance of a normative mode. By identifying a single preferred form, it provides clear and unambiguous guidance, particularly in contexts such as formal writing or official communication, where variation may be undesirable. For users who "do not want to choose," this type of guidance is essential: it offers a form that can be used with confidence, without requiring knowledge about variation or underlying linguistic distinctions. This approach is intentionally limited in scope. It does not aim to represent the full range of linguistic variation or to provide detailed information about the status of variants. Rather than replacing such a system, a multi-mode approach integrates it into a broader framework, allowing users to follow clear guidance or explore variation.

A second key area for development concerns the systematic representation of linguistic variation. The examples discussed in this paper show that variation in Frisian operates along multiple dimensions, including differences between standard and non-standard usage, the presence of contact-induced forms alongside Frisian variants, and regional or dialectal variation. In addition, variation may involve cases where no fully equivalent form exists, leading speakers to adopt Dutch lexical items in Frisian discourse. A revised label system could therefore aim to make such distinctions more explicit and more transparent. Rather than relying on loosely defined or inconsistently applied categories, labels could be structured around a limited number of clearly motivated dimensions, such as standardness, borrowing status, register, and medium. Crucially, such labels would primarily serve to contextualise non-standard or marked forms, while unmarked forms function as the default and do not require explicit labelling. It is important to recognise that not all variation can be captured through clear-cut categories. Differences between forms may be gradual, context-dependent, or unevenly distributed across speakers. Labels should therefore not function as prescriptive judgments, but as interpretive tools that enable users to understand how different forms are used in practice.

The proposed model also highlights the need for further research into the users of Frisian lexicographic resources. As the survey data discussed above indicate, dictionary use is uneven and often indirect: users may access lexical information through translation tools rather than through direct consultation or rely on their own linguistic competence in everyday communication. Dictionaries nevertheless remain an important reference point in situations of uncertainty. A better understanding of user profiles, needs, and expectations is therefore essential. Different users may seek different types of information: some prioritize clarity and correctness, others wish to explore variation, and still others require support in language learning. A flexible lexicographic system must be able to accommodate these different perspectives.

Finally, the increasing integration of lexicographic resources with other language technologies offers further opportunities for development. The widespread use of translation tools suggests that dictionary data are often embedded in broader workflows of writing and communication. This raises the possibility of aligning dictionary output with different types of linguistic variation, for instance by adapting translations to different contexts or user profiles. While such developments remain exploratory, they point to a more dynamic and interconnected lexicographic ecosystem.

In conclusion, the findings presented in this paper point to the need for a more flexible, user-oriented, and empirically grounded approach to Frisian lexicography. Dictionaries can support users more effectively by making variation visible, structuring it transparently, and providing different types of guidance depending on context and need. Frisian thus offers an instructive case for minority-language lexicography more broadly. It shows how online dictionaries can mediate between variation and standardization, supporting both linguistic diversity and user confidence without reducing this complexity to a single, uniform representation of correctness.

References

Atkins, S. B. T. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press.

Bergenholtz, H., Gouws, R. H. (2010). A Functional Approach to the Choice between Descriptive, Prescriptive and Proscriptive Lexicography. *Lexikos*, 20, 26–51. https://doi.org/10.5788/20-0-131

Breuker, P. (2001). West Frisian in language contact. In H. H. Munske (Ed.), *Handbuch des Friesischen/ Handbook of Frisian Studies* (pp. 121–129). Max Niemeyer.

Breuker, P., Dyksma-Hogeveen, M. G., Hiemstra, S. T., Scholten, K. C., Visser, W. & Zantema, A. (1984). *Foar de taalspegel: Koart oersjoch fan Hollânske ynslûpsels yn it Frysk*. Afûk.

De Haan, G. J. (1997). Contact-induced changes in modern West Frisian. In G. J. de Haan & O. de Vries (Eds.), *Dedicated to Bo Sjölin* (Us Wurk, 46) (pp. 61–89).

Dekker, S. V., Stefan, M. H., Van Seijen, F., Dijkstra, J. E., Heeringa, W. J., La Roï, C. & Kircher, R. (2025). *Taal yn Fryslân: In nije koers*. Fryske Akademy.

Duijff, P., Van der Kuip, F. (2018). Lexicography in a Minority Language: A Multifunctional Online Dutch-Frisian Dictionary. *International Journal of Lexicography*, 31(2), 196–213. https://doi.org/10.1093/ijl/ecy009

Duijff, P., Van der Kuip, F. (Eds. in chief), De Haan, R. & Sijens, H. (Eds.) (2008). *Frysk Hânwurdboek*. Fryske Akademy/Afûk.

Frysker.nl (n.d.). *Online platform for the Dutch-Frisian dictionary (ONFW)*. Retrieved May 26, 2026, from https://frysker.nl

Fuertes-Olivera, P. A. (2025). *A guide to practical online lexicography*. Routledge.

Jongbloed–Faber, L. (2021). *Frisian on social media: The vitality of minority languages in a multilingual online world* [Doctoral dissertation, Maastricht University]. LOT.

Klinkenberg, E. L., Jonkman, R. J. & Stefan, M. H. (2018). *Taal yn Fryslân. De folgjende generaasje*. Fryske Akademy.

Lew, R. (2011). Studies in dictionary use: Recent developments. *International Journal of Lexicography*, 24(1), 1–4. https://doi.org/10.1093/ijl/ecq044

Lew, R. (2015). Dictionaries and their users. In P. Hanks & G.-M. de Schryver (Eds.), *International handbook of modern lexis and lexicography*. De Gruyter.

Lew, R., De Schryver, G.-M. (2014). Dictionary users in the digital revolution. *International Journal of Lexicography*, 27(4), 341–359. https://doi.org/10.1093/ijl/ecu011

Merkuur, A. A. (2021). Changes in modern Frisian verbal inflection. [Doctoral dissertation, externally prepared, Universiteit van Amsterdam]. LOT.

Provinsje Fryslân. (2020). *De Fryske taalatlas 2020: De Fryske taal yn byld*. Provinsje Fryslân.

Sjölin, B. (1976). *"Min Frysk": Een onderzoek naar het ontstaan van transfer en "code-switching" in gesproken Fries*. Frysk Ynstitút.

Stefan, M. H. (2022). *Spoken Frisian: Language Contact, Variation and Change* [Doctoral dissertation, externally prepared, Universiteit van Amsterdam]. LOT.

Stefan, M. H., & Klinkenberg, E. L. (2019). *De rol en it belang fan it Frysk foar it ûnderwiis yn Fryslân*. Provinsje Fryslân.

Stefan, N., Klinkenberg, E., Versloot, A. (2026). L1 and L2 speakers' competence in a minority language (Frisian) compared. *Taal en Tongval*, 78(1), 64–104. https://doi.org/10.5117/TET2026.1.004.STEF

Tarp, S. (2008). *Lexicography in the Borderland between Knowledge and Non-Knowledge: General Lexicographical Theory with Particular Focus on Learner's Lexicography*. Max Niemeyer.

Contact information

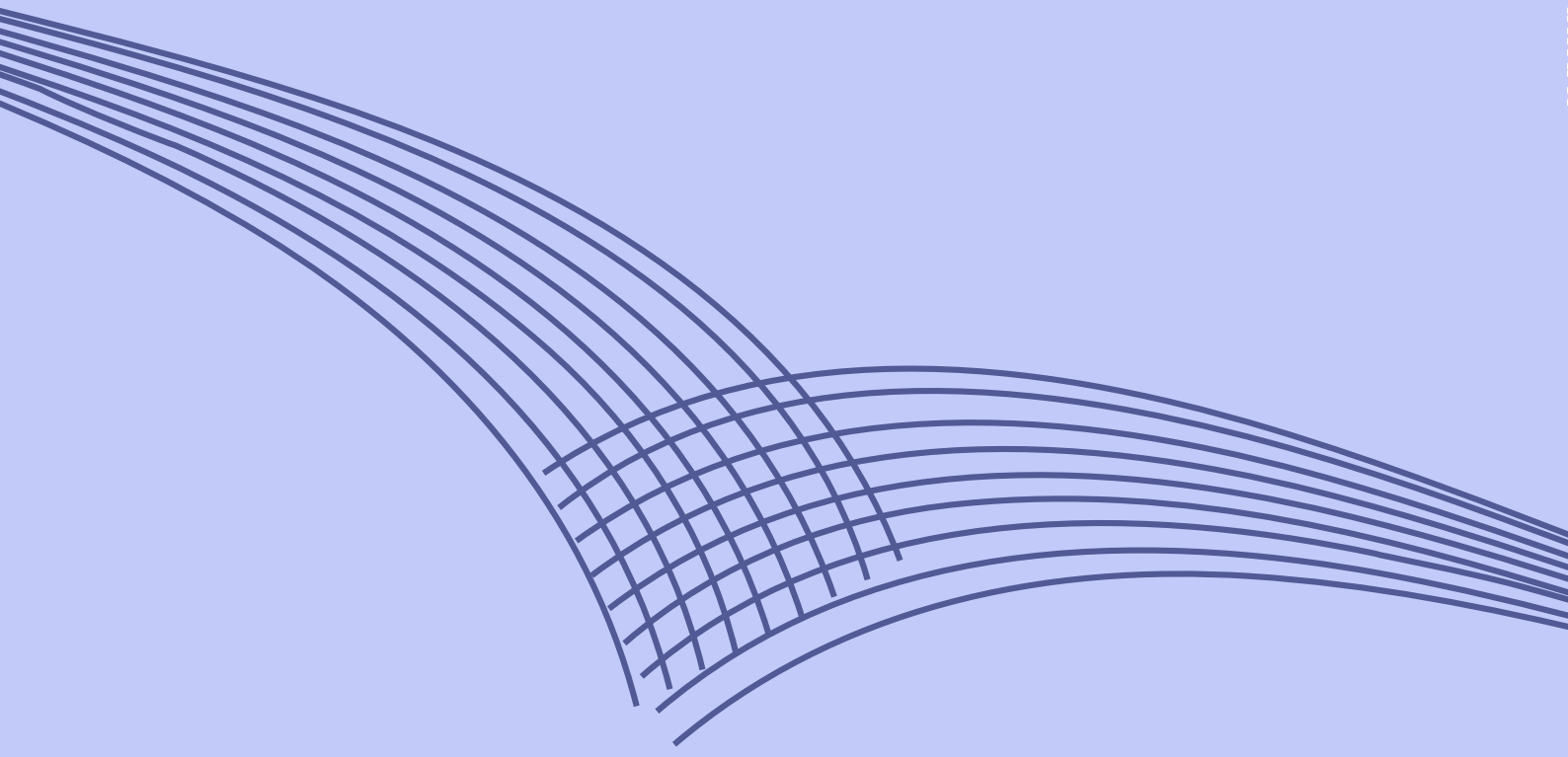
Nika Stefan
Fryske Akademy
nstefan@fryske-akademy.nl

Johan van der Zwaag
Fryske Akademy
jvanderzwaag@fryske-akademy.nl

Hindrik Sijens
Fryske Akademy
hsijens@fryske-akademy.nl

SPECIALISED & TERMINOLOGICAL LEXICOGRAPHY

EURALEX 2026



Martina Bajčić & Ana Ostroški Anić

DEFINING LEGAL CONCEPTS IN THE AI AGE

An Exploratory Study into Influencer Practices

Abstract Definitions play a crucial role in the acquisition, transfer, or translation of specialized knowledge. In the legal profession, precise definitions of complex concepts facilitate seamless knowledge management, application, and interpretation. However, despite growing reliance on artificial intelligence (AI) for this task, the real-world feasibility and user acceptability of AI-generated legal drafting remain largely unexamined. When addressing this gap, it is critical to consider that the use of authoritative legal terminology is central to the validity of any legal definition. Indeed, precise terms facilitate the application and interpretation of core concepts. However, new legal domains lack stable legal terms and conceptually demarcated concepts. Resolving the inherent tension between linguistic evolution and the need for legal certainty make the legal domain a unique use case that necessitates domain-specific task adaptation for general-purpose large language models (LLMs). With that in mind, this paper explores the potential of LLMs to generate legal definitions perceived as clear, precise, and usable by domain users through an exploratory study of a generated dataset of key influencer-related concepts. The results underscore the critical importance of data quality and precise prompt engineering when leveraging LLMs for complex, domain-specific tasks.

Keywords Legal concepts; definitions; user evaluation; AI; LLMs

1 Introduction

In just one month after its release, ChatGPT has accumulated an estimated 100 million users, making it the fastest growing application ever (Barryman & Ziegler, 2025, p. 1). Revolutionizing information retrieval processes, OpenAI's ChatGPT and other LLM-backed applications are consulted for multiple purposes, including accessing and acquiring knowledge. Definitions play a crucial role in specialized knowledge acquisition, transfer, description or translation. They help users to cognitively anchor the meanings of concepts within a domain. Beside legal terms which offer points of access (Biel, 2009) to legal knowledge, and legal concepts that frame legal knowledge (Bajčić, 2017), definitions of legal concepts play an instrumental role not only in accessing, acquiring and transferring legal knowledge, but also in legal translation and terminology management. Despite their unequivocal relevance, especially from the user's perspective, the potential of AI-generated definitions of legal concepts has only recently begun to be investigated (Zilli et al., 2025), unlike AI-powered term extraction, dictionary compilation, or AI-generated definitions of concepts from other specialized domains (Nahod, 2024; Silalahi, 2026). At the same time, a recent survey conducted among law students revealed they too rely on LLM-backed applications to define legal concepts, as corroborated by 86 % of the respondents.[1] This is hardly surprising considering not only the cognitive ease of using AI tools, but also the anthropomorphism of AI which may make its outputs feel authoritative to students (Callister, 2025).

Against this backdrop, this paper explores the potential of AI tools in defining legal concepts. Since law is a high-stakes domain, legal definitions must convey credible information using authoritative, authentic terminology. To evaluate how effectively AI tools can meet these rigorous standards, we conducted an exploratory study involving key user groups by testing an

[1] https://moodle.srce.hr/2025-2026/mod/feedback/show_entries.php?id=4852116

AI-curated dataset of terms and definitions compiled for this purpose. Our aim was to gain critical insights into users' evaluations of the generated definitions. The results underscore the paramount importance of high-quality data and precise prompt design for complex, domain-specific tasks.

The study is guided by two research questions:

RQ1: To what extent do domain users (legal scholars and, in particular, law students) perceive NotebookLM-generated definitions of established legal concepts as clear, precise, and usable?

RQ2: Can the same generative process produce comparably well-received definitions of emerging, as-yet unregulated concepts for which no authoritative source exists?

1.1 Domain and task demarcation

This study focuses on an emerging legal domain centred on influencer activities. This field is inherently hybrid, blending novel, undefined legal terms with cross-disciplinary concepts from computer science, economics, and marketing (Bajčić, 2024), creating a uniquely challenging terminological landscape. Compounding this complexity, digital influencer marketing has surged over the past decade, yet the term *influencer* entered the Oxford English Dictionary in 2019.[2] The influencer market has grown exponentially over the recent years, especially after the COVID-19 pandemic, signalling a history rapidly in the making. Influencers engage in many different activities of digital marketing requiring clear advertising disclosures.

In the eyes of EU consumer protection law, influencers act as traders. Consequently, provisions of EU consumer protection law apply to them. Nevertheless, according to a recent report by the European Commission, 77% of influencers violate the need to disclose *paid partnerships* when endorsing products on different platforms.[3] Red flags are also raised due to the fact that they use different wordings instead of authoritative legal terms, which results in explicit violation of their legal obligation. Over the last years, a handful of EU legislative acts have been adopted with a view of regulating the behaviour of influencers and potential effects thereof on consumers (e.g. Digital Services Act, Digital Fairness Act, Unfair Commercial Practices Directive). At the same time, new terms are constantly created in light of the fast-paced development and the advent of new influencer practices and concepts that are yet to be regulated and defined by the law such as *aura farming, echo influencer, organic content, advertorial*, etc. The creation of new terms is a natural corollary of knowledge expansion (Kaguera, 2022, p. 457). Yet, while terminology is generally dynamic, flexible, and prone to conceptual shifts, legal terminology demands long-term stability and consistent application to adhere to legislative drafting rules and uphold legal certainty (Bajčić, 2024). Consequently, the use of authoritative legal terms, i.e. those actively recognized within the domain, is instrumental to the clarity of legal definitions. These precise terms are immediately recognizable to practitioners, facilitating the application and interpretation of core concepts. Resolving the inherent tension between linguistic evolution and legal rigidity presents a formidable terminological challenge, rendering the legal domain a unique use case that necessitates domain-specific task adaptation for general-purpose language models.

[2] Oxford University Press. (n.d.). Influencer, n., 2.b. In Oxford English dictionary. Retrieved February 5, 2026, from https://doi.org/10.1093/OED/4737815373

[3] European Commission Sweep Report, Available at: https://ec.europa.eu/commission/presscorner/detail/en/ip_24_708

2 Methodology

Considering the novelty of this legal domain, we decided to compile the terminological dataset by retrieving legislative texts from EU consumer protection law, materials from the web portal Influencer Legal Hub, and scholarly articles using the LLM-based application NotebookLM.[4] Whereas existing, pre-compiled corpora offer empirical stability, they are often insufficient for newer domains in which there are not enough expert resources and where state-of-the-art terms have not been integrated. In contrast, LLMs might yield more data but their performance requires continuous evaluation. Devising clear methodological steps that can be replicated and checked for robustness is therefore crucial in any LLM-based application. This study proposes a method that combines automatic means of compiling and defining terminological data with human verification by terminologists and legal experts. The experiment datasets (prompts, terms, definitions of concepts, participant instructions and survey results) are available in open access in the institutional repository of the author's institution.[5]

2.1 Data collection

The process of creating the terminological dataset used for the testing consisted of several steps. First, using the web scraping method, 30 sources were retrieved and verified for credibility and relevance by the authors. Subsequently, three sources were omitted for lack of relevance. The following prompt was used for web scraping:

> *Scrape the web to retrieve relevant legal information about influencers. Retrieve legislative documents (EU consumer protection law), Influencer Legal Hub and scholarly articles.*

After the relevant data were scraped, 20 most relevant legal terms for influencer activity were elicited using the following prompt:

> *Role: You are a legal terminology expert. Task: Scan the uploaded sources to identify the 20 most relevant and frequent legal terms pertaining to the activity of influencers.*

The selected terms were validated for relevance by terminologists and EU law experts. The list includes standard terms belonging to consumer protection law, such as *trader, right of withdrawal, joint liability, undertaking and information duties*, but also new terms *like influencer, influencer agent, commercial collaboration* and *dark patterns*. As a next step, a prompt was designed to craft legal definitions meeting the requirements of legal precision and use of authoritative legal terms and their role in law as discussed above, as well as in keeping with key prompt engineering guidelines (Barryman & Ziegler, 2025, p. 11). Interestingly enough, NotebookLM defined the terms in the first prompt iteration when it was asked to provide terms only, even though it wasn't instructed to do so.

Next, the definition role prompt was used in order to retrieve legal definitions of the extracted terms:

> *Role: you are a legal terminology expert. Task: Define these legal terms. Legal definitions must be precise and include only authoritative legal terms actually used in legislative acts.*

[4] NotebookLM: Google. (2026). NotebookLM. https://notebooklm.google.com

[5] The data can be reached here: https://repository.pravri.uniri.hr/object/pravri:4718. The dataset includes the list of legal concepts concerning influencer activities, generated definitions of these concepts, a benchmark of authoritative legal definitions, and non-binding related definitions and/or explanations. The survey data contains students' evaluations of definitions as well as the evaluations of legal experts.

However, the definitions generated in the second iteration differed from the definitions the model had provided together with the list of terms. For example, *right of withdrawal* was first defined as "the consumer's statutory right to return an online purchase or cancel a service within 14 days," whereas the legal definition prompt identified it as "the statutory right for a consumer to change their mind and return an online purchase within 14 days of receiving the goods or access to a service."

The latter definition received positive evaluation by both legal scholars (average grade: 3) and students (average grade: 4.2). Another term for which the model provided different definitions is *hidden advertising*, which was initially defined as

> "the illegal practice of using paid editorial content in the media to promote a product without making that clear to the consumer".

Following the legal definition prompt, it was defined as

> "the illegal practice of failing to disclose a commercial relationship, characterized legally as a misleading omission of material information that a consumer needs to take an informed decision."

The second definition of *hidden advertising* displays more legal nuance, precision and is expressed in more legal terms. These examples demonstrate the importance of including domain-specific metadata into the prompt in particular when applying general-purpose language models to perform specific domain tasks. In general, the terminological validation of the generated definitions showed that they are slightly less formal than traditional statutory definitions. Although they include overall legal phrases and wording, as well as legal terms, they also include additional elements like embedded explanations, synonyms and vague expressions (e.g. *typically*), as will be analysed in the following sections.

Finally, the model was prompted to provide terms and definitions of concepts that are not legal in the strict sense but that arise in the context of influencer marketing activities. The following prompt was used:

> *Task: Scan the uploaded sources to identify 20 most relevant terms pertaining to the activity of influencers (such as paid partnership, barter, sponcon). Include all terms, not just legal.*

Following the evaluation of the terms and definitions generated by the model, we selected 10 non-legal terms to complement the 20 legal terms that had been selected previously. The final dataset used for testing consisted of 30 legal and non-legal terms, together with their NotebookLM-generated definitions.

2.2 Research aim, scope and participants

Our primary research aim was to examine the clarity and legal precision of NotebookLM-generated definitions of legal concepts, which was evaluated from a user perspective through two empirical studies. In this context, achieving clarity and precision requires the use of authoritative legal terms that accurately capture the underlying legal purpose. Consequently, the evaluation focused on whether the generated text maintained the clarity and precision necessary for legal compliance and understanding. The first empirical study was a pilot study involving two EU law scholars who rated the definitions to assess the robustness of our approach. The purpose of this pilot evaluation was to verify that the definitions were not invalid (i.e., evaluated negatively), misleading, or entirely hallucinated, as AI-generated content fundamentally requires human oversight. Subsequently, a second study was conducted with 13 law students enrolled in the 4th year of study at the University of Rijeka, including Croatian

students as well as Erasmus students from Greece, Italy, and Poland. Law students were chosen because they represent the population most likely to consult LLM-based tools for legal information. As noted in the Introduction, 87.5% of 70 surveyed second-year law students report relying on LLM-backed applications to define legal concepts. Evaluating the tool's output from the perspective of its actual end users allowed us to assess not only its technical adequacy, but also its comprehensibility and usability.

Both the definitions and the instructions were provided entirely in English. The evaluation scale ranged from 1 – *I do not understand the definition at all*, to 5 – *the definition is legally valid, clear, and precise; I fully understand it and can apply it*. The tested sample included 30 definitions of legal concepts, ranging from established and better-known concepts among domain experts (e.g., *trader, undertaking, information duties*) to emerging legal concepts (e.g., *influencer, hidden advertising, dark patterns*). The sample also included the definitions of entirely new non-legal concepts that can be described as unregulated domain concepts that only started to appear (in domain content), and that are yet to be legally regulated (e.g., *echo influencer, dropshipping, aura farming*). In this way, we could test if new concepts can also be validly defined, and whether existing data could be used to create new data.

2.3 Results

All the definitions, both of legal and emerging concepts, were rated positively by the students. In general, the definitions of established, well-known legal concepts were rated with grades ranging from 3 to 4 by both user groups. While it is not possible to describe all of the results here for space constraints, Table 1 shows the average grades for the better-known legal concepts:

TABLE 1 average grades of definitions

user group	Trader	undertaking	joint liability	right of withdrawal
EU law scholars	4.5	2.5	4.5	3
law students	4	4	3.9	4.2

Some definitions were rated equally by both groups, such as the definition of the key concept of *influencer* (4.5):

"*influencer* (electronic commercial influence): any natural or legal person, who, for a fee, uses their reputation among their audience to communicate content to the public by electronic means for the purpose of promoting, directly or indirectly, goods, services, or any clause whatsoever."

The definition of *undertaking* was the only generated definition in which the concept is explicitly identified as 'a legal concept':

"a legal concept used in market practice codes to describe any natural person or legal entity pursuing an economic purpose in a sustainable manner."

Starting a definition with the phrase *a legal concept* is considered poor practice in legislative drafting. All the other definitions directly stated what the concepts are by referring to superordinate concepts rather than meta-describing their role. Additionally, this definition introduces subjective qualifiers (*in a sustainable manner*) and vague references (*market practice codes*),

failing to provide the precise boundary and legal certainty required for legislative enforcement, which explains the lower rating given to the definition by the scholars. Other slightly lower rated definitions included those for *dropshipping*, which was awarded a grade of 3 from scholars and 4.5 from students, *sponcon* (graded with 3.6 by students, but with 4.5 by scholars), and *addictive design* (3.5 by scholars, and 3.9 by students):

> *dropshipping*: a commercial practice where influencers promote products for which fulfilment is handled by a third party; influencers under certain laws are responsible for ensuring these products are available and legal
>
> *addictive design*: interface features engineered to maximize engagement time, such as infinite scroll, autoplay, and endless reward loops
>
> *sponcon* (sponsored content): advertising material embedded in the materials created by an influencer that typically does not differ from the organic content they regularly produce

These three definitions follow the customary structure of the definitions of legal concepts and adhere to the principles of traditional terminological definitions in which the concept is defined by its relation to a superordinate concept (*practice, interface* and *material*) and the characteristics that set it apart from related concepts. However, they also introduce new non-legal terms, explanations and even vague expressions like *typically* which are not normally part of statutory definitions. One possible explanation for introducing these elements might be the lack of legal sources from which definitions could be used. Another reason might be the conceptual instability and fluctuation in the use of certain terms that are not yet stable in their concept systems.

