GI_Forum 2018, Issue 2
Page: 50 - 57
Short Paper
Corresponding Author:
abenito@usal.es
DOI: 10.1553/giscience2018_02_s50

Creating Meaningful Narratives in Collections of Historical Lexical Data

Alejandro Benito[1], Antonio G. Losada[1], Roberto Therón[1], Amelie Dorn[2] and Eveline Wandl-Vogt[2]

[1]University of Salamanca, Spain
[2]Austrian Academy of Sciences, Vienna, Austria

Abstract

Historical dictionaries are colossal spatio-temporal artefacts comprising thousands of interrelated concepts and hosting a wide range of answers for cultural and/or historical inquiries. In our approach, the combination of thematic maps with network analysis and real-time textual querying of semantically-enriched lexicographical data serves as an entry point for the visual exploration of a large collection of records. Our tool aims to improve the comprehension of big data through visualization, thus helping the user to reach meaningful conclusions and acquire valuable insights into linguistic and other cultural issues in fast, easy ways.

Keywords:

data visualization, e-lexicography, exploratory analysis, user-centred design, dialectology

1 Introduction

Data visualization can be extremely helpful in research in the humanities, as it can accelerate workflows and the extraction of meaningful conclusions. It is also key to connecting intimately with the end-user, appealing to known and unknown factors of human psychology. Given the self-inquiring nature of the Digital Humanities (DH), this emotional involvement is not only desirable but often required in order to obtain quality results. Therefore, this calls for more prominent research into non-text-based approaches in the Humanities, as suggested by Champion (2015). This paper describes the current state of development of a novel visual exploration tool, which is the result of a two-year collaboration between partners in a multidisciplinary DH project, exploreAT! - Exploring Austria's Culture Through the Language Glass (Wandl-Vogt, Kieslinger, O'Connor & Therón, 2015). The project is based on a rich collection of words of the Bavarian dialects of Austria (Bairische Mundarten in Österreich) (Dorn, Wandl-Vogt, Bowers, Piringer & Seltmann, 2016) and relies heavily on data visualization to offer unique insights into the corpus for expert and non-expert users. The corpus analysed in this study is