3 Discussion

The main advantage of using LLMs is the possibility to fine-tune them to perform specific tasks. To augment LLMs, it is necessary to provide them with the right datasets, instructions, examples and make the application first retrieve content relevant to the problem (Barryman & Ziegler, 2025, p. 105). Retrieving proper legal terminological data for emerging domains can be problematic as there are no legislative acts enacted yet, or no settled case law. To lever this, we included terminologically informed instructions for defining legal concepts and for defining new concepts that are yet to become legally regulated. In this way, we tested if the tool can mimic the dataset to define new undefined, unregulated concepts related to influencers and unseen terms not present in the scraped data, and potentially create synthetic terminological data in the form of legal definitions for new (legal) concepts based on the model of the previously defined legal concepts.

In general, the definitions were evaluated positively, demonstrating the tool's ability to generate definitions that domain users perceive as clear, precise, and usable. The generated definitions nevertheless exhibited features not normally present in statutory definitions. For instance, some of the definitions (of both legal and non-legal concepts) embedded term variants or synonyms within brackets, as illustrated in Table 2:

TABLE 2 examples of synonyms included in the generated definitions

defined concept	sponcon	in-kind compensation	joint liability	aura farming	product seeding
term variants in brackets	sponsored content	consideration	shared liability	fake engagement	hub seeding

For the concepts *sponcon* and *joint liability*, the term variants the model presented in brackets along with the main term can be considered variants in form (*sponsored content*) or synonyms (*joint liability* vs. *shared liability*). In other examples, however, the existence of another term for the defined concept is proof of the terminological variation and instability that is present in the domains whose terminology has not yet been fully established. In the process of harmonization, when conceptual coherence hasn't been established yet (ISO 860, 2007), it is common that several terms for one concept are used parallelly, particularly in certain contexts or in different registers. For example, *sponcon* can be considered to be more of a colloquial term or an abbreviated form of the more appropriate legal term *sponsored content*.

The definition of *aura farming* is all the more interesting because it offers three terms for allegedly the same concept, namely *aura farming, fake engagement* and *crowdturfing*:

> *aura farming (fake engagement)*: the unethical practice of purchasing fake followers, likes, or plays from a "fake engagement service" to artificially inflate perceived influence and increase market value. This is also known as crowdturfing.

Since the understanding of emerging concepts in dynamic domains like social media communication can change over a short period of time, a closer analysis of these terms that the model had grouped around the concept of *aura farming* revealed that the terms do not in fact refer to the same concept. As defined by Merriam Webster, *aura farming* refers to the act of doing something to look cool, impressive, or stylish (MW),[6] and it is mostly used as a slang term for trying to create an aura of effortless coolness or charisma. *Crowdturfing*, on the other hand, is indeed a term closely related to *fake engagement* since it denotes a "phenomenon wherein individuals generate fake engagement using their authentic profiles, behind a monetary reward" (Tricomi, 2024, p. V). *Crowdturfing* might therefore be defined as a subordinate concept to *fake engagement*, which is a broader concept that includes both human and artificial engagement on social media, e.g., bots, click farms, and AI-generated content.

In addition to providing term variants in brackets for certain defined concepts, Notebook LLM generated a few definitions in which additional explanations were also included in the definitions, providing for additional conceptual demarcation of the concepts. Such practices, though not typically part of traditional legislative drafting, as witnessed in our legal definitions benchmark, might further understanding of the defined concepts, provided they are legally accurate. Table 3 illustrates examples of additional explanations given in the definitions:

TABLE 3 examples of explanations included in the generated definitions

defined concept	hidden advertising	influencer	parasocial relationship	impressum	synthetic content
additional explanation	covert/surreptitious	electronic commercial influence	interaction	provider identification	mandatory labelling

Furthermore, examples were embedded in brackets within the main definition, such as for the concept of *advertorial*:

> *advertorial*: editorial content in the media used to promote a product where a trader has provided consideration with an asset value for the publication (such as cash or the exchange of copyright-free photos).

[6] https://www.merriam-webster.com/slang/aura-farming

The evaluation of the 30 definitions suggests that they were generally perceived by users as clear and legally precise, receiving predominantly positive ratings across the evaluation criteria. Regarding the student evaluation data, it is difficult to discern clear patterns or statistically notable differences between their assessments of legal versus non-legal concepts. In fact, the average grades assigned by students to more established legal concepts ranged from 3.9 to 4.5, where mirrored by the average grades for non-legal concepts ranging from 3.6 to 4.5.

These preliminary findings indicate that the tool can perform adequately when defining novel concepts that are not yet legally regulated or defined. They also underscore its viability not only for domain-specific educational tasks but also for the generation of new synthetic data. However, these findings reflect user perceptions rather than an objective validation of the definitions' legal accuracy, which needs to be tested against a larger benchmark that includes authoritative legal sources as well as definitions provided by legal experts. The tool's efficacy within low-resource languages remains an area requiring further investigation and rigorous evaluation.

3.1 Limitations

The limitations of the conducted study concern the relatively small sample size of both the analysed definitions and the participants. While human evaluators of AI applications are essential to verify their usefulness, expert evaluators are difficult to find. For this study, it was deemed instrumental to include EU law experts considering the purpose of checking the credibility of the dataset and then have students evaluate it. To address the size constraints, future research should replicate this experiment on a larger scale. This would glean deeper insight into how AI tools perform when defining emerging concepts within high-risk domains, particularly having in mind that defining new concepts is hindered by the existence of several term variants for one concept, reflecting the conceptual fluctuation in the domain. Additionally, replicating the experiment in other domains would further test the generalizability of the devised methods. Beyond expanding the quantitative scope, conducting subsequent structured or semi-structured interviews with specific user groups would provide concrete qualitative findings regarding the feasibility of the proposed approach. Such analysis should also address the need for users of AI tools to develop contextualized AI literacy and terminological competence tailored to the specific domain. This oversight is paramount, given that even specialized legal databases integrating AI models remain susceptible to hallucinations (Magesh et al., 2025). Consequently, rigorous human oversight and evaluation remain critical safeguards.

4 Conclusion

The findings of this exploratory study offer food for thought for deploying LLMs in defining and accessing complex domain knowledge. The methods we devised for the purpose of defining new legal concepts can be replicated and applied to other domains as well. In the application of AI-based tools, it is critical to address vital risks from the outset, most notably the need for verification and evaluation. Even though LLMs excel at natural language processing tasks due to their capability to learn the context of the text (Malenšek et al., 2025), they are probabilistic and make mistakes, thus underscoring the need for critical evaluation of their output (Barryman & Ziegler, 2025, p. 81). Irrespective of the small scale of this study, it marks a significant research step by tackling a relevant pedagogical and information retrieval topic in the age of generative AI, and proposes a novel methodology to capture and define terms in the constantly changing and growing digital environment putting users at the frontline of AI application. As citizens today rely on tools to learn about domain concepts and access information about their rights and duties, there is a need for definitions of new concepts, as well as for conceptual and terminological clarity of new domains. However, to successfully

navigate digital sources and tools users need to harness both (contextualized) AI literacy and terminological competence so to critically evaluate the generated content. Future research should compare several AI tools and incorporate large-scale quantitative studies as well as semi-structured interviews with users to better understand their experiences and their needs for new skills when using these tools to access legal knowledge.

References

Bajčić, M. (2017). *New Insights into the Semantics of Legal Concepts and the Legal Dictionary*. John Benjamins.

Bajčić, M. (2024). Terminološka varijacija u pravnim tekstovima Europske unije: Dinamičnost nasuprot dosljednosti pravnoga jezika [Terminological variation in EU legal texts: Dynamicity vs. consistency of legal language]. *Fluminensia: časopis za filološka istraživanja*, 36(2), 375–409. https://doi.org/10.31820/f.36.2.3

Callister, P. D. (2025). Generative AI and finding the law. *Law Library Journal*, 117(1), 1–62. http://dx.doi.org/10.2139/ssrn.4608268

Barryman, J. & Ziegler, A. (2025). *Prompt Engineering for LLMs. The Art and Science of Building Large Language Model-Based Applications*. O'Reilly.

Biel, Ł. (2009). Organization of background knowledge structures in legal language and related translation problems. *Comparative Legilinguistics: International Journal for Legal Communication*, 1, 176–189. https://doi.org/10.14746/cl.2009.01.13

ISO 860 (2007). *Terminology work — Harmonization of concepts and terms*. International Organization for Standardization.

Kaguera, K. (2022). Terminological growth. In P. Faber & M.-C. L'Homme (Eds.), *Theoretical perspectives on terminology: Explaining terms, concepts and specialized knowledge* (pp. 457–476). John Benjamins.

Magesh, V., Surani, F., Dahl, M., Suzgun, M., Manning, C. D., Ho, D. E. (2025). Hallucination-free? Assessing the reliability of leading AI legal research tools. Journal of Empirical Legal Studies. Advance online publication. https://doi.org/10.1111/jels.12411

Malenšek, M., Završnik, A., Krajnc, S., Križnar, P., Bajec, M., Žitnik S. (2025) Towards Contradiction Detection in Legal Texts: Corpus Preparation and Contradiction Extraction. Slovenščina 2.0, 13(2), 179–209. https://doi.org/10.4312/slo2.0.2025.2.179-209

MW: Merriam Webster. Retrieved May 25, 2026, from https://www.merriam-webster.com/

Nahod, B. (2024). Can we substitute field experts with customized large language model in processing specialized languages? – A case study. In K. Š. Despot, A. Ostroški Anić & I. Brač (Eds.), *Proceedings of the XXI EURALEX International Congress* (p. 725–739). European Association for Lexicography.

Silalahi, S., Ahmad, T. & Studiawan, H. (2026). Towards terms definition extraction in drone forensics using large language model. *Intelligent Systems with Applications*, 30, Article 200655. https://doi.org/10.1016/j.iswa.2026.200655

Tricomi, P. P. (2024). *Advancing Social Network Analytics: Resilience and Security*. [Doctoral dissertation, University of Padua]. Padua Research Archive.

Zilli, L., Corazza, M., Palmirani, M. & Sapienza, S. (2025). In R. Markovich et al. (Eds.), *Multilingual Legislative Definitions Retrieval and Generation Using LLM and Agentic AI. Legal Knowledge and Information Systems JURIX 2025: The Thirty-eighth Annual Conference* (pp. 251–262). SAGE Publications.

Contact information

Martina Bajčić
University of Rijeka, Faculty of Law
martina.bajcic@uniri.hr

Ana Ostroški Anić
Institute for the Croatian Language
aostrosk@ihjj.hr

Acknowledgements

This work was supported by the Croatian Science Foundation (Grant No. IP-2025-02-3140 GenAI-LT) and as part of the project *Semantic Frames in the Croatian Language* (SEF), funded by the European Union - NextGenerationEU.

Riccardo Billero

LEXICOGRAPHY IN THE AGE OF AI

Training Systems with Human Expertise in Art Terminology

Abstract The emergence of artificial intelligence represents the latest in a series of technological challenges that have periodically reshaped the relationship between human expertise and technological tools. This paper takes as its point of departure the conviction that AI cannot exist, function, or evolve independently of human intelligence; lexicography, as a discipline situated at the intersection of linguistic knowledge, cultural competence, and technological practice, is particularly well placed to explore how human expertise can guide AI systems rather than be displaced by them. Drawing on the resources of the *Lessico dei Beni Culturali* (LBC) project, a multilingual lexicographic initiative at the University of Florence covering Italian cultural heritage terminology across Italian and seven further languages, the paper investigates how human-curated reference materials and expert-designed prompts can systematically improve AI performance on culturally embedded art terms. Three representative lemmas (*cartone*, *grazia*, and *maniera*) were tested across three conditions, namely baseline, expert-prompted, and LBC-constrained, using three major AI systems (Claude Sonnet 4.6, Gemini 3.5 Flash, and GPT-5). The results confirm a consistent gradient: corpus evidence not only improves output quality but, in some cases, reveals lexicographic realities that expert prompting actively conceals, most strikingly the absence of a stable specialised equivalent for *maniera* in Spanish, which the LBC corpus exposes as a structural lexical gap rather than a straightforward translational choice. The study argues that the value of expert lexicographic knowledge does not diminish in AI-assisted workflows but becomes more concentrated, shifting from the direct production of entries to the design of the knowledge infrastructure within which AI systems can operate effectively.

Keywords Specialized lexicography; artificial intelligence; human expertise; art terminology; human-AI collaboration; cultural heritage; multilingual lexicography; LBC project

1 Introduction

In the 1980s, the first personal computers began to enter everyday life, marking a turning point in the relationship between humans and technology. Approximately a decade later, the introduction of the Internet further accelerated this transformation, profoundly reshaping communication, access to information, and knowledge production. Both developments generated fears concerning the progressive marginalisation of human agency and expertise in favour of machines. With hindsight, these concerns were not entirely unfounded: while technological progress has brought enormous benefits, it has also contributed to the automation of numerous tasks and, in many sectors, to significant changes in professional practice. The idea of anticipating technological change and preventing its more disruptive consequences has accordingly acquired central importance, and it is no coincidence that the relationship between computational tools and linguistic expertise has become an increasingly prominent theme at conferences such as EURALEX (Rees et al., 2024; Henrichsen, 2024).

Today, we are confronted with yet another technological challenge that is frequently perceived in similar terms: artificial intelligence. Much like the first personal computers and early Internet connections, AI is often viewed as a potential threat to professional competences built up over decades of disciplinary practice. To a certain extent, this perception is understandable, as AI systems are developing at an extremely rapid pace and are increasingly capable of performing tasks traditionally associated with human intelligence (Degaonkar et al., 2024).

However, such a view risks overlooking a fundamental aspect of artificial intelligence: AI cannot exist, function, or evolve independently of human intelligence. AI systems are still entirely

dependent on human input, whether in the form of data selection, system design, or the processes commonly referred to as training and instruction. Artificial intelligence does not learn autonomously; rather, it operates within constraints defined by human decisions, expertise, and interpretative frameworks. From this perspective, AI should not be seen exclusively as a threat but also as a resource and a challenge whose potential can be meaningfully exploited only in close and continuous interaction with human knowledge.

It is against this background that the present paper investigates how human intelligence can actively guide artificial intelligence in the development of applications for specialised lexicography. In general, lexicography represents a particularly relevant field for such investigation, situated as it is at the intersection of linguistic knowledge, cultural competence, and technological implementation (de Schryver, 2023). The role of the expert lexicographer as designer of both the knowledge base and the methodology for its exploitation has been extensively theorised in the literature (Atkins & Rundell, 2008; Fuertes-Olivera & Tarp, 2014); what remains less well understood is how this expertise can be operationalised to guide AI systems in ways that extend rather than bypass human judgment. While general-purpose AI systems have made considerable progress in handling everyday language, they continue to struggle with highly specialised domains characterised by strong culture-specific features and historically layered terminology (Leo, 2024). As a concrete and demanding test case, the paper focuses on the specialised lexicon of art, a domain that is not only highly technical but also deeply embedded in historical, cultural, and aesthetic traditions that current AI systems are ill-equipped to navigate without expert guidance (Khan et al., 2024).

To explore the limitations and possibilities of AI in this context, we draw on the resources developed within the *Lessico dei Beni Culturali* (LBC) project, a multilingual lexicographic initiative at the University of Florence devoted to the terminology of Italian cultural heritage. The three lemmas selected for this study (*cartone*, *grazia*, and *maniera*) are key terms of Renaissance art discourse, and of Vasarian lexicon in particular. As such, they exemplify precisely the intersection of linguistic, cultural, and translational complexity that has motivated previous work within the LBC research group: the polysemy and translation history of individual terms across the parallel corpora of the Vite (Ballestracci, 2023; Ballestracci et al., in press), the development of multilingual parallel corpora as tools for specialised lexicographic description (Flinz et al., 2024), and the broader methodological questions raised by corpus-based approaches to artistic vocabulary. Through the analysis of these three lemmas as a pilot study, we aim to obtain preliminary evidence useful for designing future phases of work in which AI-assisted methods can be systematically integrated into the LBC lexicographic workflow.

The paper is structured as follows: Section 2 presents the LBC project and its resources; Section 3 situates the study within current debates on AI and specialised lexicography; Section 4 describes the methodology of our comparative analysis; Section 5 presents and discusses the results; Section 6 draws conclusions and outlines directions for future work.

2 The *Lessico dei Beni Culturali* project

The *Lessico dei Beni Culturali* (LBC) – Lexicon of Cultural Heritage – project is based at the University of Florence and is conducted in collaboration with researchers from several Italian and international universities. Its aim is to address a significant gap in European lexicographic provision: the absence of dedicated multilingual tools for the terminology of Italian cultural heritage (Billero and Nicolás Martínez, 2017). This is a domain of enormous international relevance, given the global interest in Italian art, architecture, and museum collections, yet one that has remained remarkably underserved by existing lexicographic resources (Fuertes-Olivera & Tarp, 2014).

The project covers eight languages: Chinese, English, French, German, Italian, Portuguese, Russian, and Spanish. Its lexicographic scope encompasses three main categories of terms, each posing distinct challenges for description and translation. The first comprises common nouns of art, that is technical terms denoting techniques, materials, styles, and genres (*affresco*, *chiaroscuro*, *pietra serena*, *predella*), many of which originate in Italian and have been borrowed or calqued into other European languages, making Italian the primary source for the harmonisation of lexical usage in art history. The second category comprises proper nouns specific to Florence and its history: artists, patrons, place names, monuments, and titles of artworks. The third, and perhaps most theoretically interesting, consists of Florentine *realia*, comprising culturally bound items, including foods, festivals, and local practices, whose referents are unique to the Florentine context and resist direct translation into other languages.

What makes the LBC corpus particularly distinctive, however, is not only its thematic scope but the breadth of its diachronic range and the variety of genres it brings together. The corpus spans roughly five centuries of art-historical writing, from the foundational Renaissance treatises that established the vocabulary of Western art history, beginning with Leon Battista Alberti's *Della pittura* and *I dieci libri dell'architettura*, Leonardo da Vinci's *Trattato della pittura*, and, above all, Vasari's *Le Vite de' più eccellenti pittori, scultori e architettori* (1568), to nineteenth-century critical writing by Stendhal and Jacob Burckhardt, and the contemporary tourist guides to Florence published by Bonechi. This diachronic range is not merely a philological curiosity; it has direct consequences for the lexicographic challenges the project addresses. As Billero (2026) observes, the same term may carry substantially different denotations, connotations, and pragmatic values across registers and centuries: *maniera*, for instance, functions as a precise and theoretically loaded Vasarian aesthetic concept in sixteenth century treatises, but operates as a generic stylistic descriptor in contemporary tourist publications. This kind of register-conditioned semantic variation is precisely the sort of nuance that AI systems trained on general-purpose corpora are least equipped to handle.

From a technical standpoint, the corpus texts have undergone automatic linguistic annotation, including tokenisation, part-of-speech tagging, and lemmatisation, and domain-specific terms have been extracted and linked to lemma entries in the project's lexicon interface (Billero, 2020b; Billero et al., 2020), searchable via a local installation of NoSketchEngine at *corpora.lessicobeniculturali.net* (Farina et al., 2024). The corpus is available in Italian and six further languages, encompassing both texts originally composed in Italian and translations from other European languages, a multilingual structure that raises specific methodological questions for lexicographic analysis (Farina & Billero, 2018). A curated collection of KWIC concordances, selected and annotated by domain experts, is accessible at *lexicon.lessicobeniculturali.net* for four target languages, including Spanish, German, and French: the three languages on which the present study focuses.

The project has so far produced three interconnected resources: a searchable corpus (Billero et al., 2020), four lemma indexes for individual target languages, and a prototype dictionary currently under development (Farina & Flinz, 2020). The Spanish, German, and French lemma indexes are currently the most fully developed, comprising several hundred lemmas, each documented with corpus-based frequency data and contextual examples (Billero, 2020a). More recently, the LBC has been proposed (Billero, 2026) as a candidate for publication within the Linguistic Linked Open Data (LLOD) ecosystem (Khan et al., 2022), with a roadmap for its encoding in the OntoLex-Lemon framework (McCrae et al., 2017) and alignment with resources such as the Getty Art and Architecture Thesaurus, the Getty Union List of Artist Names, and Wikidata, a step that would embed the LBC within a rich and interoperable semantic infrastructure comparable to that developed for Latin in the LiLa Knowledge Base (Passarotti et al., 2024), and that would position it alongside recent linked lexicographic initiatives addressing

cultural heritage terminology in multilingual contexts (Khan et al., 2024). It is this combination of expert-curated lexicographic data, extensive diachronic and multilingual coverage, and structured digital infrastructure that makes the LBC an exceptionally well-suited testbed for the investigation we present here.

3 AI and specialised lexicography: Opportunities and challenges

The application of AI to lexicographic tasks has attracted growing scholarly attention in recent years. General-purpose large language models (LLMs) have demonstrated considerable capabilities across a range of language-related tasks, including definition generation, synonym identification, and cross-lingual transfer (de Schryver, 2023; Rees et al., 2024). For everyday vocabulary and frequent collocations, AI-generated outputs can achieve a quality broadly comparable to human-written entries, particularly when evaluated by non-specialist users. It is tempting, on this basis, to conclude that AI might soon be capable of automating significant portions of lexicographic work.

The picture changes substantially, however, when AI systems encounter the terminology of specialised domains. The lexicon of art and cultural heritage presents a set of challenges that are particularly acute and that illuminate, in an especially clear way, the limitations of AI systems operating without expert guidance.

Consider, first, the problem of historical layering: a word such as *maniera* does not simply denote "style." In its primary art-historical usage, it refers to a specific aesthetic category whose meaning evolved between centuries, with connotations that differ across European critical traditions. A system trained on general-purpose text data will inevitably flatten these distinctions and, as the LBC corpus reveals, may not find a stable specialised equivalent in the target language: *maniera* in its Vasarian sense has no single-word equivalent in languages such as Spanish or German, requiring explanatory paraphrase rather than terminological rendering, a fact only corpus evidence can establish.

Consider, second, the problem of cultural embeddedness. Numerous art terms are so deeply rooted in a specific cultural and historical context that they resist translation in any direct sense. The Florentine *realia* that the LBC project has taken as one of its central concerns, such as *pietra serena* or *commesso di pietre dure*, refer to objects and practices so specific to their context that equivalent terms in other languages either do not exist or function as mere approximations, carrying different connotations and different degrees of cultural specificity.

Third, many art terms operate as false friends across European languages: *cartone* in Italian art-historical discourse refers to a full-scale preparatory drawing, as in the celebrated cartoons produced by Leonardo and Michelangelo; yet AI systems trained on general-purpose corpora routinely conflate this meaning with the common everyday sense of "cardboard" or "cardboard box" (Khan et al., 2024; Leo, 2024). This is not a trivial error; it strikes at the heart of the term's art-historical meaning and renders the output useless for lexicographic purposes.

These difficulties are not incidental features of art terminology but reflect a structural characteristic of specialised language more broadly: meaning in a specialised domain is constituted not only by the words themselves but by the disciplinary traditions, historical contexts, and interpretive communities within which they circulate. General-purpose AI systems, trained primarily on everyday language data, simply lack the contextual scaffolding required to navigate this complexity reliably. The critical question, then, is not whether AI can be applied to

specialised lexicography: it clearly can, but under what conditions, and with what forms of human guidance, it can do so in a way that is genuinely useful.

4 Methodology

4.1 Term selection

Three lemmas were selected from the LBC Spanish, German, and French lemma indexes for comparative analysis. The selection was designed to represent distinct types of lexicographic challenge, so that the study could illuminate not merely a single type of AI difficulty but a range of failure modes and their corresponding solutions.

Cartone (Spanish: *cartón*; German: *Karton*; French: *carton*). In Italian art-historical discourse, *cartone* denotes a full-scale preparatory drawing, such as those produced by Leonardo and Michelangelo in preparation for their major works. In everyday usage, however, and therefore in the general-purpose corpora on which AI systems are trained, *cartone* primarily means "cardboard" or "cardboard box". The Spanish *cartón* (32 corpus occurrences in the LBC), the German *Karton* (70 occurrences), and the French *carton* (167 occurrences) all exhibit the same structural ambiguity, with the additional complication that French also attests the Italianate form *cartone* alongside *carton*. This lemma thus allows us to test whether AI systems can distinguish the technical art-historical meaning from the dominant everyday one.

Grazia (Spanish: *gracia*; German: *Anmut*; French: *grâce*). A central aesthetic concept in Renaissance art theory, *grazia* denotes a quality of effortless beauty and natural elegance that writers from Castiglione to Vasari considered the highest achievement of artistic mastery, something that cannot be acquired through technical skill alone but must appear as a gift of nature. The Spanish *gracia* (181 LBC occurrences) and the French *grâce* (472 occurrences) maintain a degree of semantic proximity to the Italian, but the German *Anmut* (155 occurrences) carries different connotations, more corporeal and less spiritually inflected. This lemma allows us to examine how AI systems handle culturally embedded aesthetic concepts that resist mechanical equivalence.

Maniera (Spanish: *estilo/à la manera*; German: *Manier*; French: *manière*). Perhaps the most conceptually rich of the three lemmas, *maniera* is a cornerstone of Renaissance and Mannerist art criticism, denoting in Vasari's usage a specific aesthetic category that encompasses both individual artistic style and broader period characteristics. What makes this lemma particularly revealing for our purposes is the profoundly different fate it meets in the three target languages of the LBC corpus. In German, *Manier* (333 corpus occurrences) functions as a recognizable technical term that preserves the historical specificity of the Italian concept. In French, *manière* (794 occurrences), the transparent cognate, performs the same function. In Spanish, however, the corpus reveals a striking absence: rather than developing a stable specialized equivalent for the Vasarian concept of *maniera*, the Spanish translations consistently absorb it into the generic term *estilo* (368 occurrences), which serves as the default rendering for any notion of 'style' across the corpus : architectural styles (*estilo gótico*, *estilo dórico*) and individual artistic styles (*estilo de Giotto*, *estilo de Rafael*) and literary style alike. The locution *à la manera* (15 occurrences) appears as an alternative that preserves the Italian form but remains marginal. This pattern is not a deliberate translatorial choice but a structural lexical gap: *maniera* in its Vasarian sense has no stable specialized equivalent in Spanish. An AI system operating without access to the LBC corpus would almost certainly propose *manera* as the Spanish equivalent, the transparent cognate, without any awareness that the corpus attests neither *manera* nor any other term as a

stable technical equivalent, and that the concept is effectively untranslatable in the target language without explanatory paraphrase.

4.2 Experimental conditions

Each lemma was tested under three conditions, designed to measure the incremental effect of human expertise on the quality of AI output. The conditions were conceived to replicate, in a controlled way, three real-world scenarios that a lexicographer might encounter when working with AI tools.

Condition 1: Baseline: The AI system was given a minimal, domain-neutral prompt, asking simply for a definition and translation of the lemma into Spanish, German, and French, without any additional contextual guidance. This condition simulates the experience of a non-specialist user approaching the system without domain knowledge.

Condition 2: Expert-prompted: The same request was accompanied by an expert-designed prompt that specified the art-historical domain, the relevant historical period (the Italian Renaissance), and the key semantic distinctions to be observed, for instance the difference between the technical and everyday meanings of *cartone*, or the asymmetry between cognate and non-cognate translation choices for *maniera*. This condition simulates the work of a lexicographer who brings domain expertise to the interaction but does not yet have access to curated reference materials.

Condition 3: LBC-constrained: The AI was provided with curated reference material drawn directly from the LBC resources, including attested translations, corpus frequency data, and contextual examples drawn from the expert-annotated KWIC concordances. It was then asked to produce a lexicographic entry consistent with this evidence. This condition simulates the ideal scenario in which AI operates as a human-guided system, with expert-curated resources explicitly shaping its output.

The three AI systems selected for this study were Claude Sonnet 4.6 (Anthropic), Gemini 3.5 Flash (Google DeepMind), and GPT-5 (OpenAI), all queried via their respective web interfaces in May 2026. They were chosen as representative of the current frontier of publicly accessible LLMs, spanning three major developers and covering a range of architectural approaches. No fine-tuning or system-level customisation was applied; all interactions were conducted through standard user-facing interfaces, replicating the conditions under which a practising lexicographer would typically engage with AI tools.

4.3 Evaluation criteria

Outputs were evaluated by a domain expert against four criteria:

1. **Semantic accuracy:** does the output correctly identify the specialised meaning of the lemma?
2. **Translational appropriateness:** does the proposed equivalent reflect the choices made by expert translators in the LBC corpus?
3. **Contextual adequacy:** does the output provide sufficient historical and disciplinary context for the lemma to be usable in a lexicographic resource?

4. **Consistency with LBC resources:** does the output align with the corpus-attested usage documented in the LBC lemma indexes and concordances?

These four criteria were selected on the basis of their direct relevance to the lexicographic tasks at stake: semantic accuracy and translational appropriateness address the core lexicographic output, while contextual adequacy and consistency with LBC resources reflect the specific requirements of corpus-based specialised lexicography and the need for alignment with existing LBC documentation.

4.4 Limitations

The present study is best understood as a proof of concept rather than a comprehensive empirical survey. Several limitations should be acknowledged explicitly before the results are interpreted.

First, the lemma sample is intentionally small. Three lemmas were selected to represent distinct types of lexicographic challenge, namely false-friend disambiguation, culturally embedded aesthetic concepts, and structural lexical gaps, and the selection was guided by theoretical considerations rather than statistical sampling. While this approach allows for depth of analysis, it does not support broad generalisations about AI behaviour across the full range of art-historical vocabulary, let alone across specialised lexicography as a whole. The gradient proposed in Section 5.4 should accordingly be read as an interpretive hypothesis grounded in the three cases examined, and one that will require testing against a larger and more systematically sampled term set in future work.

Second, the language coverage, while multilingual, remains partial. The LBC project encompasses eight languages in total; the present study focuses on Spanish, German, and French, which are the three target languages for which the LBC lemma indexes and expert-annotated concordances are most fully developed. Results may differ for language pairs in which the LBC corpus is less densely documented, or for languages that are more typologically distant from Italian.

Third, the evaluation of AI outputs, although conducted against explicit criteria by a domain expert, remains essentially qualitative. A single evaluator introduces the possibility of interpretive bias, and the criteria applied, while principled, have not been subjected to inter-rater reliability testing. This is a significant methodological limitation. The findings reported in Section 5 should accordingly be read as the interpretive judgements of a single domain expert rather than as objectively verifiable scores, and the patterns identified, while consistent across the three lemmas and three AI systems, cannot be considered established until replicated by independent evaluators working with the same materials and criteria. Future work should involve multiple evaluators and, where feasible, quantitative scoring protocols.

Fourth, and most importantly from a reproducibility standpoint, LLMs are subject to continuous updating, and outputs may vary across model versions, query dates, and interface configurations. All queries reported here were conducted using Claude Sonnet 4.6, Gemini 3.5 Flash, and GPT-5 in May 2026; results obtained with later versions of the same systems may differ, possibly substantially. This is an inherent limitation of empirical work involving commercial AI systems, and one that the field has not yet resolved satisfactorily. We flag it here not as a reason to discount the findings but as a call for methodological transparency that we hope others working in this area will adopt.

Notwithstanding these limitations, the patterns observed across the three lemmas and three conditions are sufficiently consistent, both internally and across the three AI systems tested, to support the central argument of the paper: that expert-curated lexicographic resources make a measurable and qualitatively significant difference to the output of AI systems operating in specialised domains. These limitations should therefore be understood as defining the scope of the present investigation rather than undermining its findings.

The present pilot study also aims to contribute to the broader conversation within the EURALEX community on the role of digital resources and AI tools in lexicographic practice, and an invitation to other specialised lexicography projects to test analogous frameworks against their own curated corpora. Whether the patterns observed here prove specific to art-historical vocabulary or reflect more general properties of AI systems when operating in culturally embedded specialised domains is a question that only comparative replication across different fields can answer.

5 Results and Discussion

5.1 Cartone

Under Condition 1, all three systems (Claude Sonnet 4.6, Gemini 3.5 Flash, and GPT-5) produced broadly similar outputs, and the pattern was instructive. None of the systems ignored the art-historical meaning of *cartone* entirely, but all three relegated it to a position of secondary importance. The typical output opened with the everyday meaning ("cardboard," "cardboard box," "carton") and listed the art-historical sense as one of several possible meanings, alongside *cartone animato* ("cartoon"). Crucially, the translations offered in the baseline condition were identical across contexts: *cartón*, *Karton*, and *carton* were proposed as equivalents without any disambiguation between the everyday material and the Renaissance drawing. The result was technically accurate but lexicographically inadequate: a user consulting a specialist dictionary on the basis of these outputs would have no means of understanding that, in art-historical discourse, the lemma denotes exclusively the full-scale preparatory drawing.

Under Condition 2, the improvement was substantial and consistent across all three systems. Each produced a technically precise definition that correctly identified the art-historical meaning as primary in the specialist domain, described the function of the *cartone* in Renaissance studio practice (with reference to the *spolvero* technique and the works of Michelangelo and Raphael), and offered a justified translational recommendation. The outputs were richer and more contextually appropriate than those of Condition 1, though differences in depth and precision were visible between systems: Claude's output was the most terminologically detailed, drawing on a wider range of art-historical sources and explicitly distinguishing the *cartone* from related preparatory formats such as the *modello* and the *schizzo*. Gemini 3.5 Flash and GPT-5 produced solid but slightly less differentiated accounts.

Under Condition 3, all three systems anchored their outputs directly to the LBC corpus evidence, with notable convergence in both the definition offered and the translational choices made. The corpus concordances proved particularly effective in this case because they demonstrated, without any explicit instruction, that *cartone* in art-historical context appears systematically in passages of high artistic prestige: Michelangelo's cartoon for the *Battle of Cascina* is described as studied by generations of artists; working without *cartons* is characterised, in the French source, as a "méthode déplorable". All three systems correctly foregrounded the art-historical sense, provided corpus-grounded justifications for the equivalents *cartón*/*Karton*/*carton*, and produced entries that were directly usable as a basis for LBC lexicographic work.

5.2 Grazia

The *grazia* case revealed a subtler and, in some ways, more instructive pattern of AI difficulty. Under Condition 1, no system confused *grazia* with an unrelated meaning: all three recognised it as an aesthetic concept and identified the relevant translational range (*gracia*, *Anmut/Grazie*, *grâce*). The problem with the baseline outputs was not misidentification but semantic flatness. All three systems conflated the art-historical sense of *grazia* as a theoretically loaded Renaissance aesthetic category with the general sense of physical elegance or charm. The crucial distinction between *bellezza* (beauty as measurable proportion) and *grazia* (beauty as ineffable quality, linked to *sprezzatura* and to the Neoplatonic tradition) was absent from all three baseline outputs. The German equivalents were particularly revealing: Claude proposed both *Anmut* and *Grazie* without distinguishing between them, while GPT-5 and Gemini 3.5 Flash diverged, one preferring *Anmut* and the other *Grazie*, without either justifying the choice.

Under Condition 2, the theoretical dimension of *grazia* was handled considerably better by all three systems. The connection to Castiglione's *sprezzatura*, the contrast with *bellezza*, and the role of *grazia* in Vasari's critical vocabulary were all identified. For the German equivalent, both Claude and Gemini settled on *Anmut* as the preferred art-historical term, noting its later elaboration by Schiller; GPT-5 preferred *Grazie* as the more historically precise borrowing. This divergence is itself a significant finding: even with expert prompting, the systems reached different conclusions on a lexicographic question, which is precisely the kind of problem that corpus evidence is needed to resolve.

Under Condition 3, the LBC concordances provided the decisive input. The systematic co-occurrence of *Anmut* with *Weichheit*, *himmlisch*, and *Kolorit* in the German corpus confirmed *Anmut* as the attested equivalent in the translation tradition the LBC is documenting. In Spanish, the corpus consistently attests *gracia* in the precise aesthetic sense, in close association with *esbeltez*, *belleza*, and *proporción*. In French, *grâce* appears paired with *beauté*, *douceur*, and *vivacité* in a semantic cluster that closely mirrors the Italian. All three systems, when provided with this evidence, produced entries that correctly identified the translational choices and grounded their justifications in corpus usage rather than in general lexicographic knowledge.

5.3 Maniera

The *maniera* case is the most revealing of the three and the one that most clearly demonstrates what corpus evidence can contribute that expert prompting alone cannot. Under Condition 1, all three systems proposed *manera* as the Spanish equivalent, *Manier* for German, and *manière* for French. The German and French choices are broadly correct; the Spanish choice is the transparent cognate, and it is the one that the LBC corpus does not support as a specialised term.

Under Condition 2, the situation improved for German and French, where all three systems produced well-developed entries acknowledging the historiographic weight of the term and its relationship to the broader category of Mannerism. For Spanish, however, the expert-prompted outputs remained anchored to *manera* as the primary equivalent, with *estilo* mentioned as a possible alternative. The LBC corpus evidence makes clear that the relationship is reversed: *estilo* is the dominant form in the corpus (368 occurrences) and *manera* is marginal, while *à la manera* functions as a loan-locution in specific adverbial constructions. No system, operating from expert prompting alone, could conclude that Spanish presents a structural lexical gap rather than a straightforward terminological choice.

Under Condition 3, the corpus evidence produced a qualitatively different kind of output. All three systems correctly identified the lexical gap, distinguished between *estilo* as the generic default and *à la manera* as the contextually marked loan-form, and acknowledged that the Vasarian concept of *maniera* is not fully lexicalized in Spanish. Claude Sonnet 4.6 and GPT-5 both recommended that specialist dictionaries consider retaining *maniera* as an unassimilated technical borrowing, on the model of terms such as *contrapposto* or *sfumato*, which Spanish art history preserves in their Italian form. This is a recommendation that no corpus-naive system could have generated, since it requires knowing both that *manera* is the expected cognate and that the corpus avoids it in this sense.

5.4 Discussion

Taken together, the three cases support a clear and consistent conclusion: the quality of AI output in specialised lexicographic tasks is directly proportional to the quality and specificity of the human guidance provided. This is not a trivial finding. It suggests that the value of expert lexicographic knowledge does not diminish in an AI-assisted workflow; on the contrary, it becomes more concentrated and more decisive. What changes is not the importance of expertise but its point of application. Rather than producing lexicographic entries directly, the expert lexicographer designs the knowledge scaffolding: the selection of materials, the construction of prompts, the curation of reference resources within which AI systems can operate effectively.

The three systems compared here (Claude Sonnet 4.6, Gemini 3.5 Flash, and GPT-5) performed broadly similarly across the three conditions, with differences of degree rather than kind. Under Condition 1, all three produced outputs adequate for general-purpose use but insufficient for specialist lexicographic purposes. Under Condition 2, all three improved substantially, with the most visible gains in terminological precision and historical contextualisation. Under Condition 3, all three produced outputs that a specialist would recognise as lexicographically grounded, with the corpus concordances providing the decisive evidence that neither baseline queries nor expert prompting could supply. The convergence of the three systems under Condition 3, and their divergence under Condition 2, is significant: it suggests that the LBC corpus evidence functions as a stabilising constraint, reducing the system-to-system variability that expert prompting alone does not eliminate.

A further observation concerns the differential value of corpus evidence across the three lemma types. For *cartone*, corpus evidence confirmed and grounded a distinction that expert prompting had already captured. For *grazia*, corpus evidence resolved a genuine ambiguity (*Anmut* versus *Grazie* in German) that expert prompting had left open. For *maniera*, corpus evidence revealed a structural lexical gap that expert prompting had concealed, by directing systems toward the transparent cognate *manera* rather than toward the more complex and accurate account that the corpus makes possible. This gradient, from confirmation through disambiguation to revelation, maps directly onto a practical typology of lexicographic tasks for which AI systems require greater corpus-grounded guidance.

6 Conclusions

The relationship between AI and specialised lexicography is best understood not as one of replacement but of structured and necessarily asymmetrical collaboration. AI systems operating without domain-specific guidance produce outputs that, while adequate for common vocabulary, are systematically inadequate, or at least, as the present study suggests, prone to inadequacy, for the culturally embedded, historically layered terminology of domains such as art and cultural heritage. The reasons for this inadequacy are structural: they reflect the

fundamental dependence of specialised meaning on disciplinary traditions and interpretive contexts that general-purpose training data cannot capture. The LBC project offers a model, and a concrete set of resources, for how human expertise can be encoded in lexicographic infrastructure that AI systems can then exploit effectively. The lemma indexes, the expert-annotated KWIC concordances, and the ongoing LLOD publication roadmap collectively represent a form of codified expert knowledge that, as our results suggest, makes a measurable and significant difference to the quality of AI-generated lexicographic output.

The three-condition experimental framework presented here, namely baseline, expert-prompted, and LBC-constrained, provides a replicable methodology for investigating this interaction across different lemma types, language pairs, and specialised domains. Our results suggest that expert-designed prompts consistently improve AI performance, and that the most significant gains occur when AI systems access corpus-attested evidence rather than abstract domain descriptions alone. Whether this pattern holds across a broader range of term types and language pairs remains a question for future investigation, but the consistency of the finding across three distinct AI systems suggests preliminary robustness. From the perspective of lexicographic practice, this means that the future of AI-assisted lexicography lies not in the replacement of human expertise but in its transformation: the expert lexicographer becomes, above all, the designer of the human-guided systems through which AI can be made to serve the needs of specialised lexicographic work.

Future work will extend this analysis to the full range of LBC language pairs, explore the integration of the LLOD-modelled LBC corpus with AI systems via RAG pipelines, and investigate the potential of AI tools for semi-automated updating and expansion of the LBC dictionary. More broadly, this study contributes to an emerging framework for human-guided AI lexicography, one in which technological capability and human linguistic and cultural competence are understood not as competing forces but as complementary resources, each indispensable to the other.

References

Atkins, B. T. S., & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press.

Ballestracci, S. (2023). Da un segno a tanti segni. L'emergere della polisemia del termine "disegno" nelle traduzioni tedesche delle Vite di Vasari. In V. Zotti & M. Turci (Eds.), *Nuove strategie per la traduzione del lessico artistico: da Giorgio Vasari a un corpus plurilingue dei beni culturali* (pp. 21–38). Firenze University Press.

Ballestracci, S., Flinz, C., Palilla, G., Giri, G., & Billero, R. (in press). Le Vite del Vasari nel corpus LBC italiano-tedesco: i corpora paralleli per la linguistica e la traduzione specialistica. In F. Funari, A. Farina, & D. Henkel (Eds.), *Giorgio Vasari, langue des arts et traduction*. Firenze University Press.

Billero, R. (2020a). Cultural heritage lexicon: A case study. In A. Pano Alamán & V. Zotti (Eds.), *The Language of Art and Cultural Heritage: A Plurilingual and Digital Perspective* (pp. 86–103). Cambridge Scholars Publishing.

Billero, R. (2020b). Implementazione di software per la gestione dei corpora LBC. In R. Billero, A. Farina, & M. C. Nicolás Martínez (Eds.), *I Corpora LBC. Informatica Umanistica per il Lessico dei Beni Culturali* (pp. 19–32). Firenze University Press.

Billero, R. (2026). Towards a Linguistic Linked Open Data Resource for Italian Cultural Heritage: The Lessico dei Beni Culturali Corpus. In J. P. McCrae et al. (Eds.), *Proceedings of the 10th Workshop on Linked Data in Linguistics (LDL-2026) @ LREC 2026* (pp. 32–38). ELRA.

Billero, R., Farina, A., & Nicolás Martínez, M. C. (2020). *I Corpora LBC. Informatica Umanistica per il Lessico dei Beni Culturali*. Firenze University Press.

Billero, R., & Nicolás Martínez, M. C. (2017). Nuove risorse per la ricerca del lessico del patrimonio culturale: corpora multilingue LBC. *CHIMERA: Romance Corpora and Linguistic Studies*, 4(2), 203-216.

Degaonkar, S. V., Kadam, S., & Rane, N. M. (2024). Can AI Replace Human Intelligence? *International Journal of Advanced Research in Science, Communication and Technology*, 4(1), 114–119.

de Schryver, G.-M. (2023). Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography*, 36(4), 1–33.

Farina, A., Billero, R., & Nicolás Martínez, M. C. (2024). Presentation of the LBC database. In I. Natali (Ed.), *The LBC English Corpus*. Firenze University Press. E-ISBN: 979-12-215-0307-4. https://corpora.lessicobeniculturali.net/en/.

Farina, A., & Billero, R. (2018). Comparaison de corpus de langue "naturelle" et de langue "de traduction": les bases de données textuelles LBC, un outil essentiel pour la création de fiches lexicographiques bilingues. In D. F. Iezzi, L. Celardo, M. Misuraca (Eds.), *Proceedings of the 14th International Conference on Statistical Analysis of Textual Data (JADT 2018)* (pp. 108–116). UniversItalia.

Farina, A., & Flinz, C. (2020). LBC-Dictionary: A multilingual cultural heritage dictionary. Data collection and data preparation. In Z. Gavriilidou, M. Mitsiaki, & A. Fliatouras (Eds.), *Lexicography for Inclusion. Proceedings of the XIX EURALEX Congress, Vol. 1* (pp. 371–379). Ljubljana University Press.

Flinz, C., Henkel, D., Zotti, V., & Ballestracci, S. (2024). A multilingual parallel corpus for the lexical information system LBC: Recent progress and future perspectives. In K. Š. Despot, A. Ostroški, & I. Brač (Eds.), *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress* (pp. 316–328). Institute for the Croatian Language.

Fuertes-Olivera, P. A., & Tarp, S. (2014). *Theory and Practice of Specialised Online Dictionaries: Lexicography versus Terminography*. De Gruyter.

Henrichsen, P. (2024). Make Each Morph Count. A New Approach to Computational Lexicography for Text Processing. In K. Š. Despot, A. Ostroški, & I. Brač (Eds.), *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress* (pp. 349–355). Institute for the Croatian Language.

Khan, A. F., Chiarcos, C., Declerck, T., Gifu, D., González-Blanco García, E., Gracia, J., Ionov, M., Labropoulou, P., Mambrini, F., McCrae, J. P., Pagé-Perron, É., Passarotti, M., Ros Muñoz, S., & Truïcă, C.-O. (2022). When linguistics meets web technologies: Recent advances in modelling linguistic linked data. *Semantic Web*, 13(6), 987–1050.

Khan, A. F., Castro Salgado, A., McCrae, J. P., Liyange, C., Rani, P., Costa, R., Anuradha, I., Kr. Ojha, A., & Frontini, F. (2024). Cultural HeritAge and Multilingual Understanding through lexiCal Archives (CHAMUÇA). In K. Š. Despot, A. Ostroški, & I. Brač (Eds.), *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress* (pp. 475–484). Institute for the Croatian Language.

Leo, H. Y. (2024). Do Indonesian Speakers Feel 'pain'? NSM and Corpus-Based Approach to the Cross-Linguistic Concept of 'pain' in Bahasa Indonesia. In K. Š. Despot, A. Ostroški, & I. Brač (Eds.), *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress* (pp. 109–120). Institute for the Croatian Language.

McCrae, J. P., Bosque-Gil, J., Gracia, J., Buitelaar, P., & Cimiano, P. (2017). The OntoLex-Lemon model: Development and applications. In I. Kosem et al. (Eds.), *Electronic Lexicography in the 21st Century. Proceedings of the eLex 2017 Conference* (pp. 587–597). Lexical Computing.

Passarotti, M., Mambrini, F., & Moretti, G. (2024). The Services of the LiLa Knowledge Base of Interoperable Linguistic Resources for Latin. In C. Chiarcos, K. Gkirtzou, M. Ionov, F. Khan, J. P. McCrae, E. Montiel Ponsoda, & P. Martín Chozas (Eds.), *Proceedings of the 9th Workshop on Linked Data in Linguistics @ LREC-COLING 2024* (pp. 75–83). ELRA and ICCL.

Rees, G., & Gibert, I. (2024). A Textbook or ChatGPT: Which Helps Novice Programmers Most with Unknown Terms? In K. Š. Despot, A. Ostroški, & I. Brač (Eds.), *Lexicography and Semantics. Proceedings of the XXI EURALEX International Congress* (pp. 265–273). Institute for the Croatian Language.

Contact information

Riccardo Billero
Free University of Bozen
riccardo.billero@unibz.it

Acknowledgments

I would like to express my sincere gratitude to Carolina Flinz, whose encouragement first prompted me to submit this work to the XXII EURALEX International Congress 2026 and without whom I would never have done so. I am also grateful to Sabrina Ballestracci for her careful reading of the manuscript, whose comments greatly improved the final text. Warm thanks go to all the members of the *Lessico dei Beni Culturali* project, whose years of dedicated work in building the lexicographic infrastructure have made this research possible.

Marek Blahuš, Miloš Jakubíček,
Katarína Petreková & Yingtian Wu

AUTOMATIC DRAFTING OF TERMINOLOGICAL DICTIONARIES WITH ONE-CLICK DICTIONARY FOR 34 LANGUAGES

Abstract We present an evaluation of a system for automatic drafting of terminological dictionaries using the One-Click Dictionary feature of Sketch Engine. We bootstrapped five small corpora in Arabic, Czech, English, German and Spanish and evaluated the performance of the methodology on these five corpora yielding five terminological dictionaries for five different terminological domains; the system currently supports 34 languages in total.

Keywords Automatic dictionary drafting; terminology; Sketch Engine; Lexonomy; One–Click Dictionary

1 Introduction

Recent advances in NLP technologies including the availability of large language models significantly improved the quality of automatically created dictionaries. While general language dictionaries, both mono- and bilingual, still require significant amount of human post-editing to achieve publishing quality, terminological dictionaries represent a target much easier to achieve: their entries are (or at least should be) typically monosemous as each term should principally correspond to a single concept. This removes what is frequently seen as the most difficult lexicographic task: establishing on a word's sense division as well as crafting of further entry components such as collocations, examples or translations in a sense-specific way.

In this paper we present a processing pipeline available for 34 languages (Afrikaans, Arabic, Bosnian, Bulgarian, Chinese, Croatian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Italian, Japanese, Korean, Latvian, Lithuanian, Maori, Norwegian, Polish, Portuguese, Romanian, Russian, Serbian, Slovak, Slovenian, Spanish, Swedish, Ukrainian, Vietnamese) which allows users of the system to automatically create a dictionary draft from a reasonably large (typically just dozens of millions of words) domain corpus. The pipeline is integrated in the Sketch Engine corpus management system (Kilgarriff et al., 2014a) and consists of the following four steps:

1. building a domain corpus from users' texts or using web search
2. finding terms in the domain corpus using a language-specific term grammar
3. drafting a dictionary using the One-Click Dictionary feature
4. exporting the dictionary into the Lexonomy dictionary editing system

2 Building a domain corpus

The source evidence for the dictionary is gathered from a domain corpus supplied either by the user or created using web search. The former scenario is typical for large corporate users that

possess vast amount of internal documents describing their field and are looking for fast ways to create their own internal glossary. The latter option focuses on individual users or academic terminology and is based on the WebBootCaT technology (Baroni et al., 2006) which allows users to quickly gather texts from internet based on search seed words.

In our experiment we focused on the latter option and created five corpora for five domains. Languages were chosen with different language families in mind, domains were chosen according to the respective evaluators' subject knowledge. Details of the corpora can be found in Table 1.

TABLE 1 Description of corpora created in the experiment

Language	Size (tokens)	Domain
Arabic	11,124,898	football
Czech	15,763,412	book publishing
English	12,008,065	photography
German	6,544,280	automotive
Spanish	11,261,104	astronomy

We also provide an overview of the seed words used when crawling the web for corpus data using WebBootCaT. In accordance with the WebBootCaT best practices, a few seed words were identified for each language at the beginning, out of words that are typical of the domain (i.e., not necessarily of high frequency, but rarely used outside of the domain).

These seed words were used as the tool's input for the first pass, after which the process was iterated several more times, each time with a new set of seed words automatically discovered in the corpus data collected in the previous pass. This enabled us to grow the corpora to the desired sizes while keeping their content aligned with the selected domain. Following are the full lists of the used seed words:

- Arabic (86 seeds): *FIFA, أتلتيكو مدريد, أرسنال, أرسنال، يوفنتوس, أشرف حكيمي, الأهلي, الاتحاد, الاتحاد الآسيوي لكرة القدم, الاتحاد الأسيوي لكرة القدم, الاتحاد الأفريقي لكرة القدم, الاتحاد الإفريقي لكرة القدم, الاتحاد الدولي لكرة القدم, الترجي, الدوري الإسباني, الدوري الإنجليزي, الدوري السعودي, الدوري المصري, الدولي الإطالي, الدولي الإيطالي, الرجاء, الزمالك, الفيفا, المرمى, النصر, الهلال, الوداد, الوقت الإضافي, انتر ميلان, باريس سان جرمان, باريس سان جير مان, بايرن, بايرن ميونخ, برشلونة, بروسيا دورتموند, بطاقة الحمراء, بطاقة حمراء, بطاقة صفراء, بطولة كأس العالم, بوروسيا دورتموند, تسجيل هدف, تسديدة, تسلل, تشلسي, تشيلسي, جارسيا المرمى, حارس المرمى, حجم الأصلي للصورة, حكيم زياش, دوري أبطال, دوري أبطال أوروبا, رأسي, ربع النهائي, ركل, ركلة جزاء, ركلة حرة, ركلة ركنية, ركل حرة, رياض محرز, ريال مدريد, سادیو مانی, شوط أول, شوط ثاني, شوط ثاني الوقت الإضافية, كأس آسيا, كأس الأمم الأفريقية, كأس العالم, كأس العالم للأندية,كرة القدم, كريستيانو رونالدو, كريم بنزيمة, كيليان مبابي, ليفربول, ليونيل ميسي, مانشستر سيتي, مانشستر يونايتد, مبادرة كرة القدم, مباراة كرة القدم, مباشرة حر, محمد صلاح, منتخب المغرب, منطقة الجزاء, ميلان, ميونخ, نهاية الوقت, نيمار, هلال*

- Czech (59 seeds): *akvizice rukopisů, antikva, autotyp, bohemikum, copyright, čtverčík, distributor, ediční plán, formát, fraktura, gramáž papíru, hřbet knihy, iniciála, inkunabule, ISBN, kalandr, kaligrafický, kapitálek, kapitálka, knihkupec, knihtisk, knižní blok, knižní trh, knižní vazba, korektura, kurzíva, litera, maloobchodní cena knihy, marketingové strategie nakladatelství,*

marketingový mix, minuskula, minuskule, nakladatel, patitul, pečeť, písemnictví, pomlčka, produkce knižních titulů, propagace knihy, průměrná cena, průměrný náklad, prvotisk, předmluva, předsádka, signet, signum, spojovník, státní podpora knižního průmyslu, struktura knižního trhu, štoček, švabach, tiráž, tiskař, tištěná kniha, titul, trendy knižního trhu, typografický, vakát, verzálka

- English (194 seeds): *aperture, aperture priority, APS-C, artificial light, aspect ratio, astrophotography, autofocus, auto mode, backlighting, ball head, batch editing, black-and-white photography, blending, blue hour, blur, blurred background, blurring, bokeh, bracketing, brightness, built-in flash, bulb mode, camera, camera body, camera lifespan, card reader, cataloging, clone stamp, color photography, compact camera, composition, contrast, conversion, copyright, couple photography, cropping, data backup, data recording, depth of field, developer, digital camera, display, distracting background, DSLR, EXIF, exposure, exposure mode, exposure time, exposure triangle, exposure value, extender ring, external flash, family photo, filter, final photo, fixed focal length, fixed lens, fixer, focal length, focus point, foreground, freezing motion, glare, golden hour, gradation filter, gradient filter, graphics card, halogen bulb, hard drive, HDR, histogram, hyperfocal distance, chromatic aberration, image, image quality, image sensor, image stabilization, incandescent bulb, interface, interior, ISO, JPEG, landscape photography, lens, lens distortion, lens hood, light-sensitive, light source, live view, local editing, long exposure, macro lens, macro photography, magnetic filter, main subject, manual focus, manual mode, memory card, minimalist photography, mirrorless camera, monitor, monitor calibration, motion blur, mount, mount adapter, natural light, negative, neutral density filter, noise, noise reduction, open aperture, operating system, optical stabilization, outdoor photography, overexposed image, panning, panorama, parameter, perspective, photo atmosphere, photo background, photo download, photo export, photogenic, photograph, photographed scene, photographed subject, photography equipment, photography school, photo import, photo layers, photo management, photo processing, photo resolution, photo series, photo sorting, pixel, PNG, polarizing filter, portrait, preview, product photography, RAM, raw development, RAW format, raw photo, read speed, reflection, reflector, reportage photography, reproduction, retouching, retouching tool, RGB, ring light, self-timer, sensor, sensor sensitivity, shadow, sharpening, sharpness, shutter, shutter button, shutter priority, shutter release, shutter speed, smudging, softbox, softening light, software, sRGB, storage, street photography, studio, technical quality, telephoto lens, thematic photography, TIFF, time-lapse, tonality, tripod, tripod plate, underexposed image, USM, variable filter, video mode, viewfinder, vignetting, wedding photography, white balance, wide-angle lens, wildlife photography, zoom lens, 35mm film*

- German (37 seeds): *Abgasnorm, Akkufüllung, Assistenzsysteme, Außenspiegel, Direktschaltgetriebe, Einparkhilfe, Fahrgastraum, Fahrstil, Fahrzeugklasse, Familienauto, Federungskomfort, Fernlichtassistent, Feuchtigkeitssensor, Fünf-Liter-Marke, Geschwindigkeitsbegrenzer, Heckscheibe, Höchstgeschwindigkeit, Infotainmentsystem, Innenspiegel, Karosserie, Kofferraumvolumen, Kraftstoffverbrauch, Kurvenräuber, Limousine, Mittelkonsole, Motorabdeckung, Neuwagenpreis, Panoramadach, Plug-in-Hybrid, Reichweite, Rückfahrkamera, Serienausstattung, siebensitzig, Spurhalteassistent, Vierzylinder, Vorderrad, Winterreifen*

- Spanish (73 seeds): *agujero negro, agujero negro supermasivo, año luz, año luz de distancia, astronomía, astronómico, astrónomo, Betelgeuse, campo magnético, cinturón de asteroides, cometa, constelación, cuerpo celeste, cúmulo, cúmulo abierto, cúmulo de galaxias, cúmulo estelar, cúmulo globular, elemento más pesado, enano blanca, enano roja, energía oscura, esfera celeste, estrella, estrella cercana, estrella de neutrones, estrella masiva, exoplaneta, expansión del universo, explosión de supernova, formación estelar, fusión nuclear, galaxia, galaxia espiral, gigante roja, Hubble, interestelar, Júpiter, longitud de onda, luminosidad, luz de distancia, magnitud aparente, masa solar, materia oscura, NASA, nebulosa, nebulosa planetaria, neutrino, neutrón, NGC, nube de gas, nube de gas y polvo, objeto astronómico, planeta enano, planeta extrasolares, planeta terrestre, polvo interestelar, protoestrella, púlsar, radiación, rayo gamma, sistema binario, sistema*

estelar, sistema planetario, sistema solar, supergigante, supernova, telescopio, telescopio espacial, telescopio Hubble, universo observable, velocidad de la luz, velocidad de rotación

3 Term finding

Having domain corpus texts, by uploading them into Sketch Engine, each corpus gets automatically analyzed by a part-of-speech tagger and a lemmatizer. Such morphological annotation is used for finding candidate terms using a language specific term grammar – a morphosyntactic description of grammatically valid terms (typically noun phrases) in the form of corpus queries in the CQL language (Kilgarriff et al., 2014b).

The term grammars are developed to achieve high coverage on existing term bases such as IATE and so far there are term grammars for 34 languages, which is the present restriction on the languages supported. In One-Click Dictionary, the extracted terms are used as input (i.e., headwords) for further dictionary drafting. This is similar to how terminologists routinely use the same algorithm through a dedicated interface called OneClick Terms to extract terminology and then build term bases or glossaries out of it using specialized terminology management tools (Blahuš et al., 2023).

4 One-Click Dictionary

Finally, a functionality of Sketch Engine called One-Click Dictionary (Jakubíček et al., 2021) is used to create a dictionary draft automatically (with "one click") from a corpus. The dictionary headwords are sourced from the most frequent single- and multi-word terms in the source corpus as identified by the respective term grammar. Further entry components are generated from the source corpora:

- word forms; based on the automatic lemmatization of the corpus
- example sentences; based on the GDEX functionality of Sketch Engine
- collocations; based on the word sketch functionality of Sketch Engine
- thesaurus; based on the thesaurus functionality of Sketch Engine
- images; sourced from royalty-free databases such as Wikimedia Commons and Pixabay
- definitions; based on a large language model which takes all the other entry components (and the domain specification as well, of course) as input for consideration, leading to very high accuracy of the generated definitions. The system allows the user to choose from dozens of models provided by the OpenRouter platform. In this experiment, we used the GPT-5 model released in August 2025.

Each of the steps that sources data from the corpus comes with a settings allowing the user to select the number of example sentences, collocations, thesaurus items etc., including selection criteria as available elsewhere in the Sketch Engine corpus management system. Figure 1 shows the components used in our experiment.

As explained earlier, no word sense induction or disambiguation is performed as we build upon the assumption that a terminological entry should always correspond only to a single sense for a given fixed domain.

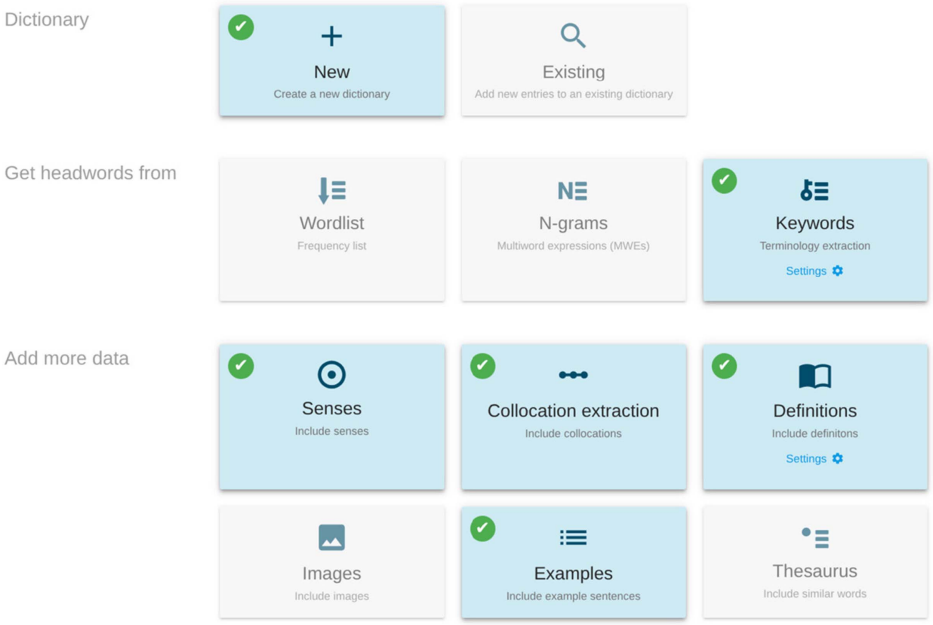


FIGURE 1 One-Click Dictionary user interface

The fully automatic dictionary drafting using One-Click Dictionary with the above settings takes less than 60 minutes for the first 1000 dictionary entries.

5 Lexonomy

The ready-made dictionary draft is made available in Lexonomy (Jakubíček et al., 2018), a web-based dictionary editor, where it can be reviewed, optionally post-edited (preserving access to the underlying corpus evidence in Sketch Engine) and published. In Figure 2 you can find a sample entry *exposure time* generated for the photography domain from the English corpus.

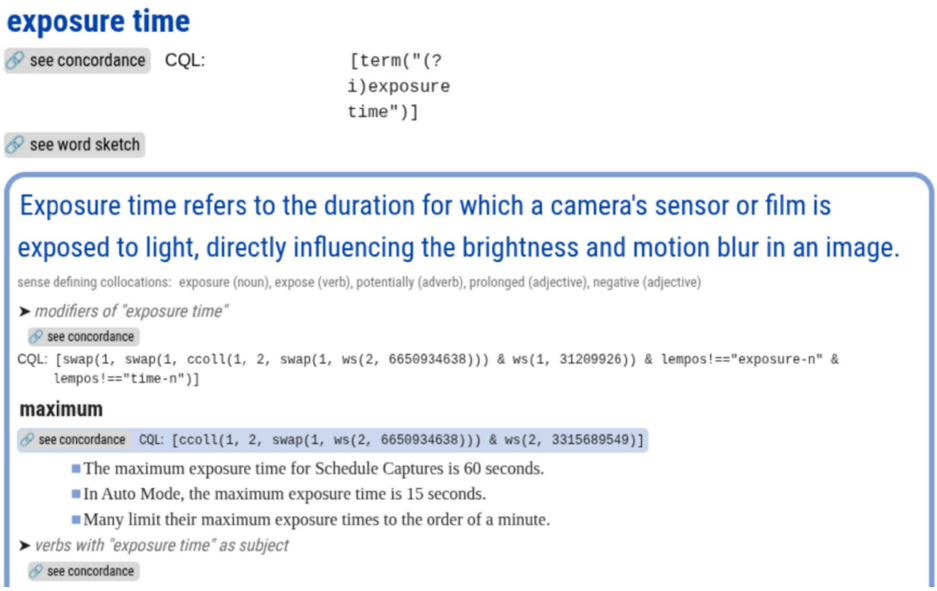


FIGURE 2 Sample English entry from the photography corpus

6 Evaluation

We evaluated the automatically created dictionary drafts as for the plausibility of headwords and accuracy of examples and definitions. Each dictionary was assessed by a single reviewer who was a proficient speaker of the respective language and domain expert for the respective topic. Only a hundred top-ranked terms of each dictionary have been evaluated. Table 2 presents results of the evaluation for the respective languages.

TABLE 2 Evaluation results assessing the quality of the dictionary drafts as for headwords, examples and definitions. The percentages for examples and definitions are calculated from plausible headwords only. Use the ID to access the corpus in Sketch Engine (in the Select corpus page, or in a URL of type https://app.sketchengine.eu/?corpname=preloaded/ID) or the dictionary in Lexonomy (in the list of public dictionaries, or in a URL of type https://www.lexonomy.eu/ID)

Language	Headword plausibility	Definition accuracy	Example accuracy	Corpus and dictionary ID
Arabic	88.0 %	97.7 %	91.1 %	euralex2026_ocd_ar
Czech	84.0 %	90.5 %	57.9 %	euralex2026_ocd_cs
English	97.0 %	96.9 %	71.1 %	euralex2026_ocd_en
German	88.0 %	96.6 %	79.4 %	euralex2026_ocd_de
Spanish	94.0 %	94.7 %	73.4 %	euralex2026_ocd_es
Average	*90.2 %*	*95.3 %*	*74.6 %*	

A headword was considered plausible if it was a domain term for the respective domain. Minor deviations from the expected citation form were tolerated, such as incorrect suffixes or wrong letter case, and proper nouns were accepted as terms as long as they identified a specific named entity in the domain. Semantic overlaps between terms, such as synonymy or hyponymy (e.g. the Spanish *agujero negro* 'black hole' vs. *agujero negro supermasivo* 'supermassive black hole'), were disregarded during evaluation.

A definition or example was considered accurate if the reviewer finds it suitable for dictionary inclusion without further changes. While the system tries to come up with multiple senses (and therefore multiple definitions) for each headword, we found it sufficient to consider only the first one, supported by the expected monosemy of terms.

The number of generated examples depends on the number of identified collocations. In order to balance evaluation costs and representativeness, we assessed three examples for each headword: the first one, the last one, and one from the middle of the list.

Public access is given to all the corpora used (in Sketch Engine) and the respective drafted dictionaries (in Lexonomy).

7 Conclusions and future work

In this paper we presented a minimal framework that demonstrates rapid development of terminological dictionaries using domain seed words, corpus bootstrapping and automatic dictionary drafting. The framework could be extended to additional languages and domains in a straightforward way, providing stable benchmarking for the task of automatic dictionary drafting for terminological dictionaries.

References

Baroni, M., Kilgarriff, A., Pomikálek, J. & Rychlý, P. (2006). WebBootCaT. Instant domain-specific corpora to support human translators. In *Proceedings of the 11th annual conference of the European association for machine translation*.

Blahuš, M., Jakubíček, M., Cukr, M., Kovář, V. & Suchomel, V. (2023). Development of evidence-based grammars for terminology extraction in OneClick Terms. In *Proceedings of the Electronic Lexicography in the 21st Century, Brno, Czech Republic* (pp. 27–29). Lexical Computing.

Jakubíček, M., Kovář, V. & Rychlý, P. (2021). *Million-click dictionary: Tools and methods for automatic dictionary drafting and post-editing*. In Z. Gavriilidou, L. Mitits & S. Kiosses (Eds.), *EURALEX XIX* (pp. 65–67). SynMorPhoSe Lab, Democritus University of Thrace.

Jakubíček, M., Měchura, M., Kovář, V. & Rychlý, P. (2018). Practical Post-editing Lexicography with Lexonomy and Sketch Engine. In J. Čibej, V. Gorjanc, I. Kosem & S. Krek (Eds.), *The XVIII EURALEX International Congress: Lexicography in Global Contexts. Book of Abstracts* (p. 65).

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014a). The Sketch Engine: ten years on. *Lexicography*, 1. https://doi.org/10.1007/s40607-014-0009-9

Kilgarriff, A., Jakubíček, M., Kovář, V., Rychlý, P. & Suchomel, V. (2014b). Finding terms in corpora for many languages with the Sketch Engine. In S. Wintner, M. Tadić & B. Babych (Eds.), *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 53-56).

Contact information

Marek Blahuš
Lexical Computing, Brno, Czech Republic
marek.blahus@sketchengine.co.uk

Miloš Jakubíček
Lexical Computing, Brno, Czech Republic
milos.jakubicek@sketchengine.eu

Katarína Petreková
Lexical Computing, Brno, Czech Republic
Faculty of Arts, Masaryk University, Brno, Czech Republic
katarina.petrekova@sketchengine.eu

Yingtian Wu
Lexical Computing, Brno, Czech Republic
wyt15858810001@163.com

Barbara Heinisch

DICTIONARIES, GLOSSARIES AND TERMINOLOGY DATABASES IN THE AGE OF ARTIFICIAL INTELLIGENCE

Rethinking Boundaries and Functions

Abstract The increasing integration of large language models (LLMs) into (multilingual) specialized communication is challenging traditional distinctions between dictionaries, glossaries and terminology databases. While these resources differ in structure, purpose, theoretical foundations and user groups, LLMs increasingly access their underlying data as machine-readable 'knowledge'. This paper examines how approaches ranging from prompt-based glossary integration and retrieval-augmented generation (RAG) to terminology-augmented generation (TAG), API-based access and knowledge-graph-based systems are affecting the relationship between language resources and AI. It argues that reducing rich lexicographic and terminological resources to simple equivalence lists risks losing contextual and conceptual information, whereas structured access can preserve and exploit their descriptive richness. At the same time, resource typologies remain relevant: resource type and metadata provide information about a resource's theoretical basis, purpose, intended users and epistemic status, particularly when LLMs combine multiple resources that may provide competing information. Therefore, resource type, underlying data and resource function should be considered together. This perspective highlights the continuing importance of lexicographic and terminological resources in AI environments, suggesting an increasing role for lexicographers and terminologists in creating language data and organizing knowledge for users and machines alike.

Keywords Lexicography; terminology work; knowledge graph; large language models; retrieval-augmented generation; terminology-augmented generation

1 Introduction

Dictionaries, glossaries and terminology databases have long been understood as distinct types of lexical resources, each grounded in specific theoretical traditions, structural principles and communicative purposes. Dictionaries typically aim to describe the lexical system of a language in its semantic, grammatical and pragmatic complexity (Hartmann, 2016), glossaries tend to offer selective, domain-bound explanations of terms for specific user groups (Bremmer et al., 2023); and terminology databases are concept-oriented resources (Budin, 2022) often designed to support standardization, consistency and multilingual knowledge management. These distinctions have provided a stable framework for lexicographic and terminological practice for decades. However, the rapid rise of language technologies driven by artificial intelligence (AI) is fundamentally challenging both the boundaries between these (and other) resource types and the assumptions that underlie their design and use.

In general, lexicography, terminology, ontology and knowledge graphs represent complementary approaches to the organization and representation of knowledge, differing primarily in their focus and level of abstraction. The key distinction between lexicography and terminology work is the semasiological versus the onomasiological approach (Humbley & Warburton, 2025): Lexicography is concerned with the description of lexical units and their meanings, usage, grammatical properties and communicative functions within a language. Terminology focuses on concepts within specialized domains and the terms that designate them. It emphasizes concepts as the starting point of terminology work for communication in languages for specific purposes (and often considers standardization). While lexicography is generally

semasiological, proceeding from words (lemmas) to meanings, terminology is typically onomasiological, proceeding from concepts to their terms (Humbley & Warburton, 2025; Costa, 2013). Both disciplines, however, create structured resources that capture semantic relations, definitions and contextual information.

Ontologies and knowledge graphs extend these traditions into the domain of formal knowledge representation. Ontologies as "formal and explicit specifications of shared conceptualizations of a domain" (Budin, 2005, p. 103) enable consistent interpretation by computational systems. Knowledge graphs operationalize such structures by connecting entities, their properties and their relations (Hogan et al., 2022). Lexicographic and terminological resources may serve as valuable inputs for ontologies and knowledge graphs, providing linguistic labels, definitions and semantic relations. Conversely, ontological and graph-based knowledge representations offer new possibilities for using lexicographic and terminological resources in AI-driven applications, such as large language models (LLMs).

2 Lexicographic and terminological resources for machines rather than users

In contemporary AI-supported workflows (most notably in neural machine translation, content generation and multilingual communication), lexical resources are increasingly used by machines rather than directly by users. Neural machine translation systems (for example, DeepL) often rely on glossaries that require lexical data to be reduced to simple 1:1 'equivalence' across languages (Absolon, 2025). In this process, complex dictionary entries and richly structured terminological records are flattened into minimal 'equivalence' lists that disregard different meanings, contexts, registers, collocations and pragmatic information. While such reduction enables technical efficiency, it raises concerns about the long-term impact on lexicographic and terminological quality. The risk is not merely information loss at the point of use, but a gradual reshaping of resource creation itself, driven by the perceived requirements of AI tools rather than by linguistic or communicative needs.

While technologically efficient, this reduction risks a 'poorification' of lexicography and terminology work, as semantic and/or pragmatic information is stripped away; leading to a situation where information-rich resources are flattened to fit the technical constraints of specific AI applications. From a disciplinary perspective, this trend challenges core principles of both lexicography and terminology work, including the commitment to user- and purpose-oriented description. (For good reason, different types of dictionaries have emerged, including bilingual (translation) dictionaries and specialized dictionaries as well as terminology databases). It also calls into question the traditional differences between dictionaries, glossaries and terminology databases, as all three are increasingly treated as interchangeable 'lexical data sources' within AI systems, regardless of their original purpose or internal logic.

At the same time, recent advances in generative AI challenge the assumption that lexical resources must be simplified to be machine-usable. Unlike earlier rule-based or statistical systems, generative models are capable of interacting with external 'knowledge' sources in more flexible and interpretive ways (Lewis et al., 2020). Emerging approaches based on automatic programming interfaces (APIs), such as model context protocols (MCP, 2026) allow large language models to access lexicographic and terminological resources dynamically, retrieve information selectively and adapt it to specific (communicative) tasks. In such scenarios, dictionaries and terminology databases no longer need to be pre-emptively simplified to fit rigid input formats. Instead, they can be consulted as structured data sources whose full descriptive richness can be used (Agrawal et al., 2024; Di Nunzio, 2025; Lackner et al., 2025b;

Li et al., 2024) for generating texts meeting the requirements of the relevant purpose. This development can be illustrated through different approaches to specialized translation using large language models.

2.1 Glossaries as resources for LLMs

The most straightforward scenario involves glossaries or bilingual word lists. Depending on their size, such resources can be incorporated directly into the prompt of a large language model as plain-text input, typically consisting of source-language and target-language term pairs. For larger glossaries, retrieval-augmented generation (RAG) (Lewis et al., 2020) can be employed to provide the model with access to relevant entries at inference time. Since glossaries and bilingual word lists are commonly stored in generic formats such as TXT or CSV, they can be integrated into LLM-based workflows with relatively little technical effort.

2.2. Terminology databases as resources for LLMs

Terminology databases present a different situation. Compared with glossaries, they are typically larger, concept-oriented and more richly structured, containing definitions, subject fields, usage notes, contextual information and multilingual equivalents. They are often maintained in specialized terminology management systems and stored in dedicated formats such as TBX. Consequently, the direct use of terminology databases within conventional RAG architectures may require the export or adaptation of terminological data into formats suitable for retrieval systems, potentially resulting in the loss of valuable information. To address these limitations, the concept of terminology-augmented generation (TAG, also referred to as TermRAG) (Di Nunzio, 2025; Lackner et al., 2025a) has emerged. TAG enables large language models to access terminological resources directly, typically via APIs, thereby preserving the richness of the terminological resource while ensuring the consistent use of domain-specific terminology. Terminology-augmented generation offers several advantages over retrieval-augmented generation (RAG), particularly in terminology-intensive and multilingual environments (Lackner et al., 2025a): By leveraging direct access to terminology databases, TAG achieves faster processing (by retrieving only the relevant terminological information rather than larger text passages) while enabling real-time retrieval of information from terminology databases. Unlike the probabilistic retrieval mechanisms commonly employed in RAG systems, TAG benefits from deterministic access through exact searches and controlled retrieval processes, ensuring greater consistency and reliability. Furthermore, the concept-oriented and linguistically structured nature of terminological data provides a flexible data model that aligns well with the requirements of large language models. Through real-time API integration, TAG can access terminology databases without the overhead associated with embedding and retrieving large document collections (Lackner et al., 2025a).

2.3 Dictionaries as resources for LLMs

A complementary scenario concerns dictionaries. Unlike glossaries, dictionaries typically contain richer linguistic information, including definitions, sense distinctions, grammatical properties, collocations, usage examples, register labels and pragmatic notes. Dictionaries are traditionally designed for human consultation, but LLMs can also exploit lexicographic information to support tasks such as word-sense disambiguation, context-sensitive lexical choice, register adaptation and stylistic consistency (including in translation tasks). In the field of natural language processing, frequently used lexicographic resources such as WordNet often

contain example sentences that are short, schematic and limited in contextual richness (Almeman et al., 2024). Therefore, they offer only limited support for modelling realistic language use. In contrast, dictionary sources such as the Oxford English Dictionary (Oxford University Press) typically provide more informative examples with richer contextual information and clearer usage patterns. At the same time, LLMs such as ChatGPT can generate example sentences that are, in some cases, even more effective for computational understanding of word meaning (Almeman et al., 2024). This highlights an emerging complementarity between traditional lexicographic practice and AI-generated data for lexicographic purposes. A major advantage of dictionaries is that, in contrast to WordNet, they are more diverse in form, content and purpose (Tarp, 2017), reflecting real-world language variation across different user needs and communicative situations.

2.4 *Knowledge graphs as resources for LLMs*

Knowledge graphs (KGs) are an effective approach for organizing and querying interconnected data. They provide a semantic framework that supports complex queries and reasoning and have been widely applied in areas such as search, recommendation systems and data integration. Structurally, KGs represent entities and concepts as nodes connected by relationships and enriched with properties (Ibrahim et al., 2024). In knowledge graphs data are stored as networks of interconnected entities rather than as (often) isolated entries or records. LLMs can exploit these relationships to retrieve not only individual terms, but also the surrounding concept structures, including hierarchical relations, (domain) dependencies and semantic constraints. This enables context-aware language generation in LLMs. Knowledge graphs can be integrated into LLMs at various stages, such as during training, output verification and inference (Agrawal et al., 2024). In either case, KGs can help reduce the occurrence of hallucinations text output generated by LLMs (Agrawal et al., 2024).

The integration of LLMs with knowledge graphs combines the linguistic capabilities of LLMs with the structured, verifiable data contained in knowledge graphs, resulting in more accurate, context-sensitive and reliable LLM-based text generation. Knowledge graphs enrich LLMs with up-to-date factual data, support contextual reasoning and personalization and improve transparency by making underlying knowledge structures and reasoning paths more explicit (increasing the explainability and transparency of LLMs) (Ibrahim et al., 2024).

3 Outlook

The increasing integration of large language models into text generation workflows is reshaping the traditional distinction between dictionaries, glossaries, terminology databases and knowledge graphs. While these resources are differentiated by purpose, structure and user orientation, their role in AI-supported text generation is increasingly determined by the functions they can perform rather than their typological categorization.

At a more general level, the increasing availability of structured lexical resources in computationally accessible formats suggests a shift in perspective from resource types to resource functions. Rather than asking whether a dataset constitutes a dictionary, glossary or terminology database, the more relevant question becomes what type of (lexical) knowledge it encodes, how this knowledge is structured and under which conditions it can be effectively reused by LLMs. This perspective also extends to more complex forms of knowledge representation such as knowledge graphs, which further enable LLMs to exploit relational and contextual knowledge for more context-sensitive text generation.

Although the underlying data contained in lexicographic and terminological resources may be more directly relevant to LLMs than the particular formats or interfaces through which these data are presented, the distinction between resource types remains analytically and practically significant. Dictionaries, glossaries, terminology databases and knowledge graphs did not emerge as arbitrary alternative formats for identical information, they are grounded in different theoretical traditions, data-modelling principles, purposes and user groups. These differences influence not only how the resources are created and used, but also the content, structure and epistemic status of the data made available to LLMs.

This becomes particularly important when multiple of these resources are accessed simultaneously for a specific generation task. If, for example, an LLM is provided with data from several dictionaries, terminology databases and knowledge graphs that offer (partially) conflicting terms, definitions or concept relations, the system requires information about, for example, the domain, purpose and scope of the respective resources in order to determine how their contents should be prioritized. In this context, metadata describing the domain and purpose of lexicographic and terminological data becomes particularly important. Resource type is a central element of these metadata, as it reflects the intended domain, scope and purpose of a dataset and can therefore help guide the prioritization of its contents. Metadata may thus be as important as the data themselves. Metadata can provide the (contextual) information necessary to distinguish between resources and to guide the selection and weighting of their contents according to the communicative situation and task. Thus, a prompt requesting the generation of a specialized text could specify not only which external resources should be consulted, but also the relevant domain, target audience, language variety or intended function of each resource. The distinction and explicit description of resource types therefore remain important in LLM-supported text generation because resource typology can provide essential contextual information for the requirements of the specific communicative situation and the integration of heterogeneous language resources.

This development provides an opportunity to reconsider, and in some cases reaffirm, the functional roles of dictionaries, glossaries and terminology databases in AI-mediated environments that increasingly serve both users and computational systems. The boundary between lexicography and terminology, long debated at a theoretical level (Bergenholtz & Tarp, 2010; Costa, 2013; Costa et al., 2023; Riggs, 1989), thus acquires new practical relevance in the age of AI. Questions of sense versus concept, description versus prescription and usage versus standardization are no longer purely academic but directly affect how lexical resources are integrated into AI systems.

This development has implications for data modelling, interoperability and governance, as metadata describing a resource's type, purpose and domain become increasingly important when LLMs combine multiple, potentially conflicting external resources. These metadata can help AI systems determine which resource is most relevant to a particular communicative context, for example by prioritizing terminology for expert audiences or more general lexical choices for non-specialist communication.

Specialized communication demonstrates that language and knowledge are inseparable, as the quality and clarity of specialized discourse depend on how effectively underlying knowledge is structured and expressed through linguistic means (Costa, 2013; Humbley & Warburton, 2025). In this sense, lexicographers, terminologists and knowledge graph engineers play a key role in mediating between linguistic expression and structured knowledge by creating and structuring rich 'knowledge' resources to ensure that (specialized) communication fulfils its purpose.

References

Absolon, J. (2025). Redefining Terminology Work: The Role of Glossaries and TMs in Customized AI-Driven Machine Translation. *Jazykovedný Časopis*, 76(1), 237–246.

Agrawal, G., Kumarage, T., Alghamdi, Z. & Liu, H. (2024). Can Knowledge Graphs Reduce Hallucinations in LLMs? : A Survey. In K. Duh, H. Gomez & S. Bethard (Eds.), *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)* (pp. 3947–3960). Association for Computational Linguistics.

Almeman, F. Y., Schockaert, S. & Espinosa Anke, L. (2024). WordNet under Scrutiny: Dictionary Examples in the Era of Large Language Models. In N. Calzolari, M.-Y. Kan, V. Hoste, A. Lenci, S. Sakti & N. Xue (Eds.), *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)* (pp. 17683–17695). ELRA and ICCL.

Bergenholtz, H. & Tarp, S. (2010). LSP Lexicography or Terminography? The Lexicographer's Point of View. In P. A. Fuertes-Olivera (Ed.), *Lexicographica. Series maior: Vol. 136. Specialised dictionaries for learners* (pp. 27–38). De Gruyter.

Bremmer, R. H., Seiler, A. & Pons-Sanz, S. M. (2023). Function of glossaries. In A. Seiler, C. Benati & S. M. Pons-Sanz (Eds.), *Medieval Glossaries from North-Western Europe*. Brepols Publishers.

Budin, G. (2005). Ontology-driven translation management. In H. V. Dam, H. Gerzymisch-Arbogast & J. Engberg (Eds.), *Text, translation, computational processing: Vol. 7. Knowledge systems and translation* (Vol. 7, pp. 103–124). M. de Gruyter.

Budin, G. (2022). Wüsters Allgemeine Terminologielehre – ein Grenzgebiet zwischen Sprachwissenschaft, Logik, Ontologie, Informatik und den Sachwissenschaften. In D. Candel (Ed.), *Eugen Wüster Et La Terminologie De L'école De Vienne* (pp. 243–294). Société d'histoire et d'épistémologie des sciences du langage.

Costa, R. (2013). Terminology and Specialised Lexicography: two complementary domains. *Lexicographica*, *29*(2013), 29–42. https://doi.org/10.1515/lexi-2013-0004

Costa, R., Salgado, A., Ramos, M., Almeida, B., Silva, R., Carvalho, S., Khan, F., Tasovac, T., Khemakhem, M. & Romary, L. (2023). A crossroad between lexicography and terminology work: Knowledge organization and domain labelling. *Digital Scholarship in the Humanities*, *38*, i17-i29. https://doi.org/10.1093/llc/fqad022

Di Nunzio, G. M. (2025). Terminology-Augmented Generation (TAG): Foundations, Use Cases, and Evaluation Paths. *Journal of Digital Terminology and Lexicography*, 1(1), 96–104. https://doi.org/10.25430/PUPJ.JDTL.1752566034

Hartmann, R. R. K. (2016). *Teaching and Researching Lexicography*. Routledge.

Hogan, A., Blomqvist, E., Cochez, M., D'amato, C., Melo, G. de, Gutierrez, C., Kirrane, S., Gayo, J. E. L., Navigli, R., Neumaier, S., Ngomo, A.-C. N., Polleres, A., Rashid, S. M., Rula, A., Schmelzeisen, L., Sequeda, J., Staab, S. & Zimmermann, A. (2022). Knowledge Graphs. *ACM Computing Surveys*, *54*(4), 1–37. https://doi.org/10.1145/3447772

Humbley, J. & Warburton, K. (2025). Terminology throughout history: An introduction. In K. Warburton & J. Humbley (Eds.), *Terminology throughout History* (pp. 1–24). John Benjamins Publishing Company.

Ibrahim, N., Aboulela, S., Ibrahim, A. & Kashef, R. (2024). A survey on augmenting knowledge graphs (KGs) with large language models (LLMs): models, evaluation metrics, benchmarks, and challenges. *Discover Artificial Intelligence*, 4(1). https://doi.org/10.1007/s44163-024-00175-8

Lackner, A., Vega-Wilson, A. & Lang, C. (2025a). An Evaluation of Terminology-Augmented Generation (TAG) and Various Terminology Formats for the Translation Use Case. *Journal of Digital Terminology and Lexicography*, 1(2). https://doi.org/10.25430/pupj.jdtl.1763379938

Lackner, A., Vega-Wilson, A. & Lang, C. (2025b). Terminology Augmented Generation: A Systematic Review of Terminology Formats for In-Context Learning in LLMs. In *3rd International Conference on "Multilingual digital terminology today. Design, representation formats and management systems" (MDTT)* (pp. 1–9).

Lewis, P., Perez, E., Piktus, A., Petroni, F., Karpukhin, V., Goyal, N., Küttler, H., Lewis, M., Yih, W., Rocktäschel, T., Riedel, S. & Kiela, D. (2020). Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan & H. Lin (Eds.), *Advances in Neural Information Processing Systems* (Vol. 33, pp. 9459–9474). Curran Associates, Inc.

Li, M., Miao, S. & Li, P. (2024). *Simple Is Effective: The Roles of Graphs and Large Language Models in Knowledge-Graph-Based Retrieval-Augmented Generation*. https://doi.org/10.48550/arXiv.2410.20724

MCP. (2026). *What is the Model Context Protocol (MCP)?* Retrieved May 30, 2026, from https://modelcontextprotocol.io/docs/getting-started/intro

Oxford University Press. Oxford English Dictionary (OED). Retrieved from https://www.oed.com

Riggs, F. W. (1989). Terminology and Lexicography: Their Complementarity. *International Journal of Lexicography*, 2(2), 89–110. https://doi.org/10.1093/ijl/2.2.89

Tarp, S. (2017). The concept of dictionary. In P. A. Fuertes-Olivera (Ed.), *The Routledge Handbook of Lexicography* (pp. 237–249). Routledge.

Contact information

Barbara Heinisch
Eurac Research, Italy
barbara.heinisch@eurac.edu

Acknowledgements

This paper was funded by the EC-MCSA Seal of Excellence Programme of the Autonomous Province of Bozen/Bolzano – Department for Innovation, Research and University, project *University terminology in German in the age of ChatGPT* (UniTermGPT).

Bruno Nahod & Marijana Tuta Dujmović

FROM TEXT TO TERMS

TermScout, an AI Module for Early Terminology Processing

Abstract This paper presents TermScout, a GPT-based tool developed to support the initial stage of terminology management by identifying candidate terms in specialised texts. The study evaluates TermScout using a corpus focused on cochlear implantation, assessing both English term extraction and the generation of Croatian equivalents. TermScout identified 100 English candidate terms from a specialised textbook, while Croatian equivalents were generated either directly by TermScout or through a subsequent process utilising TermAI. An independent term validation tool was used to assess all results. The tool demonstrated strong performance in English term extraction, achieving a validity rate of 92%. In contrast, generating Croatian equivalents proved more challenging, with validity rates of 22% for TermScout and 29% for TermAI. These findings indicate that concept identification and bilingual term generation are distinct tasks that require distinct linguistic and terminological expertise. The study underscores the potential of AI-assisted workflows for terminology management, while emphasising the continued need for expert validation.

Keywords Terminology management; candidate-term identification; large language models; bilingual terminology; AI-assisted terminology processing

1 Introduction[1]

The identification of specialised concepts and their associated terms constitutes a fundamental step in terminology management. Consequently, automatic term extraction plays a central role in terminology research. Traditional approaches employ statistical measures, frequency counts, collocation analysis, linguistic patterns, or comparisons between specialised and general language corpora (Ahmad et al., 1994; Frantzi et al., 2000). While these methods facilitate the identification of potential terms, they frequently yield numerous irrelevant results, necessitating expert filtering and validation (Nazarenko & Zargayouna, 2009).

Contemporary terminology theory asserts that terminology management extends beyond the identification of repeated words or phrases. Recent communicative and sociocognitive approaches emphasize the underlying concepts within specialised language (Cabré, 1999; Temmerman, 2000; Faber, 2012). Terminology work primarily concerns the structuring and organisation of concepts, rather than focusing solely on individual words. Therefore, a candidate term is considered valuable not only due to its frequency of occurrence, but also because of the significance of its associated concept.

This distinction is particularly significant in highly specialised fields, where lexical variation often reflects nuanced differences in expert knowledge. As specialised knowledge expands, identifying the most relevant concepts becomes increasingly challenging. Terminologists must not only extract potential terms from texts but also determine which terms warrant further analysis.

Within Croatia's national terminological database, Struna (Institut za hrvatski jezik, n.d.), terminology work has traditionally depended on close collaboration of domain experts and

[1] The authors used GPT-based tools for language editing, proofreading, and stylistic revision of the manuscript. All research design, data collection, analysis, interpretation of results, and conclusions remain the sole responsibility of the authors.

terminologists. Experts identify relevant concepts and potential terms from both literature and professional practice, while terminologists analyse these concepts, consult the corpus, and process terms in accordance with established guidelines (Nahod, 2016). Although this approach has yielded high-quality resources across multiple fields, it remains time-intensive and difficult to scale. As demand for terminological resources continues to increase while expert-driven workflows remain resource-intensive, there is growing interest in leveraging technology to support certain aspects of the process.

1.1. Large language models in terminology processing

Generative artificial intelligence is exerting a growing influence on lexicography and terminology, attracting increased scholarly attention. De Schryver (2023) demonstrated that large language models can perform a range of lexicographic tasks, but also identified significant challenges, including hallucinations, inconsistencies, and difficulties distinguishing authentic from generated information. Jakubíček & Rundell (2023) reported that GPT-based systems can assist with post-editing and lexicographic processes, although expert supervision remains necessary. Lew (2024) contends that artificial intelligence will transform the production of dictionaries and language resources, not by replacing lexicographers, but by reallocating certain tasks from humans to computational systems.

Research focused on term extraction has also yielded promising outcomes. Chataut et al. (2024) compared various large language models for keyword extraction in specialised domains and found that GPT-based systems could accurately identify relevant domain-specific expressions. However, the study also highlighted persistent challenges, including candidate validation, addressing domain-specific complexities, and distinguishing between superficially plausible terms and genuinely established ones.

These findings have significant implications for terminology management. In contrast to many lexicographic tasks, terminology work involves specialised concepts that are typically defined with greater precision by domain experts. Consequently, some challenges observed in general-language lexicography may be less pronounced in specialised contexts. Nevertheless, identifying appropriate concepts and reaching consensus on correct terms generally requires expert knowledge and formal validation.

1.2. From TermAI to TermScout

The application of large language models to Croatian terminology management began with TermAI, a custom GPT (OpenAI, 2025a, 2025b) designed to generate early-stage (raw) terminological units. Previous research demonstrated that TermAI could produce draft definitions, conceptual links, and translation equivalents of sufficient quality to serve as a foundation for further processing (Nahod, 2024; Nahod & Vukša Nahod, 2024). These findings indicate that certain aspects of terminology management may be partially automated without significant loss of quality.

Nonetheless, these studies identified a persistent bottleneck in the process. Although AI systems can assist in generating terminological information, the identification of candidate terms continues to depend primarily on domain experts, who must manually review specialised literature to select concepts for further analysis.

To address this challenge, TermScout was developed as a GPT-based tool to support the initial stage of terminology processing: candidate term extraction. In contrast to traditional term extraction systems that prioritise recall, TermScout emphasises precision. Its objective is to provide a concise list of conceptually relevant candidate terms for subsequent processing, rather than exhaustively listing all possible terms within a text.

1.3. Aim of the study and research questions

The present study evaluates the performance of TermScout in the domain of cochlear implantation. This domain was selected because of its interdisciplinary nature, rapid technological development, extensive reliance on English-language scientific literature, and considerable terminological variation. These characteristics make cochlear implantation a particularly demanding environment for terminology management and therefore an appropriate test domain for evaluating AI-assisted term extraction.

Unlike previous evaluations of TermAI, which focused primarily on the generation of terminological information, the present study investigates the effectiveness of candidate-term identification and bilingual term generation. More specifically, the study examines not only the quality of English source-term extraction but also the generation of Croatian equivalents and the potential contribution of a secondary GPT-based workflow involving TermAI.

The study addresses the following research questions:

1. Can TermScout successfully identify domain-specific English terms from a specialised source text?
2. Can TermScout generate Croatian equivalents of sufficient quality to support bilingual terminology processing?
3. Are Croatian equivalents generated by TermScout comparable to the ones generated by TermAI?

By addressing these questions, the study contributes to the growing discussion on the role of artificial intelligence in terminology management and explores the extent to which large language models can support the transition from specialised texts to structured terminological resources.

2 TermScout and AI-Assisted Terminology Processing

TermScout was developed as a specialised tool to support the initial stage of terminology processing, specifically the identification of candidate terms in specialised texts. The system was conceived in response to challenges encountered during the development of AI-assisted terminology workflows for the Croatian national terminological database, Struna.

TermScout was developed to assist in identifying terms prior to more detailed processing. In the current workflow, TermScout serves as the first step. The candidate terms identified by TermScout are subsequently processed by TermAI, generating raw terminological units. Final decisions and validation of results remain the responsibility of human experts.

The integration of TermScout and TermAI reflects a broader transformation in terminology management practices. Instead of requiring experts to oversee every stage, artificial intelligence now assists with repetitive and time-consuming tasks. This enables terminologists

and domain specialists to concentrate on evaluating concepts, validating terms, and establishing standards.

2.1. Candidate-Term Identification and Scoring

TermScout operates exclusively on materials supplied by users, including PDF documents, text files, and web resources. This method ensures transparency in the extraction process and prevents the system from accessing uncontrolled external sources. Subsequently, the system detects the language and prepares the text for analysis. At present, TermScout supports English and Croatian documents.

After preprocessing, the system identifies candidate noun phrases using syntactic, semantic, and structural indicators. TermScout defines a term as a noun or noun phrase that denotes a distinct concept within a specific domain. Consequently, the extraction process prioritises conceptual relevance over mere frequency of occurrence.

The system employs both positive and negative indicators to identify candidate terms. Positive indicators include proximity to definitional phrases such as is a, defined as, refers to, or known as; inclusion in structures like type of, kind of, or category of; frequent occurrence in specialised contexts; presence in headings, glossaries, or lists; and formation as multiword noun phrases. The system particularly targets noun phrases containing two to five words, as specialised concepts are frequently expressed in this manner.

To minimise noise, the system applies several exclusion criteria. It removes generic vocabulary, procedural language, boilerplate text, captions, legal disclaimers, and any phrases that do not constitute noun phrases. Additionally, expressions that occur frequently but lack conceptual specificity are excluded. The objective is to identify a concise set of highly relevant candidate terms for further analysis. The scoring process integrates linguistic, contextual, statistical, and semantic evidence:

$$S = (0.25D) + (0.15T) + (0.20F) + (0.15S) + (0.10M) + (0.10C) - (0.05X)$$

where D represents definitional context, T taxonomic indicators, F normalised frequency, S semantic similarity, M multiword structure, C consistency of usage, and X exclusion penalties. The weights were empirically assigned during system development and are intended as heuristic indicators rather than statistically optimised parameters.

All feature values are normalised to the range [0, 1]. Candidates with scores below 0.55 are excluded from further consideration. The remaining candidates are ranked by confidence, and the highest-ranking terms are selected for output. To maintain manageability, the system presents up to 100 candidate terms per extraction cycle. This methodology aligns with TermScout's emphasis on precision, prioritising a concise and relevant set of terms over exhaustive extraction.

2.2. Bilingual Processing and Integration with TermAI

Following the identification of candidate terms, TermScout generates a structured bilingual entry for each term. The output includes the source-language term, a preliminary genus–differentia definition, available synonyms, a proposed target-language equivalent, frequency data, and complete source references in APA format.

The bilingual feature operates differently depending on the source language. For English documents, TermScout identifies English terms and proposes Croatian equivalents, while for Croatian documents, the process is reversed. When a verified bilingual equivalent is available in existing resources or databases, it is presented directly. Otherwise, the system generates a provisional equivalent based on the language model's contextual and semantic analysis, explicitly marking it as inferred rather than verified. The results are organised in a bilingual table to facilitate subsequent processing and expert review.

The candidate list is subsequently forwarded to TermAI, which enriches the entries with additional details, including refined definitions, conceptual relationships, and alternative equivalents. TermScout and TermAI collaborate within an AI-assisted workflow: TermScout identifies concepts, while TermAI contributes to developing raw terminological units.

This distinction is significant in terminology work. Identifying key concepts within specialised texts and assigning accepted equivalents in another language constitute distinct tasks. The former involves recognising specialised concepts in the source text, whereas the latter requires familiarity with expert conventions in the target language. Accordingly, this study examines both the effectiveness of source term extraction and the quality of bilingual equivalent generation.

TermScout's design aims to support, not replace, expert-led terminology management. It is meant to speed up the early steps of processing specialised texts, while keeping terminologists and domain experts central in checking and organising specialised knowledge.

3 Materials and Methods

Cochlear implantation is a highly interdisciplinary field that is constantly developing. Consequently, professionals in this domain employ a broad and continually evolving set of specialised terms. The necessity for standardised terminology is particularly evident in Croatia. Although cochlear implantation has been performed in Croatia since 1996 and is now a well-established clinical practice, the terminology employed continues to vary across publications, educational materials, public resources, and clinical settings.

These factors position cochlear implantation as an appropriate domain for evaluating AI-assisted terminology extraction. The field's complexity, rapid evolution, reliance on English-language sources, and pronounced terminological variation present a challenging environment for assessing whether automatically extracted terms can facilitate more effective terminology management.

3.1. Corpus and Source Material

The evaluation corpus consisted of the textbook Cochlear Implants: Audiologic Management and Considerations for Implantable Hearing Devices (Wolfe, 2020). Three primary factors motivated the selection of this textbook. First, it serves as a comprehensive and authoritative source of current knowledge in the field. Second, it encompasses terminology from multiple subdomains associated with cochlear implantation. Third, specialised textbooks typically provide clear definitions and explanations, which facilitate the evaluation of terminology-extraction systems that rely on definitional and taxonomic cues (Bowker & Pearson, 2002). Furthermore, specialised texts constitute an important source of terminological evidence because they reflect the conceptual organisation and linguistic practices of expert communities (Pearson, 1998).

The entire textbook was utilised as the source for extracting candidate terms. All extraction procedures employed consistent TermScout settings and scoring parameters.

3.2. Study design

The study evaluated the effectiveness of an AI-assisted terminology workflow for extracting English terms and generating Croatian equivalents. Initially, TermScout processed the source textbook to identify domain-specific English terms. The system generated a ranked list of 100 candidate terms, each paired with a Croatian equivalent during extraction. To determine whether a secondary AI-assisted workflow could enhance bilingual terminology generation, TermScout's English term list was provided to TermAI. TermAI generated Croatian equivalents for the same English concepts, resulting in a third dataset.

This experimental design enabled evaluation of three outputs: source-language term extraction, direct bilingual generation via TermScout, and bilingual generation via a secondary TermAI-assisted workflow.

3.3. Validation procedure

An independent GPT-assisted term-validation tool, developed specifically for this study, was utilised to evaluate all three datasets. The aim was to determine the degree to which the extracted or generated candidates aligned with established terms in professional contexts.

Each candidate was classified into one of four validation categories:

TABLE 1 Validation categories

Category	Interpretation
Exact match	Candidate corresponds directly to an attested term
Close match	Candidate represents a near-equivalent, accepted variant, or partially acceptable designation
Suspect	Candidate is weakly supported or terminologically questionable
False positive	Candidate cannot be validated as a domain term

Two types of validity were assessed. The first measure, strict validity, included only candidates that were exact matches. The second measure, moderate validity, included both exact and close matches. Considering bilingual terminology often displays different forms, morphological variations, and partial overlaps, moderate validity was especially important for evaluating Croatian equivalents.

In addition to validation categories, the tool also provided information on the frequency of each candidate's occurrence in reference sources. These frequency bands were subsequently used to analyse the relationship between candidate frequency and validation success.

3.4. Statistical Analysis

The analysis was conducted at two distinct levels. First, descriptive statistics were employed to compare the distribution of validation categories across the three datasets. The analysis specifically examined strict validity rates, moderate validity rates, and the proportion of false positives.

Second, inferential statistics were applied to examine the relationship between term frequency and validation success. Given the ordinal nature of validation outcomes, Spearman's rank-order correlation was used to assess the association between appearance-frequency groups and validation quality. Chi-square tests were used to determine whether validation results differed significantly across appearance-frequency groups. Effect sizes were calculated using Cramér's V.

Finally, the two Croatian datasets were directly compared. Validation categories were converted to a four-point ordinal scale, assigning the highest score to exact matches and the lowest to false positives. As both Croatian lists corresponded to the same set of English concepts, the Wilcoxon signed-rank test was employed to evaluate differences between the two workflows.

By integrating descriptive and inferential analyses, the study assessed the overall effectiveness of TermScout, compared the bilingual workflows, and evaluated whether appearance frequency serves as an indicator of terminological validity.

4 Results

4.1 Validation of Extracted English Terms

The first dataset consisted of 100 English candidate terms extracted by TermScout from Wolfe's (2020) textbook. The candidates were evaluated using an independent semi-automatic term-validation tool and classified as exact matches, close matches, suspect candidates, or false positives.

The results indicate a high level of extraction accuracy. Of the 100 extracted candidates, 85 were classified as exact matches and seven as close matches. Six candidates were classified as suspect, while only two were identified as false positives.

TABLE 2 Validation results for the TermScout English list

Validation category	Count	Percentage
Exact match	85	85.0%
Close match	7	7.0%
Suspect	6	6.0%
False positive	2	2.0%

Using the strict validity measure, which includes only exact matches, the English list achieved a validity rate of 85%. Using the moderate validity measure, which includes both exact and close matches, validity increased to 92%.

4.2 Validation of Croatian Equivalents

The second dataset consisted of Croatian equivalents generated directly by TermScout for the extracted English concepts. Validation results differed substantially from those observed for the English list.

Only five candidates were classified as exact matches and seventeen as close matches. Twenty-five candidates were categorised as suspect, while fifty-three were classified as false positives.

TABLE 3 Validation results for the TermScout and TermAI Croatian list

Term Scout Validation category	Count	Percentage	Term AI Validation category	Count	Percentage
Exact match	5	5.0%	Exact match	10	10.0%
Close match	17	17.0%	Close match	19	19.0%
Suspect	25	25.0%	Suspect	17	17.0%
False positive	53	53.0%	False positive	54	54.0%

The strict validity rate was therefore 5%, while the moderate validity rate reached 22%.

The third dataset consisted of Croatian equivalents generated by TermAI from the English candidate list extracted by TermScout. Validation results showed a modest improvement compared with the Croatian list generated directly by TermScout. Ten candidates were classified as exact matches and nineteen as close matches. Seventeen candidates were categorised as suspect, while fifty-four were classified as false positives. The strict validity rate increased to 10%, while the moderate validity rate increased to 29%.

4.3 Appearance Frequency and Validation Success

The validation tool additionally assigned appearance-frequency information to each candidate. Appearance bands were used to examine whether frequency evidence was associated with validation success.

For the English dataset, candidates occurring more frequently achieved substantially higher validation rates than candidates with little or no appearance evidence. Moderate validity increased from 14.3% among candidates with minimal appearance evidence to 100% among candidates occurring in the higher-frequency bands.

A Spearman rank-order correlation demonstrated a significant positive association between appearance frequency and validation quality ($\rho = 0.435$, $p < .001$). A chi-square test further confirmed a significant relationship between appearance-frequency category and moderate validity, $\chi^2(3) = 63.12$, $p < .001$, with a large effect size (Cramér's $V = 0.79$).

A similar pattern was observed for the Croatian equivalents generated by TermScout. Candidates with higher appearance frequencies achieved substantially better validation outcomes than candidates with little or no appearance evidence. In this dataset, all candidates lacking appearance evidence were classified as false positives.

The association between appearance frequency and validation quality was statistically significant ($\rho = 0.954$, $p < .001$). A chi-square test likewise demonstrated a significant relationship between frequency category and validation outcome, $\chi^2(2) = 41.11$, $p < .001$, with a large effect size (Cramér's $V = 0.64$).

For the Croatian equivalents generated by TermAI, appearance frequency again showed a positive relationship with validation quality. Spearman's correlation remained significant ($\rho = 0.342$, $p < .001$), while the chi-square test confirmed a significant association between frequency category and validation outcome, $\chi^2(1) = 9.51$, $p = .002$. The corresponding effect size was smaller (Cramér's $V = 0.31$) than in the previous datasets.

5 Discussion

5.1 Concept Identification versus Concept Transfer

The results indicate a marked distinction between extracting terms in the source language and generating bilingual equivalents. The English candidate list demonstrated a validity rate of 92%, whereas the Croatian lists were significantly lower at 22% and 29%. These findings suggest that identifying specialised concepts in one language and assigning accurate equivalents in another constitute fundamentally different tasks.

This distinction aligns with established principles in terminology management. Extracting terms involves identifying key concepts in specialised texts by recognising recurring ideas and linguistic patterns within a specific field. In contrast, generating bilingual equivalents requires both a comprehensive understanding of the concept and familiarity with its conventional designation in another language and professional context.

These findings support the idea that terminology processing is more than just a word-based task. Communicative and sociocognitive approaches say that specialised communication is built around concepts, not just words (Cabré, 1999; Temmerman, 2000; Faber, 2012). TermScout performed well in English extraction, showing that large language models can find important concepts in specialised texts. But the much lower results for Croatian equivalents show that finding a concept does not always mean it can be easily transferred to another language.

This distinction is particularly significant in fields characterised by rapid technological advancement and extensive international communication. In such domains, new concepts frequently emerge before standardised terminology is established. Consequently, multiple competing terms may exist for the same concept, rendering the generation of equivalents considerably more challenging than the extraction of terms from the source language.

5.2 English Extraction and Croatian Equivalent Generation

The English dataset demonstrated strong performance, indicating that TermScout is effective in identifying candidate terms. The system identified numerous validated domain-specific terms and produced few false positives. These findings indicate that combining definitional indicators, taxonomic cues, contextual information, semantic similarity, and frequency-based evidence is an effective strategy for identifying specialised terms within a text.

The results for Croatian equivalents differed significantly. Although many generated terms appeared linguistically correct, a substantial number failed validation. This outcome highlights

a key limitation of AI-assisted terminology generation: language models may produce expressions that appear appropriate but do not correspond to terms actually used by professionals.

Several factors may contribute to this discrepancy. First, English terms were extracted directly from the source, whereas Croatian equivalents were generated through translation. Second, Croatian terminology for cochlear implantation is less standardised than its English counterpart. Third, Croatian exhibits greater morphological complexity and offers more methods for term formation, which can result in generated expressions that differ from those commonly used, even if their meanings are equivalent.

5.3 Appearance Frequency as a Confidence Indicator

A principal finding of the study is the association between term frequency and the likelihood of validation. Across all three datasets, candidates with higher frequencies were significantly more likely to be confirmed as exact or close matches. This relationship was statistically significant in every analysis and particularly pronounced in the Croatian datasets.

This finding holds practical implications for AI-based terminology management. Term frequency can serve as an additional indicator of confidence when evaluating automatically extracted or generated terms. Candidates with low or zero frequency are more likely to require manual review, whereas frequently occurring candidates are more likely to correspond to established terms.

This observation is particularly relevant given that LLMs frequently generate information that appears plausible but lacks evidential support. Although term frequency analysis cannot entirely resolve this issue, it can assist experts in prioritising terms for review. In practical terminology management, appearance frequency may function as a quality-control measure between automatic extraction and expert validation. Terminology-support systems could benefit from closer integration of corpus evidence, terminological databases, and language-model outputs. Rather than relying solely on generative capabilities, AI-assisted workflows may achieve greater reliability when generation is supplemented by independent evidence of actual usage.

5.4 Implications for AI-Assisted Terminology Management

This study contributes to recent research on the application of large language models in lexicography and terminology management (de Schryver, 2023; Jakubíček & Rundell, 2023; Lew, 2024). The findings indicate that artificial intelligence facilitates the initial stages of terminology work by streamlining the identification of candidate terms.

However, the results underscore the continued necessity of expert validation in terminology work. LLM based modules like TermScout and TermAI are not yet capable of independently managing terminology. Their primary advantage lies in accelerating the preliminary stages that typically require significant expert time.

Within the Struna framework, artificial intelligence should be regarded as a tool that supports rather than replaces experts. AI can assist in identifying candidate terms, drafting definitions, and proposing possible equivalents. Nevertheless, expert involvement remains essential for concept verification, term standardisation, and final decision-making.

These findings are consistent with previous studies of TermAI, which also concluded that AI provides useful preliminary information about terms but still requires expert review and correction (Nahod, 2024; Nahod & Vukša Nahod, 2024). The present study extends this understanding by demonstrating that AI can assist in identifying candidate terms even before descriptions are formulated.

5.5 The Evolving Role of Domain Experts

The results do not indicate that domain experts should be replaced in terminology work. Rather, they suggest a shift in the nature of expert involvement. Previously, specialists were responsible for identifying candidate terms, selecting concepts, reviewing literature, and participating in analysis. With the integration of AI, some of these initial tasks can be automated, allowing experts to concentrate on more advanced aspects of the process.

As a result, experts can allocate more time to analysing concepts, verifying terms, and establishing standards. In summary, their role will shift from content creation to a greater emphasis on review and decision-making.

This change is important from a sociocognitive point of view. Expert knowledge is still crucial because terminology is based on how specialised communities think and categorize (Nahod, 2016). While large language models can spot patterns in expert language, they do not take part in the real-life work of shaping and agreeing on concepts. Therefore, using AI in terminology management should be seen as a way to help experts share their knowledge more widely, not as a replacement for their expertise.

6 Conclusion

This study examined TermScout, a custom GPT-based tool designed to facilitate the initial stage of terminology processing by identifying candidate terms in specialised texts. Wolfe's (2020) textbook, Cochlear Implants: Audiologic Management and Considerations for Implantable Hearing Devices, served as the primary data source. The research evaluated TermScout's effectiveness in extracting English terms, generating Croatian equivalents, and assessed the contribution of a secondary AI-assisted workflow using TermAI.

The results indicated that TermScout was highly effective in identifying English domain-specific terms. The list of English candidate terms achieved a validity rate of 92%, demonstrating that most extracted terms corresponded to established terminology. These findings suggest that large language models can facilitate the identification of key terms in specialised texts and streamline the early stages of terminology processing.

In contrast, generating Croatian equivalents proved considerably more challenging. Both Croatian datasets exhibited substantially lower validation rates compared to the English dataset. Although the use of TermAI provided some improvement, the difference was not statistically significant. These results indicate that identifying specialised concepts and creating accepted terms in another language are distinct tasks requiring different types of expertise. While large language models can identify important terms in the source text, the production of reliable bilingual terminology continues to rely on professional usage, corpus-based evidence, and expert review.

Another key finding concerns term frequency within the datasets. Across all datasets, terms with higher frequency were significantly more likely to be validated. This suggests that term

frequency may serve as a useful confidence indicator in AI-assisted terminology work and could assist experts in prioritising validation efforts for terms requiring greater scrutiny.

Overall, the findings support the adoption of a hybrid approach to terminology management. TermScout is intended to assist, rather than replace, terminologists and domain experts by identifying candidate terms and reducing the workload associated with compiling terminology. In this workflow, artificial intelligence facilitates the extraction and processing of terms, while experts are responsible for analysis, validation, and standardisation. This study contributes to the ongoing discourse regarding the use of large language models in terminology management and suggests that their primary value lies in enhancing the efficiency of expert knowledge rather than replacing it.

Future research should evaluate TermScout's performance in additional specialised domains, investigate the integration of corpus-based validation methods, and assess whether domain-specific training can enhance the generation of bilingual terms. Such studies would advance understanding of how large language models can support the development of reliable and scalable terminological resources.

References

Ahmad, K., Davies, A., Fulford, H. & Rogers, M. (1994). What is a term? The semi-automatic extraction of terms from text. In M. Snell-Hornby, F. Pöchhacker & K. Kaindl (Eds.), *Translation studies: An interdiscipline* (pp. 267–278). John Benjamins.

Bowker, L. & Pearson, J. (2002). *Working with specialized language: A practical guide to using corpora*. Routledge.

Cabré, M. T. (1999). *Terminology: Theory, methods and applications*. John Benjamins.

Chataut, R., Sharma, S., Sharma, K. & Joshi, A. (2024). Large language models for domain-specific keyword extraction: A comparative evaluation. *arXiv*. https://doi.org/10.48550/arXiv.2404.02330

de Schryver, G. M. (2023). Generative AI and lexicography: The current state of the art using ChatGPT. *International Journal of Lexicography*, 36(4), 355–387. https://doi.org/10.1093/ijl/ecad021

Faber, P. (2012). *A cognitive linguistics view of terminology and specialized language*. De Gruyter Mouton.

Frantzi, K., Ananiadou, S. & Mima, H. (2000). Automatic recognition of multi-word terms: The C-value/NC-value method. *International Journal on Digital Libraries*, 3(2), 115–130.

Institut za hrvatski jezik. (n.d.). *Struna: Hrvatsko strukovno nazivlje*. https://struna.ihjj.hr/

Jakubíček, M. & Rundell, M. (2023). The end of lexicography? Can ChatGPT outperform current tools for post-editing lexicography? In M. Medveď, M. Měchura, C. Tiberius, I. Kosem, J. Kallas, M. Jakubíček & S. Krek (Eds.), *Proceedings of the eLex 2023 Conference*: *Electronic lexicography in the 21st century* (pp. 518–533). Lexical Computing CZ, s.r.o

Lew, R. (2024). Dictionaries and lexicography in the AI era. *Humanities and Social Sciences Communications*, 11(1), 1–12. https://doi.org/10.1057/s41599-024-02889-7

Nahod, B. (2016). *O umu stručnjaka*. Institut za hrvatski jezik i jezikoslovlje.

Nahod, B. (2024). Can we substitute field experts with customized large language model in processing specialized languages? A case study. In K. Š. Despot, A. Ostroški Anić & I. Brač (Eds.), *Lexicography and Semantics: Proceedings of the XXI EURALEX International Congress* (pp. 725–736). Institute for the Croatian Language.

Nahod, B. & Vukša Nahod, P. (2024). Artificial intelligence in terminology processing: Testing the robustness of a custom GPT. In *Proceedings of the 11th SWS International Scientific Conference on Arts and Humanities* (ISCAH 2024) (Vol. 11(1)). SGEM World Science. https://doi.org/10.35603/sws.iscah.2024/fs11.21

Nazarenko, A. & Zargayouna, H. (2009). Term extraction: A state of the art. In *Terminology and Knowledge Engineering 2009* (pp. 203–216).

OpenAI. (2025a). *GPTs*. OpenAI. https://openai.com/chatgpt/gpts/

OpenAI. (2025b). *GPT-4.5 system card*. OpenAI. https://openai.com/

Pearson, J. (1998). *Terms in context*. John Benjamins Publishing Company.

Temmerman, R. (2000). *Towards new ways of terminology description: The sociocognitive approach*. John Benjamins.

Wolfe, J. (2020). *Cochlear implants: Audiologic management and considerations for implantable hearing devices* (3rd ed.). Plural Publishing.

Contact information

Bruno Nahod
Institute for Croatian Language
bnahod@ihjj.hr

Mirjana Tuta Dujmović
SUVAG - Polyclinic for Hearing and Speech Rehabilitation
mtuta@suvag.hr

Acknowledgements

This paper was created within the framework of the project TermAI: Development and Modernization of Croatian Terminology, which is funded by the European Union – NextGenerationEU. The views and opinions expressed are solely those of the authors and do not necessarily reflect the official positions of the European Union or the European Commission. Neither the European Union nor the European Commission can be held responsible for them.

Deepak Solanki

USER-ORIENTED SPECIALIZED LEXICOGRAPHY

Corpus-Driven Models for Targeted Dictionary Design

Abstract Lexicographic practice has often balanced comprehensive coverage with systematic ordering, privileging alphabetical arrangement and encyclopedic completeness over clearly defined user needs. This paper argues that user-oriented specialized dictionaries, supported by corpus-driven and AI-augmented methods, can better serve users who require targeted lexical knowledge rather than exhaustive inventories. The study uses Śabdakalpadruma as a Sanskrit lexicographic case study and reports its conversion into a structured digital corpus of 42,824 headword entries through OCR, post-OCR cleaning, Excel inspection, SQLite structuring, JSON representation, and searchable HTML implementation. Building on this corpus, the paper develops a pilot rule-based recommender model for ritual vocabulary. Twenty ritual query terms were searched in headword and definition fields, and the extracted entries were automatically annotated through semantic tagging, co-occurrence detection, relation scoring, confidence labelling, and relevance classification. A stratified validation sample of 100 entries, with five entries from each query term, was manually evaluated. The results show that 67% of annotations were correct, 19% usable with correction, and 14% wrong. The study demonstrates a staged pathway from digitized lexicographic corpus to rule-based recommendation, human-validated training data, and future AI-assisted semantic prediction.

Keywords Specialized lexicography; corpus-driven lexicographic models; targeted dictionary design; recommender systems; rule-based lexical recommendation; Śabdakalpadruma; sanskrit lexicography; digital humanities

1 Introduction

Lexicography has historically balanced two major goals: comprehensive lexical coverage and systematic organization. In many general-purpose dictionaries, alphabetical arrangement became the dominant organizing principle because it provides quick access to individual words. However, in specialized lexicography, the user often does not require only the meaning of a single word. A student, translator, researcher, ritual specialist, literary scholar, or computational linguist may require a network of related terms, conceptual associations, domain-specific vocabulary, and contextual evidence. For such users, a dictionary is not merely a lookup tool; it is a guide to a lexical field.

Modern lexicographic theory has repeatedly emphasized the importance of user needs in dictionary design. Atkins and Rundell argue that dictionary planning must begin with the expected user and the situations in which the dictionary will be used (Atkins & Rundell, 2008, pp. 17–44). Similarly, function theory in lexicography places the user's information need at the centre of dictionary design and treats lexicographic assistance as a response to specific communicative or cognitive situations (Tarp, 2008, pp. 39–72). This paper adopts that user-oriented perspective and applies it to Sanskrit lexicographic tradition in the digital age.

The central case study of this paper is Śabdakalpadruma, a major Sanskrit lexicon traditionally valued for its rich explanations, grammatical information, etymological details, semantic distinctions, and rigorous textual citations. Unlike concise bilingual dictionaries or short glossarial resources, Śabdakalpadruma contains extended explanatory material that is particularly useful for semantic extraction and domain-based vocabulary modelling. Its internal richness

makes it suitable not only for traditional dictionary consultation but also for computational restructuring into a corpus-driven lexical resource.

The present study is based on the digital preparation of Śabdakalpadruma as part of a PhD research at JNU, New Delhi. The source material consists of five kā a-s and more than 3,000 large-format printed pages. These pages are physically and typographically denser than ordinary A4 pages, so the textual scale of the corpus is larger than the page count alone suggests. Through text extraction via OCR and subsequent cleaning, the lexicon has been transformed into a structured digital corpus containing 42,824 headword entries. The data has been organized in Excel spreadsheet, in SQLite database file, in JSON, and in searchable HTML file for user-oriented consultation.

This paper therefore moves from the question "How should a specialized dictionary be designed?" to the practical question "How can an existing Sanskrit lexicon be transformed into a user-oriented semantic recommendation system?" The answer developed here is a pilot corpus-driven recommender model that extracts a ritual-vocabulary dataset from the full database and uses it to generate curated word sets for specialized users. The paper is not presented as a fully autonomous AI system; rather, it demonstrates a staged pathway from a digitized Sanskrit lexicographic corpus to a rule-based recommender model, from recommender output to human-validated labelled data, and then to future machine-learning models for automated semantic prediction.

2 Background: User-Oriented Specialized Lexicography and Corpus Methods

Specialized lexicography differs from general lexicography because its goal is not merely to record the vocabulary of a language as a whole. It is designed for particular domains, tasks, and users. A specialized dictionary may serve students for ritual terminology, translators seeking equivalents through description, researchers of medico-botanical vocabulary, or computational linguists building semantic datasets. In each case, the usefulness of the dictionary depends on how effectively it connects the user to the relevant lexical field. This user-specific orientation is central to the theory and practice of specialized dictionaries (Bergenholtz & Tarp, 1995).

Corpus linguistics provides an empirical foundation for this task. Corpus-based methods allow the researcher to identify repeated lexical patterns, frequency distributions, co-occurrence relations, collocations, and contextual usage. Halliday and his co-authors emphasize that lexicology and corpus linguistics are closely linked because lexical meaning becomes clearer when words are studied through patterns of use rather than as isolated units (Halliday et al., 2004, pp. 1–22). Sinclair's work on concordance and collocation also shows the importance of repeated textual patterning for lexical description (Sinclair, 1991). McEnery and Brezina similarly stress the importance of empirical evidence, replication, and structured observation in corpus-based linguistic research (McEnery & Brezina 2022, pp. 145–174).

In practical lexicography, corpus evidence is especially valuable because it prevents the dictionary from depending only on intuition. Atkins and Rundell describe lexicographic evidence as a core stage in dictionary planning and entry construction (Atkins & Rundell 2008, pp. 45–76). Digital corpus tools have also shown how large lexical datasets can support dictionary-making, example selection, and word profiling in a more systematic way (Kilgarriff et al., 2014). For the present study, the Śabdakalpadruma corpus functions not as a modern usage corpus in the usual sense, but as a structured historical lexicographic corpus. Its entries, definitions, citations, and internal word associations become the evidence base for building semantic word sets.

The study also draws on the idea that a digital dictionary can become more than a passive retrieval system. In traditional print format, a user searches for a word and reads the corresponding entry. In a digital corpus-driven model, however, the system can retrieve the entry, identify associated terms, classify the domain, and recommend related vocabulary. This shift from static lookup to dynamic suggestion is central to the present paper.

3 Śabdakalpadruma as a Digital Lexicographic Corpus

The Indian lexicographic tradition has a long history of organizing vocabulary according to semantic class. Sanskrit lexicons such as Amarakośa arranged words thematically and served the needs of poets, scholars, students, and ritual-literary users. Vogel's survey of Indian lexicography shows that Sanskrit lexical traditions developed through word lists, synonymic collections, thematic organization, grammatical classification, and later expanded lexicographic compilations (Vogel, 1979, pp. 303–401). This historical background is important because the present paper does not impose an entirely foreign model on Sanskrit lexicography. Instead, it reactivates an older 'tried and tested through centuries' logic of lexical grouping through modern corpus and database methods.

Śabdakalpadruma is especially suitable for this experiment because its entries are arranged alphabetically. Which represents a major deptarture from many traditional thematic Sanskrit lexicographical models. At the same time, Śabdakalpadruma almost never gives only a single word equivalent for its entries. Its entries may contain grammatical information, etymological explanation, texual quotations, and related domain-specific references. Such features make it possible to treat each entry as a multi-layered lexical unit.

In the present study, the printed Śabdakalpadruma was converted into a digital corpus as part of my doctoral thesis. The full database currently contains 42,824 headword entries. The corpus exists in Excel file is used for inspection, manual correction, and validation. A SQLite.db file is used for query-based retrieval and computational analysis. The database has been added to a JSON structure, which supports machine-readable representation and future web or API-based applications. Also, this databse has been used to create a searchable HTML dictionary of Śabdakalpadruma, which allows users to consult the corpus directly.

4 Methodology

The methodology of this study is designed to answer three practical questions. First, how a Sanskrit lexicon converted into a structured digital corpus can be used for domain-specific lexical material extraction and user recommendation system? Second, how can the quality of the extracted and recommended material be evaluated? Third, The approach combines corpus preparation, database structuring, rule-based semantic annotation, ritual-domain extraction, co-occurrence-based recommendation, stratified validation sampling, and human review. It is deliberately transparent and reproducible so that the recommendations can be inspected rather than accepted as black-box AI output.

4.1 Research Design

This pilot dataset is then used to test semantic tagging, automatic related-word extraction, recommendation scoring, confidence labelling, and validation procedures. The design follows a controlled sequence: corpus creation, data extraction, automatic annotation, stratified validation sampling, manual checking, result analysis, and rule revision.

The present system is rule-based rather than fully autonomous. This is a methodological choice. A rule-based model allows the researcher to see why a recommendation has been produced. This is especially important for Sanskrit lexicographic material because entries may be long, citation-rich, and semantically layered. A ritual word may appear in a definition without making the entire entry strongly ritual-related. Therefore, the system treats automatic extraction as a first-stage recommendation, not as final lexicographic truth.

4.2 Implementation and Rule-Based Matching

In order to make the implementation reproducible, the pilot was implemented through a small Python workflow using sqlite3 for database access, pandas for tabular processing, openpyxl for Excel export, and regular-expression/string matching for rule-based extraction. No neural language model or external NLP library was used in the current implementation. Query terms were matched through SQL LIKE searches in the headword and definition_text fields. The automatic annotation layer used controlled keyword lists, synonym/near-synonym groups, stopword filtering, frequency counts within definition fields, and conditional scoring rules. This clarification is important because the present contribution is a rule-based lexical recommendation and validation pipeline, while AI-assisted semantic prediction is treated as a future stage built on the human-validated data produced by this pilot.

4.3 Database Structure and Lexicographic Units

The full database contains 42,824 headword entries. At the present stage, the primary operational fields include an entry identifier, the headword, and the definition or full entry text. For the recommender model, additional annotation fields were added to the pilot dataset: semantic_domain, important_words, related_words_manual, related_words_auto, relation_basis, recommendation_score, confidence, manual_validation, scholarly_notes, human_validation, and human_notes. These fields create a bridge between the historical lexicon and the computational recommender model. Table 1 summarizes these fields and their functions.

The database is treated as a multi-layered lexical corpus. At the entry level, each row represents a headword and its associated explanation. At the sense level, longer entries may contain more than one semantic value, which can be separated in future refinements. At the citation level, quoted textual evidence and source references offer contextual proximity signals. At the domain level, entries are tagged according to semantic fields such as ritual, deity, botanical, grammatical, philosophical, medical, material, social, geographical, and literary vocabulary. At the relation level, words are connected through synonymy, co-occurrence, shared domain, citation proximity, or functional relation.

TABLE 1 Database and annotation fields used in the pilot model

Field	Function
entry_id	Unique identifier for each headword entry.
Headword	Main Sanskrit lexical item.
definition_text	Full explanation or lexicographic content of the entry.
semantic_domain	Domain tag such as RIT, DEI, BOT, MAT, GRA, PHI, MED, SOC, GEO, or LIT.
important_words	Controlled vocabulary items detected inside the definition.
related_words_auto	Automatically extracted candidate related words.

TABLE 1 (*CONTINUED*)

Field	Function
related_words_manual	Scholarly verified related words, to be filled during validation.
relation_basis	Reason for relation: synonymy, co-occurrence, shared domain, citation/context, material relation, or functional relation.
recommendation_score	Numerical strength of the recommendation on a 0–20 scale.
confidence	High, Medium, or Low confidence label.
manual_validation	Automatic first-draft relevance category: Relevant, Partially relevant, or Irrelevant.
scholarly_notes	Automatically generated qualitative observation for later review.
human_validation	Human review label: Correct, Needs Correction, or Wrong.
human_notes	Human reviewer's short observation on the automatic annotation.

4.4 Dataset: Ritual Vocabulary

The full corpus is too large to evaluate completely. Therefore, a controlled pilot dataset was created from the ritual domain. Ritual vocabulary was selected because ritual subject matter is common across a wide range of Sanskrit texts, from Vedic literature to Purā ic and later textual traditions. It is also a domain in which users frequently require related lexical sets rather than isolated word meanings.

The initial ritual query list included twenty terms: अग्नि, यज्ञ, मन्त्र, हवि; आहुति, वेदी, ऋत्विज्/ऋत्विक्, सोम, देवता, हवन, होम, दर्भ, कुश, श्राद्ध, पूजा, व्रत, तर्पण, हविष्य, पुरोडाश, and स्वाहा. Each term was searched in both the headword field and the definition field of the SQLite database. If the searched term appeared in the headword field, the row was marked as a direct_headword_match. If it appeared only in the definition field, the row was marked as a contextual_definition_match. If it appeared in both places, it was marked as both_headword_and_definition_match. This distinction is important because direct entries and contextual occurrences contribute differently to the recommendation model.

The extracted rows were exported into a separate Excel file for validation and into a separate SQLite database for further computational work. This produced a domain-specific ritual-vocabulary pilot corpus derived from the full Śabdakalpadruma database. The extraction procedure also generated summary sheets showing distribution by search term and match type. These summaries help identify whether a term functions as a central ritual node, a specialized ritual item, or a weak contextual trigger. The search dimensions used in the extraction are shown in Table 2.

TABLE 2 Search dimensions used for the ritual-vocabulary pilot dataset

Search dimension	Purpose
Headword search	Finds direct dictionary entries such as यज्ञ, अग्नि, वेदी, or कुश.
Definition search	Finds entries where ritual terms occur contextually inside explanations.
Match-type classification	Distinguishes direct, contextual, and combined matches.
Excel export	Supports manual validation and scholarly review.
SQLite export	Supports co-occurrence analysis and recommender modelling.

Sanskrit query terms used in the pilot are glossed in Table 3. The table is not intended as a full philological description; it simply provides enough semantic orientation for readers to evaluate the domain assignments and the discussion of ambiguous terms.

TABLE 3 Sanskrit ritual query terms used in the pilot

Sanskrit query words	Transliteration	English meaning
अग्नि	agni	fire; fire deity
यज्ञ	yajña	sacrifice; ritual offering
मन्त्र	mantra	sacred hymn; ritual utterance
हविः	havi	oblation
आहुति	āhuti	offering into fire
वेदी	vedī	Altar
ऋत्विज्/ऋत्विक्	ṛtvij/ṛtvik	ritual priest; officiant
सोम	soma	soma plant; ritual drink; deity
देवता	devatā	deity
हवन	havana	fire offering
होम	homa	oblation ritual
दर्भ	darbha	sacred grass
कुश	kuśa	sacred grass
श्राद्ध	śrāddha	ancestral rite
पूजा	pūjā	worship
व्रत	vrata	vow; ritual observance
तर्पण	tarpa a	libation; oblation to ancestors/deceased
हविष्य	haviṣya	oblation material
पुरोडाश	puroḍāśa	sacrificial flour-cake
स्वाहा	svāhā	ritual exclamation used at offering

4.5 Automatic Semantic Annotation

After extraction, the ritual pilot dataset was automatically annotated through a rule-based procedure. A controlled semantic tag-set was used to assign domain labels. The main tags were RIT for ritual/religious vocabulary, DEI for deity/mythological vocabulary, BOT for plant/flora vocabulary, MAT for material objects, GRA for grammatical terminology, PHI for philosophical vocabulary, LIT for literary/poetic usage, MED for medical/Ayurvedic vocabulary, SOC for social/customary vocabulary, and GEO for geographical vocabulary. Combined tags such as RIT-DEI, RIT-BOT, and RIT-MAT were allowed because ritual vocabulary often overlaps with deity names, plant materials, and ritual objects. Table 4 presents the tag-set used in the automatic annotation.

The automatic annotation process examined the headword, definition text, search term, and match type. It detected controlled vocabulary items inside the definition and filled the important_words column. It then generated related_words_auto by selecting domain-relevant lexical items found in the entry. The relation_basis column was filled using categories such as Synonymy, Co-occurrence, Shared domain, Citation/context, Functional ritual relation, and Material relation. This created a preliminary machine-generated annotation layer that could then be checked by the researcher.

TABLE 4 Controlled semantic tag-set used in the automatic annotation

Tag	Meaning	Example
RIT	Ritual/religious vocabulary	यज्ञ, आहुति, हवन
DEI	Deity/mythological vocabulary	अग्नि, देवता, इन्द्र
BOT	Plant/flora vocabulary	कुश, दर्भ, पुष्प
MAT	Material object	वेदी, पात्र, आसन
GRA	Grammar/linguistic terminology	विभक्ति, धातु
PHI	Philosophy	आत्मन्, ब्रह्म
LIT	Literary/poetic usage	अलंकार, रस, मन्त्र
MED	Medicine/Ayurveda	औषधि, रोग
SOC	Social/customary vocabulary	श्राद्ध, विवाह
GEO	Place/geographical vocabulary	तीर्थ, नदी

4.6 Recommendation Scoring

The recommender model was designed as an explainable scoring system rather than a black-box neural model. Each candidate recommendation receives a score based on five signals: synonymy, co-occurrence, shared semantic domain, citation/context proximity, and match/confidence strength. The maximum score is 20. Synonymy receives up to 5 points, co-occurrence receives up to 5 points, shared semantic domain receives up to 5 points, citation/context receives up to 3 points, and match/confidence receives up to 2 points. This scale allows the system to distinguish strong lexical recommendations from weak contextual associations.

For example, in a query such as यज्ञ, words such as अग्नि, हवि; वेदी, मन्त्र, ऋत्विज्, देवता, and आहुति may be recommended because they belong to the same ritual field and occur repeatedly in relevant lexicographic contexts. In a query such as अग्नि, words such as वह्नि, अनल and पावक may receive stronger scores if they are detected as synonymic or near-synonymic lexical items. The system therefore distinguishes direct equivalence, functional ritual relation, and broader domain association. The scoring signals used for these distinctions are shown in Table 5.

TABLE 5 Recommendation scoring signals in the pilot model

Signal	Maximum score	Function
Synonymy	5	Rewards direct synonym or equivalent relation.
Co-occurrence	5	Rewards repeated occurrence in the same or related definitions.
Shared semantic domain	5	Rewards belonging to the same domain such as RIT or RIT-DEI.
Citation/context proximity	3	Rewards occurrence in citation-rich or textually connected contexts.
Match/confidence strength	2	Rewards direct headword or strong combined match.
Total	20	Maximum possible recommendation score.

4.7 Control Mechanisms and Quality Assurance

Several control mechanisms were introduced to prevent uncontrolled or misleading recommendations. First, OCR noise was reduced through post-OCR cleaning and manual inspection of the larger corpus. Second, the extraction process distinguished direct headword matches from contextual definition matches, thereby preventing all occurrences from being treated equally. Third, a controlled vocabulary list and a stop-word list were used so that common particles, grammatical fragments, and irrelevant high-frequency forms did not dominate the recommendation output. Fourth, the semantic-domain tag-set limited over-generation by forcing the system to classify suggestions within explicit lexicographic categories. Fifth, the system assigned confidence levels: High, Medium, or Low. Sixth, manual validation was made a required stage after automatic annotation. These quality-control mechanisms are summarized in Table 6.

TABLE 6 Quality-control mechanisms used to reduce recommendation errors

Risk	Control mechanism
OCR noise	Post-OCR cleaning and manual inspection of corpus samples.
Duplicate guide words or headers	Removal or filtering during corpus preparation and validation.
Accidental lexical occurrence	Distinction between direct headword and contextual definition matches.
Common irrelevant words	Stop-word filtering.
False semantic relation	Manual validation of the pilot sample.
Over-generation by AI	Rule-based filtering before AI-assisted expansion.
Weak recommendations	High, Medium, and Low confidence labels.
Missed relevant lemmas	Future recall testing through variant expansion and independent expert seed lists.

The control strategy is especially important because Sanskrit lexicographic entries can be long, citation-heavy, and semantically layered. A word may appear inside an entry without being a meaningful recommendation for the user. Therefore, automatic co-occurrence is treated as a signal, not as final proof of semantic relation. The final recommender model depends on the combination of automatic extraction and human scholarly validation.

4.8 Recall and Missed-Lemma Limitation

The pilot evaluation primarily measures precision, that is, whether the retrieved entries were correctly annotated and recommended. It does not claim to measure full recall across the entire 42,824-entry corpus. This limitation is important because Sanskrit lexical material may appear through orthographic variants, inflected forms, sandhi forms, synonymic alternatives, and OCR-distorted forms. Exact or substring matching against twenty seed terms can therefore miss entries that belong to the ritual field but do not contain one of the selected query terms. In the present pilot, this risk was partly reduced by including variant query forms such as ऋत्वजि्/ऋत्वकि् and by searching both headword and definition fields. A complete recall evaluation would require a manually prepared gold-standard list of ritual-domain entries. Future work will address this by expanding seed lists, adding variant spellings and transliteration-normalized search, testing synonym-based query expansion, and auditing unmatched entries. The present results should therefore be read as evidence for first-stage precision and usefulness, not as a claim of complete ritual-domain coverage.

4.9 Validation Sampling

The evaluation was designed as a small but measurable pilot study. After automatic annotation, the dataset was not validated by simply reading the first one hundred rows. Since Śabdakalpadruma is alphabetically arranged, a first-100-row validation would produce an alphabetical bias. For example, after sorting by entry identifier, the first entries may belong predominantly to words beginning with अ or may be strongly linked with a single query term such as अग्नि. To avoid this, the study used a stratified validation sample of 100 rows, distributed across the twenty ritual search terms. This ensured that terms such as अग्नि, यज्ञ, मन्त्र, सोम, कुश, वेदी, श्राद्ध, पूजा, and स्वाहा were represented in the validation set.

The validation sample included 5 rows from each ritual search term. Two human-review columns were added: human_validation and human_notes. The human_validation column used three labels: *correct, needs correction*, and *wrong*. *Correct* means that the automatic annotation is acceptable; *needs correction* means that the annotation is partially useful but requires editing; *wrong* means that the row is not suitable or the annotation is misleading. The human_notes column recorded brief qualitative comments, such as whether a semantic tag should be changed, whether related words were too broad, or whether the row represented an accidental keyword match. Table 7 lists the evaluation elements used in the pilot.

TABLE 7 Evaluation elements for the pilot recommender model

Evaluation element	Measurement
Extraction scale	Rows retrieved from the full 42,824-entry database.
Search-term distribution	Rows retrieved for each ritual query term.
Match-type distribution	Direct, contextual, and combined matches.
Annotation quality	Correct, Needs Correction, and Wrong labels in the stratified sample.
Recommendation strength	Average and distribution of 0–20 recommendation scores.
Confidence profile	High, Medium, and Low confidence counts.
Error analysis	Human notes explaining false or weak recommendations.
Recall Limitation	Missed lemmas acknowledged as a future evaluation requirement.

4.10 Python Workflow

The Python workflow produced three primary files at the extraction stage: i. an excel file containing the extracted rows and annotation columns for validation; ii. a dababase file to store the same pilot corpus in SQLite form; iii. another excel file summarized the extraction by search term and match type. After automatic annotation, the workflow produced annotated version of the excel file containing extracted data and a database file for the same. A sampling excel file named Ritual_validation_sample_100.xlsx was produced for stratified sample human review.

Figure 1 should be read as a methodological schematic rather than a software interface. Steps 1–8 represent the Python-based workflow: database inspection, query list preparation, retrieval, match classification, export, and rule-based annotation. Step 9 is the product of this workflow, namely the stratified validation sample. Step 10 is deliberately outside the Python workflow because it represents human intervention. Step 11 receives the human-validated output and can be used for manual evaluation, Python-based evaluation, and future machine-learning training. This staged design makes the model reproducible and avoids presenting AI as a black-box replacement for lexicographic expertise.

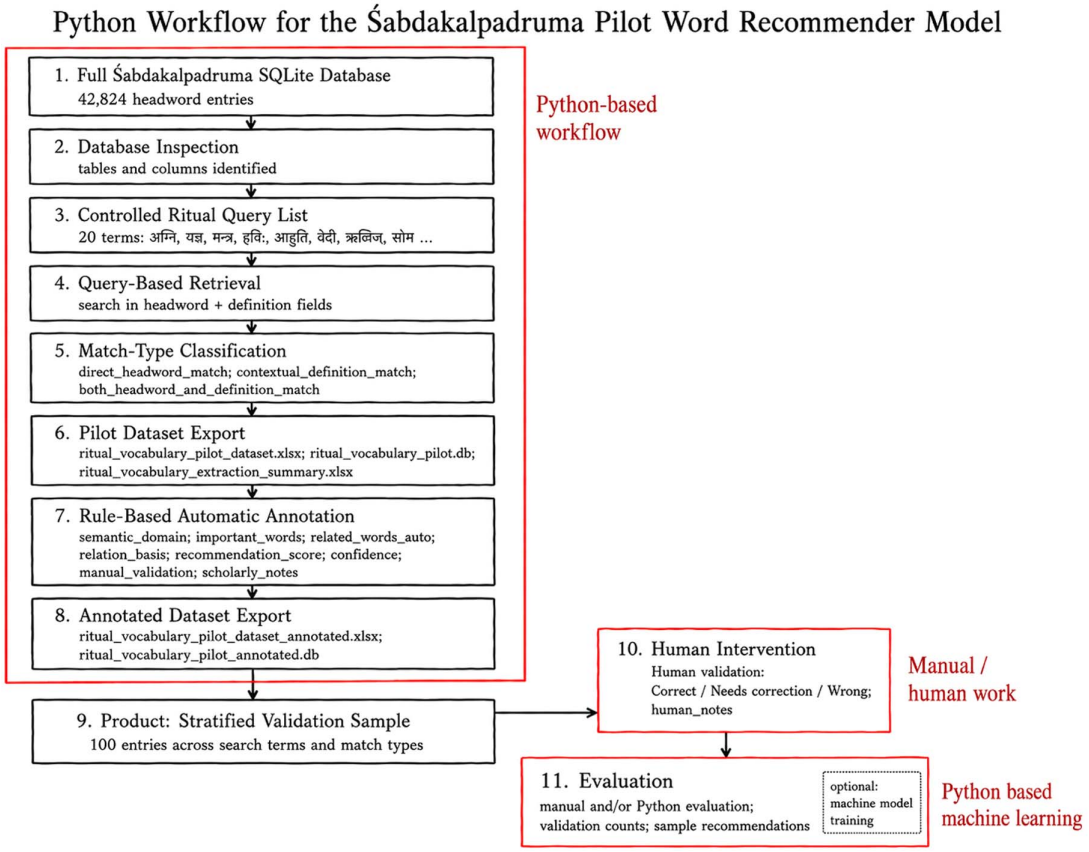


FIGURE 1 Python-based workflow from full Śabdakalpadruma database to ritual pilot extraction, automatic annotation, stratified validation sample, human validation, evaluation, and future machine-prediction model

5 Results and Evaluation

The pilot evaluation was conducted on a stratified sample of 100 entries. The sample was drawn from the annotated ritual-vocabulary dataset, with 5 entries randomly extracted for each of the 20 ritual search terms. This design avoided alphabetical bias and ensured representation across the query terms. The validation sample consisted mainly of contextual definition matches, along with a smaller number of direct headword matches and combined headword-definition matches. The structure of the validation sample is presented in Table 8.

TABLE 8 Structure of the stratified validation sample

Item	Result
Total validation entries	100
Search terms represented	20
Entries per search term	5
Contextual definition matches	92
Direct headword matches	3
Both headword and definition matches	5

As shown in Table 9, The human validation results showed that 67 entries were accepted as Correct, 19 were marked as Needs Correction, and 14 were marked as Wrong. This means that

67% of the automatically generated annotations were accepted without correction. A further 19% were not fully correct but were still usable after scholarly correction. Therefore, 86% of the model output was lexicographically useful at the pilot stage.

TABLE 9 Human validation results

Human validation	Count	Percentage
Correct	67	67%
Needs Correction	19	19%
Wrong	14	14%
Total	100	100%

These results indicate that the model is effective as a first-stage semantic recommendation and annotation system. It is not yet a fully autonomous lexicographic AI system, but it can identify useful lexical candidates with reasonable reliability. The 19% Needs Correction category is especially important because it shows that the automatic output is often not useless when imperfect; rather, it can function as a machine-generated draft requiring expert refinement.

The match-type analysis gives a more detailed view of the model's performance, as shown in Table 10. Direct headword matches and combined headword-definition matches showed full correctness in the validation sample. All 3 direct headword matches and all 5 combined matches were judged Correct. The contextual definition matches were more mixed: 59 Correct, 19 Needs Correction, and 14 Wrong out of 92 cases.

TABLE 10 Human validation by match type

Match type	Correct	Needs Correction	Wrong	Total
Direct headword match	3	0	0	3
Both headword and definition match	5	0	0	5
Contextual definition match	59	19	14	92
Total	67	19	14	100

This confirms that structural match type is an important predictor of recommendation reliability. When a ritual term occurs as the headword or both as headword and definition-field item, it is more likely to be lexicographically central. When it appears only contextually inside a definition, it may still be useful, but the relation requires filtering.

Search-term performance also showed variation, as summarized in Table 11. Terms such as देवता, श्राद्ध, and हविः performed especially well, with all 5 sampled entries judged Correct. Other terms such as यज्ञ, आहुति, पूजा, तर्पण, व्रत, हविष्य, स्वाहा, and होम also produced strong results. In contrast, terms such as मन्त्र, वेदी, कुश, सोम, पुरोडाश, दर्भ, and हवन required more correction or showed higher ambiguity.

This variation is a valuable result rather than a weakness. It shows that not all ritual terms behave in the same way inside the lexicographic corpus. Some terms are strongly domain-specific, while others are semantically mobile. मन्त्र, for example, may appear in ritual, literary, textual, and sometimes grammatical contexts. वेदी may function as a ritual object but also as a material or spatial term कुश and दर्भ may belong simultaneously to ritual and botanical

TABLE 11 Search-term performance in the validation sample

Search term	Transliteration	Human validation pattern
देवता	devatā	5 Correct
श्राद्ध	śrāddha	5 Correct
हवि:	havi	5 Correct
यज्ञ	yajña	4 Correct, 1 Wrong
आहुति	āhuti	4 Correct, 1 Wrong
पूजा	pūjā	4 Correct, 1 Wrong
तर्पण	tarpa a	4 Correct, 1 Wrong
व्रत	vrata	4 Correct, 1 Needs Correction
हविष्य	haviṣya	4 Correct, 1 Needs Correction
स्वाहा	svāhā	4 Correct, 1 Wrong
होम	homa	4 Correct, 1 Wrong
अग्नि	agni	3 Correct, 2 Needs Correction
ऋत्विज्/ऋत्विक्	ṛtvij/ṛtvik	3 Correct, 2 Needs Correction
मन्त्र	mantra	1 Correct, 4 Needs Correction
वेदी	vedī	2 Correct, 3 Needs Correction
कुश	kuśa	2 Correct, 3 Wrong
सोम	soma	2 Correct, 2 Needs Correction, 1 Wrong
दर्भ	darbha	2 Correct, 2 Needs Correction, 1 Wrong
पुरोडाश	puroḍāśa	2 Correct, 2 Needs Correction, 1 Wrong
हवन	havana	3 Correct, 2 Wrong

domains. सोम may refer to deity, plant, ritual substance, or textual context. These cases demonstrate the need for domain-sensitive filtering rather than simple keyword matching.

The automatic validation column also provided useful evidence. The model originally marked 78 entries as Relevant, 16 as Partially relevant, and 6 as Irrelevant. Human review showed that the automatic relevance classification was useful as a first-pass filter but required correction in several cases. All entries automatically marked Irrelevant were rejected in human review, showing that the model's rejection logic was useful. However, some automatically Relevant entries were judged Wrong or Needs Correction, showing that broad ritual relevance must be distinguished from precise lexicographic recommendation.

A further insight concerns semantic-domain tagging. The automatic system overused broad combined labels, especially RIT-DEI-BOT. This indicates that the first version of the tagging rules was good at detecting ritual-related lexical density but less successful in assigning fine-grained secondary tags. Human notes showed that many entries required more precise labels such as RIT-DEI-LIT, RIT-SOC-LIT, RIT-GRA, RIT-MAT, RIT-PHI-LIT, RIT-GEO, or RIT-MAT-LIT. This suggests that future versions of the model should reduce automatic multi-tagging and introduce more selective rules for secondary domains.

The recommendation score also requires recalibration. In the present version, some incorrect or correction-needed entries received scores comparable to correct entries. This shows that the score currently measures lexical density and broad domain overlap more than final

recommendation accuracy. A revised scoring system should give greater weight to direct headword match, confirmed semantic-domain fit, synonymy, and repeated co-occurrence across multiple entries. It should reduce the weight given to one-time contextual occurrence and broad domain overlap of ritual terms.

The pilot evaluation also clarifies what remains unmeasured. The results measure the precision of retrieved and annotated entries, not the recall of the entire ritual field. A future recall-oriented evaluation should begin from an independently prepared ritual vocabulary list and test how many of those lemmas are retrieved by the query system. This would allow the model to address the silent-omission problem in addition to the incorrect-annotation problem addressed here.

6 Discussion

The results show that a historical Sanskrit lexicon can be converted into a corpus-driven recommendation environment, but they also show that recommendation quality depends on careful control. The model performs well as a first-stage extraction and annotation system. Its strongest performance appears in direct headword matches and combined headword-definition matches. It is weaker when ritual terms appear only contextually inside longer definitions. This confirms that co-occurrence is useful, but not sufficient by itself.

The study's contribution is therefore both technical and theoretical. Technically, the ritual-vocabulary pilot shows how a large database can be reduced into a controlled experimental corpus. The stratified validation sample shows how the output can be evaluated with lexicographic insight. Theoretically, the study argues that AI-age lexicography should not abandon historical lexicons. When converted into digital corpora, these lexicons can become valuable raw material for training models and advancing AI-age lexicography. A recommender system reorganizes this depth according to contemporary user needs and reconnects digital lexicography with older Sanskrit thematic traditions of lexical grouping. The model does not replace the traditional lexicon; it makes selected aspects of its semantic structure more visible to users.

The AI component of the study should be understood carefully. The present model is not an unverified black-box AI system. It is a hybrid rule-based and human-validated model. At the first stage, corpus extraction and rule-based annotation identify candidate ritual vocabulary from the Śabdakalpadruma database. At the second stage, human validation classifies the output as Correct, Needs Correction, or Wrong. These labels do not merely evaluate the model; they produce labelled data for future machine-learning work.

The manually validated sample therefore has two functions. It measures the effectiveness of the present recommender model, and it supplies training logic for the next stage. The 67 Correct cases show where the rule-based model already works. The 19 Needs Correction cases show where the model is useful but needs rule refinement. The 14 Wrong cases provide negative examples that can help a future classifier learn what to reject. In this way, human validation becomes part of the AI pipeline rather than an external correction after the fact.

On the basis of the pilot results, the next version of the scoring and tagging rules should be revised. Direct headword and combined matches should receive greater weight. Contextual definition matches should receive lower initial weight unless supported by repeated co-occurrence, synonymy, or strong domain evidence. Broad multi-domain labels should be reduced, and secondary tags should be assigned more selectively. Ambiguous terms such as मन्त्र, वेदी, कुश, सोम, and दर्भ require special rules because they cross ritual, botanical, literary, material,

TABLE 12 Revisions suggested by the pilot evaluation

Problem observed	Revision needed
Overuse of broad combined tag RIT-DEI-BOT	Reduce automatic multi-tagging and apply secondary tags more selectively.
Wrong rows receiving high scores	Lower the score for weak contextual matches.
Direct matches showing full reliability	Give greater weight to direct headword and combined matches.
Contextual matches producing mixed results	Require stronger evidence before marking as Relevant.
Ambiguous terms such as कुश, दर्भ, सोम	Add domain-specific filtering for botanical, deity, and ritual-substance contexts.
मन्त्र and वेदी requiring correction	Add textual/material-context subrules.
Recall not yet measured	Develop a standard ritual vocabulary list and test missed lemmas.

and deity-related domains. Table 12 summarizes the main revisions suggested by the pilot evaluation.

This leads to a revised understanding of the model. The current stage is not the final AI system; it is the corpus-driven and human-validated foundation for it. The study demonstrates how a structured lexicographic corpus can generate recommendations, how human expertise can evaluate them, and how validated results can be used to train future machine prediction.

7 Conclusion

This paper has presented a user-oriented digital transformation of Śabdakalpadruma. The study is grounded in Śabdakalpadruma's digitized corpus of 42,824 headword entries. On this foundation, it develops a ritual-vocabulary pilot dataset and proposes a transparent recommendation model based on headword search, semantic tagging, co-occurrence analysis, relation scoring, confidence labelling, and manual validation.

The results of the stratified 100-entry validation sample show that 67% of the automatically generated annotations were correct (contexually related), 19% were usable with correction, and 14% were wrong (contexually unrelated). Thus, 86% of the output was lexicographically useful at the pilot stage. Direct headword matches and combined headword-definition matches showed promising reliability in the sample, while contextual definition matches required greater filtering. The results also showed that the first version of the semantic tagging system over-generated broad multi-domain labels and that the scoring system needs recalibration.

The main contribution of this study is to show that specialized dictionaries can move beyond static alphabetical lookup and become active lexical systems that guide users toward relevant semantic fields. In this sense, the study fulfils the claim of corpus-driven targeted dictionary design by showing how a structured lexicographic corpus can generate domain-specific lexical recommendations, how these recommendations can be validated by human expertise, and how the validated results can guide future machine-learning models for semantic prediction.

This study therefore demonstrates a staged pathway for corpus-driven targeted dictionary design: a digitization of lexicographic resources, to a rule-based recommender model; the recommender output is then evaluated through human validation; and the resulting labelled data becomes the basis for future AI-assisted semantic prediction. Future work will expand the

validation sample, refine the scoring and tagging rules, include more semantic domains such as botanical, philosophical, and grammatical vocabulary, and develop a supervised prediction model trained on human-validated recommendation data.

References

Atkins, S. B. T. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. Oxford University Press.

Bergenholtz, H. & Tarp, S. (Eds.). (1995). *Manual of Specialised Lexicography: The Preparation of Specialised Dictionaries*. John Benjamins.

Dev, R. R. (2018). *Śabdakalpadruma*. Rashtriya Sanskrit Sansthan.

Halliday, M. A. K., Teubert, W., Yallop, C. & Čermáková, A. (2004). *Lexicology and Corpus Linguistics: An Introduction*. Continuum.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. & Suchomel, V. (2014). The Sketch Engine: Ten Years On. *Lexicography*, 1, 7–36.

McEnery, T. & Brezina, V. (2022). *The Fundamental Principles of Corpus Linguistics*. Cambridge University Press.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford University Press.

Tarp, S. (2008). *Lexicography in the Borderland between Knowledge and Non-Knowledge: General Lexicographical Theory with Particular Focus on Learner's Lexicography*. Max Niemeyer.

Vogel, C. (1979). *Indian Lexicography*. Otto Harrassowitz.

Contact information

Deepak Solanki
Centre of Indian Languages, School of Language, Literature and Culture Studies, Jawaharlal Nehru University, New Delhi
deepsoul.solanki@gmail.com

Anna Vacalopoulou

A DEFINITION WALKS INTO A DICTIONARY

The Lexicography of Humor

Abstract Different disciplines approach the definition of humor from different angles: philosophy through theoretical frameworks that explain why people find things funny, psychology through behavioral and dispositional dimensions that describe how humor is experienced and expressed. This paper examines the treatment of humor in terms of its lexicographic definition, drawing on entries from six Greek dictionaries currently in circulation. Through systematic microstructural analysis, definitional components are identified and categorized, revealing significant variation in how dictionaries conceptualize humor. Definitions are broken down into components that reveal lexicographers' attitudes towards humor's nature (what it is), manifestation (how it appears), subject (what its target is), features (what characterizes it), and function (what its goal is). Results demonstrate that while lexicographers employ diverse definitional strategies, ranging from analytical to functional approaches, they align on certain core elements, particularly humor's cognitive dimension and social function. This research proposes a systematic methodology for analyzing how dictionaries handle abstract, culturally embedded concepts, providing a replicable framework applicable to other affective, cognitive, or socially constructed concepts such as irony, empathy, and creativity.

Keywords Lexicography; dictionaries; abstract concept definition; microstructural analysis; humor studies

1 Introduction

The lexicographic treatment of abstract concepts presents challenges that distinguish it from defining concrete objects or actions. Unlike terms for physical entities or observable processes, abstract psychological concepts such as humor resist straightforward definition, as they involve cognitive, emotional, social, and cultural dimensions and are shaped by context. Therefore, lexicographic choices in defining them reflect not only usage but also implicit theoretical assumptions about the phenomena.

Humor illustrates these difficulties particularly clearly. Although it has been extensively studied in philosophy, psychology, linguistics, and anthropology, and is well theorized in models ranging from classical superiority theories to contemporary cognitive approaches, these frameworks rarely appear in general-purpose Greek dictionary definitions, which prioritize accessibility over theoretical detail.

This study examines the microstructural treatment of humor in six contemporary dictionaries, addressing four questions: what definitional styles are used; what conceptual components are emphasized; how abstract meaning is balanced with linguistic description; and what patterns of similarities and differences emerge across dictionaries. The analysis is synchronic, focusing on six current general-language collegiate dictionaries and excluding specialized, bilingual, or learner's dictionaries. It also considers definitional content alongside microstructural elements such as usage notes, examples, and collocations, in order to show how dictionaries construct meaning beyond the definition itself.

2 Theoretical perspectives and lexicographic approaches

Philosophy and psychology approach the definition of humor from different angles: the former through theoretical frameworks that explain why we find things funny, the latter through behavioral and dispositional dimensions describing how humor is experienced and expressed. According to the *Internet Encyclopedia of Philosophy* (Smuts, n.d.), philosophical inquiry categorizes humor under three major frameworks: superiority theories, traceable to Plato and Aristotle, which view humor as arising from perceiving others' inferiority; incongruity theories, developed by Kant, Schopenhauer, and contemporary cognitive scientists, which emphasize the perception of logical or conceptual incongruity; and relief theories, associated with Freud and Spencer, which characterize humor as releasing psychological tension. From a psychological perspective, humor resists reduction to a single dimension.

Beyond its theoretical complexity, humor is deeply context-dependent (Roeckelein, 2002, pp. 3–4). As Tsakona (2020) argues, humor surfaces in numerous and diverse contexts that simultaneously determine how it works, its form, and its functions and consequences for interlocutors. This contextual variability has important cross-cultural dimensions: what speakers find funny, how humor is performed, and what social functions it serves vary significantly across communities and cultural settings (Carrell, 1997), This insight is directly relevant to any assessment of dictionary definitions, which by their nature take concepts out of context and present them as fixed, universal categories.

Lexicographic definitions differ fundamentally from philosophical or scientific ones. As Landau (2001, pp. 153–154) notes, dictionaries describe language use rather than establish truths about reality. Mel'čuk and Polguère (2018, p. 443) capture this tension in the slogan "Think of the word, not of the thing." A central challenge is determining what counts as word meaning. Jackson (2002, pp. 15–16) identifies meaning as an interaction of reference, associations, lexical relations, and collocation, noting that reference is especially unclear for abstract concepts. Humor belongs to this category, making its lexicographic treatment structurally difficult.

Traditional theory distinguishes definitional types such as analytical, functional, descriptive, enumerative, and synonymic. For abstract psychological terms like humor, no single strategy is sufficient (Atkins & Rundell, 2008, pp. 446–448). As Rundell (2006, p. 333) observes, dictionaries often fail to capture non-denotative meaning. Hanks (2016, pp. 100–101) further argues that analytical definitions break down for abstract entities, since they resist hierarchical classification and are characterized by vagueness, context dependence, and prototypicality rather than necessary and sufficient conditions. A further limitation is that dictionaries emphasize observable effects ("amusing," "funny") rather than underlying mechanisms, and tend to present humor as universal, overlooking cross-cultural variation.

Research on abstract lexical concepts in dictionaries has mainly come from semantics and linguistic anthropology. Wierzbicka (1996, 1999) and Goddard (1998, pp. 1–3, 96) show that abstract concepts vary across languages and encode culture-specific assumptions. These studies inform the present componential approach. However, no prior work has systematically analyzed humor across multiple contemporary dictionaries using a unified framework, making this study a novel contribution to lexicography and humor research.

3 Methodology and Analysis

The study focuses on six general language collegiate dictionaries currently in circulation (mentioned as 1 through 6), representing a range of editorial approaches and publication formats.

The analytical framework developed identifies five conceptual components that definitions may include: nature (what humor fundamentally is); manifestation (how humor appears or is expressed); subject (what humor targets or is about); features (what characterizes humor); and function (what humor aims for). This framework draws on methods of componential analysis used in lexicographic and semantic research (Zgusta, 1971, pp. 91–94; Goddard, 1998, pp. 43 on). For each dictionary the primary definition of humor was extracted, its components were identified, and specific lexical and syntactic choices in each component were noted. They were then classified by definitional style and compared across dictionaries to identify similarities and differences. The six definitions are summarized in Table 1.

4 Results

The definitions employ primarily analytical and functional styles. #1 and #2 combine analytical elements with descriptive elaboration of characteristics; #3, #4, and #5 emphasize capacities with functional components describing what these achieve; #6 adopts a more explicitly functional approach, emphasizing what humor does rather than what it is. None employs purely synonymic, enumerative, or full-sentence styles, reflecting the challenge of defining abstract phenomena.

All definitions specify humor's fundamental nature, though with significant variation. #1 and #2 frame it as a mental disposition; #3, #5, and #6 as an ability or capacity; #4 characterizes it as perception combined with expression. These choices implicitly engage with psychological debates about the different conceptualizations of humor (Martin & Ford, 2018, pp. 192–195). #2 introduces a further framing, i.e., "witty jest", suggesting humor may refer to the product rather than the capacity, a genuine polysemy that the remaining definitions evade.

Five definitions specify manifestation. #4's explicit mention of spoken and written expression acknowledges humor's multimodal nature; #6 uniquely recognizes humor as involving both production ("tell jokes") and comprehension ("understand others' jokes"), an insight more commonly found in linguistic accounts (Attardo, 1994) than in lexicographic ones.

Four definitions specify subject, ranging from "situations or people" (#1) to "life and events" (#4); #6 acknowledges "even unpleasant or serious topics" aligning with theoretical accounts of humor as a coping mechanism (Martin & Ford, 2018); #2 and #5 omit subject matter entirely, reflecting a more formal, mechanism-oriented conception.

All definitions include characteristic features, with intellectual qualities prominent throughout: "intelligence" (#1, #4), "witty" (#2, #4, #6), "subtlety" (#1), and "inventive" (#5) collectively position humor as cognitively sophisticated. The repeated reference to irony, specified as "harmless" or "subtle", implies a normative dimension distinguishing humor from mockery. Definition #5's unique reference to "incongruous elements" explicitly invokes incongruity theory (Raskin, 1985; Attardo, 1994), representing the most theoretically engaged feature description in the sample.

Four definitions specify function, dividing into cognitive-perceptual types (highlighting aspects of reality) and social-emotional types (bringing amusement, evoking atmosphere). #6's inclusion of "biting" alongside "cheerful, pleasant, and witty" acknowledges humor's capacity to serve aggressive or critical functions, consistent with superiority theory (Smuts, n.d.).

Despite surface differences, the definitions describe humor in similar ways: as a mental ability, a sign of intelligence, something that often uses irony, involves other people, and is generally

TABLE 1 A comparative presentation of the different components in the six definitions (own translations)

#	NATURE	MANIFESTATION	SUBJECT		FEATURES				FUNCTION
1	a disposition of the mind	that presents	situations	or people	with intelligence,	grace, and subtlety			highlighting the funny, comic, or even ridiculous aspects of life
2	witty jest or cheerful disposition	expressed through						harmless irony	which brings others amusement or a pleasant mood
3	the ability	to portray	various situations			with apparent seriousness and indifference,	beneath which lies satire	and harmless irony	to face reality with a cheerful attitude
4	the perception		of life and events				from a funny and comic perspective		
	as well as the expression of this view	—whether spoken or written—			in an intelligent and witty manner				
5	the ability	to provoke laughter in others			through the inventive combination of incongruous elements			and/or the use of subtle irony	
6	the ability	to tell jokes by presenting	even unpleasant or serious topics				in a comic way		evoking a cheerful, pleasant, biting, or witty atmosphere in others,
		and to understand others' jokes							

positive or harmless. Together, these ideas form a common everyday understanding of humor. Significant differences include degree of theoretical engagement (#5 explicitly invokes incongruity theory; others remain atheoretical), scope (#6 includes receptive capacity), complexity (23 to 40 words), and balance of components.

5 Discussion and conclusions

Analysis of the treatment of humor in the sample dictionaries reveals a fundamental tension between ontological and lexicographic approaches to its definition. Philosophers and psychologists ask, "What is humor really?" Lexicographers ask, "How do speakers use the word humor?" These questions occasionally align (the employment of incongruity theory in #5 being the clearest instance) but such agreement is neither necessary nor expected and may not even be desirable if theoretical precision comes at the cost of practical accessibility.

The comprehensiveness-conciseness trade-off is clearly illustrated in the research sample. At 40 words, #6 is the most complete definition, including both production and comprehension, acknowledging varied affective tones, and recognizing that humor can address serious topics. This completeness, however, comes at the cost of complexity. At 23 words, #5 achieves conciseness by focusing on mechanism and function, but excludes important dimensions. There is no consensus on the ideal balance, since decisions are shaped differently by dictionary type, intended audience, and editorial philosophy.

Cultural and linguistic specificity remains underaddressed. Research in linguistic anthropology demonstrates that emotion and attitude terms often lack perfect cross-linguistic equivalents (Wierzbicka, 1999), and features such as the emphasis on irony, or the characterization of humor as involving "grace" and "subtlety" may reflect culturally specific communicative values rather than universal properties. None of the definitions explicitly addresses cultural variation. It remains to be examined whether this is a reasonable simplification or a meaningful omission.

For lexicographic practice, findings suggest that explicit component analysis in drafting and evaluating definitions of abstract concepts is effective. Cautious incorporation of theoretical insights, such as the incongruity theory, can enhance precision without sacrificing accessibility. For terms that express multiple related senses, dictionaries should either recognize the polysemy through distinct definitions or develop definitions that adequately cover the spectrum of usage.

Ultimately, the diversity of humor definitions across the sample reflects humor's genuine complexity as a simultaneously cognitive, emotional, social, and cultural phenomenon. Perhaps no single definition can adequately capture all that humor is and does; perhaps it is, as Attardo claims (1994, pp. 3–7), an "impossible definition." In any case, the lexicographer's task is not to provide perfect definitions but to create definitions that help users make sense of such complex concepts, despite the inevitable incompleteness involved in describing such a dynamic and context-dependent human phenomenon. Future work could extend the component framework to related concepts such as irony, wit, and empathy, and incorporate corpus-based validation to assess whether the features emphasized by lexicographers align with attested patterns of use.

References

Atkins, S. B. T. & Rundell, M. (2008). *The Oxford Guide to Practical Lexicography*. OUP.

Attardo, S. (1994). *Linguistic theories of humor*. Mouton de Gruyter.

Carrell, A. (1997). Humor communities. *Humor: International Journal of Humor Research*, 10(1), 11–24.

Goddard, C. (1998). *Semantic analysis: A practical introduction*. OUP.

Hanks, P. (2016). Definition. In P. Durkin (Ed.), *The Oxford handbook of lexicography*. OUP.

HLNG. (2014). *Χρηστικό λεξικό της νεοελληνικής γλώσσας*. Charalambakis C. (Ed.). Athens: Academy of Athens.

Jackson, H. (2002). *Lexicography: An introduction*. Routledge.

Landau, S. I. (2001). *Dictionaries: The art and craft of lexicography* (2nd ed.). CUP.

LKN (1998). *Λεξικό της κοινής νεοελληνικής*. Institute for Modern Greek Studies, Manolis Triandaphyllidis Foundation. Aristotle University of Thessaloniki.

LNEG (2019). *Λεξικό της νέας ελληνικής γλώσσας*. (5th ed.). Babiniotis, G. (Ed.). Athens: Lexicology Center.

Martin, R. A., & Ford, T. (2018). *The psychology of humor: An integrative approach* (2nd ed.). Academic Press.

MEL (1997). *Μείζον ελληνικό λεξικό*. Mandala, M. (Ed.). Athens: TegopoulosFytrakis.

Mel'čuk, I. & Polguère, A. (2018). Theory and practice of lexicographic definition. *Journal of Cognitive Science*, 19(4), 417–470.

MILNEG (n.d.). *Μεγάλο ηλεκτρονικό λεξικό της νεοελληνικής γλώσσας*. Patakis. https://lexicon-neohel.patakis.gr/el/lexikon/ (Accessed 2025-02-05).

NEL (1995). *Νέο ελληνικό λεξικό*. Kriaras, E. (Ed.). Athens: Ekdotiki Athinon.

Raskin, V. (1985). *Semantic mechanisms of humor*. D. Reidel.

Roeckelein, J. E. (2002). *The psychology of humor: A reference guide and annotated bibliography*. Greenwood.

Rundell, M. (2006). More than one way to skin a cat: Why full-sentence definitions have not been universally adopted. In E. Corino, C. Marello, & C. Onesti (Eds.), *Proceedings of the 12th EURALEX International Congress* (pp. 323–337). Edizioni dell'Orso.

Smuts, A. (n.d.). Humor. *Internet Encyclopedia of Philosophy*. https://iep.utm.edu/humor/

Tsakona, V. (2020). *Recontextualizing humor: Rethinking the analysis and teaching of humor* (Language Play and Creativity, Vol. 4). De Gruyter Mouton.

Wierzbicka, A. (1996). *Semantics: Primes and universals*. OUP.

Wierzbicka, A. (1999). *Emotions across languages and cultures: Diversity and universals*. CUP.

Zgusta, L. (1971). *Manual of lexicography*. Mouton.

Contact information

Anna Vacalopoulou
Athena Research Center/Institute for Language and Speech Processing (ILSP)
avacalop@athenarc.gr

Acknowledgements

An earlier version of the research discussed in this paper was presented at the 35th ISHS conference in Kraków, Poland, on 10 July 2025 under the title "The serious business of defining humor: A dictionary-based inquiry."

INDEX OF AUTHORS

EURALEX 2026



INDEX OF AUTHORS

A

Abdelzaher, Esra 888
Abel, Andrea 3
Akinremi, Luke Omoyemi 1006
Alamdar, Abolfazl 584
Arčon, Tjaša 132
Arias-Arias, Iván 16, 245

B

Bago, Petra 493
Bajčić, Martina 1077
Beliga, Slobodan 729
Berthemet, Elena 706
Bichlmeier, Harald 508
Billero, Riccardo 1087
Black Bear, Jr., Ben 902
Blahuš, Marek 143, 1100
Boguslavsky, Igor 34
Bothma, Theo J.D. 413
Brač, Ivana 920
Brankačkec, Katja 932
Breathnach, Cormac 1019
Burkart, Rukiye Smilla 376

C

Camilleri, John J. 607
Caruso, Valeria 347, 1032
Chen, Lian 390
Chirino, Dalí 862
Cimander, Luisa 517
Colman, Lut 862
Cui, Li 46

D

Dao, Huy-Linh 390
De Meo, Anna 347
de Oliveira Maia-Pires, Flávia 405
de Schryver, Gilles-Maurice 58, 88
Delaporte, Alexander 390
Dikonov, Vyacheslav 34
Dinić Marinković, Milica 295
Domínguez Vázquez, María José 16, 245
Dringó-Horváth, Ida 648
Ducceschi, Luca 3

E

Eder, Elisabeth 117
Ellul, Dwayne 607
Elsner, Daniel 117
Eterović, Ivana 493

F

Fiedler, Sophia 717
Filipović Petrović, Ivana 729
Fischer, Hanna 103
Flörke, Ida 310
Freixas, Margarita 523
Frolova, Tatiana 34
Fuentes Morán, M.ª Teresa 217

G

Ganslmayer, Christine 661
Gantar, Polona 132
Garcia-Casado, Marta 743
Garriga Martínez, Pol 273
Gavriilidou, Zoe 676, 1043
Geyken, Alexander 758
Gouws, Rufus H. 413
Grasmanis, Mikus 788

H

Heinisch, Barbara 1107
Herman, Ondřej 768
Hernández, Esther 523
Hudeček, Lana 1054

I

Iaroshenko, Polina 944
Inshakova, Evgeniya 34

J

Jakubíček, Miloš 143, 1100
Janda, Laura A. 776
Johannsson, Ellert Thor 233, 530, 623
Jung, Hae-Yun 46

K

Kabatnik, Susanne 376
Kallas, Jelena 332
Kern, Boris 953
Kettnerová, Václava 966, 978
Kiil, Ahto 332
Kim, Agnes 537

Kolářová, Veronika 966, 978
Konstantinidou, Evanthia 676
Koppel, Kristina 177
Koppensteiner, Wolfgang 117
Kopylova, Mariia 284
Korecky-Kröll, Katharina 117
Kosem, Iztok 132
Kovář, Vojtěch 143
Kovařík, František 143
Kováříková, Dominika 776
Krawczyk, Iwona 568
Krek, Simon 132
Kroon, Martin 862
Krstev, Cvetana 295

L

Laluashvili, Tamar 162
Lameli, Alfred 103
Langemets, Margit 177
Lazursky, Alexandre 34
Lee, Jun 46
Lewandowska-Tomaszczyk, Barbara 152
Licurgo Pires, Silvio 405
Lindemann, David 425
Lorentzen, Henrik 482
Loukachevitch, Natalia 944
Lovrić Jović, Ivana 1054

M

M. Pintér, Tibor 547
Maask, Hanna 177
Makhatadze, Marine 547
Margalitadze, Tinatin 162
Marjanović, Saša 318
Marklová, Anna 259
Marković, Aleksandra 295
Marszałek, Jagoda 568
Martín-Cancela, Elena 1006
Mchedlishvili, Ketevan 162
Mederake, Nathalie 103
Meladze, Giorgi 162
Meyer, Peter 437
Middell, Gregor 758
Mihaljević, Ana 554
Mihaljević, Josip 554
Mika, Dorota 449
Mikkelsen, Agnes Aggergaard 310
Mirković-Birtašić, Dejana 318
Mírovský, Jiří 966, 978
Movsisyan, Diana 359

N

Nahod, Bruno 1114
Nam, Kilim 46
Namkung, Seol 46
Nešpore-Bērzkalne, Gunta 788
Nied Curcio, Martina 661
Nimb, Sanni 310
Norman, Nathalie Hau 310
Nouvel, Damien 390
Nowak, Krzysztof 449, 568

O

Ó Mianáin, Pádraig 1019
Okropiridze, Giorgi 162
Ölander Aadland, Emma Josefin 798
Olbrich, Michal 978
Ostroški Anić, Ana 1077

P

P. Márkus, Katalin 547, 648
Paulsen, Geda 332
Perak, Benedikt 1100
Petreková, Katarína 1100
Pődör, Dóra 547
Prangel, Esta 178
Prossliner, Linda 813

R

Raaf, Manuel 193
Rebosio, Laura 683
Rees, Geraint Paul 1032
Risberg, Lydia 177
Röhrer, Ines 193
Ruseishvili-Cartledge, Manana 547
Rysin, Andriy 463

S

Sahkai, Heete 332
Sanmarco Bande, María Teresa 16, 245
San Martín, Antonio 203
Sharifi, Saghar 584
Sigtryggsson, Jóhannes B. 530, 598
Sijens, Hindrik 1065
Skoumalová, Hana 829
Solanki, Deepak 1127
Sørensen, Nicolai Hartvig 310
Spagnol, Michael 607
Špica, Dragana 991
Srdanović, Irena 991
Stāde, Madara 788

Stanković, Ranka 295
Starko, Vasyl 463
Stefan, Nika 1065
Štěpánková, Barbora 932
Stöckle, Philipp 117
Storjohann, Petra 836
Sviķe, Silga 849

T

ten Hacken, Pius 471
Terzić, Dušica 318
Tiberius, Carole 862
Timoshenko, Svetlana 34
Torres del Rey, Jesús 217
Tóth, Ágoston 876, 888
Trap-Jensen, Lars 482
Trekker, Catherine 203
Tuta Dujmović, Marijana 1114
Tuulik, Maria 177
Two Dogs, Richard 902

U

Ulfarsdottir, Thordis 233, 530
Ullrich, Jan 902
Urbach, Nico 103

V

Vacalopoulou, Anna 1142
Vainik, Ene 332
Valcárcel Riveiro, Carlos 16, 245
van der Zwaag, Johan 1065
Vapper, Silver 177
Villa, Federico 3
Vokáčová, Martina 259

W

Wang, Zongyuan 347
Wills, Tarrin 623
Wu, Yingtian 1100

Y

Yamada, Shigeru 698
Yerznkyan, Yelena 359

Z

Žarković Eriksson, Marija 523, 636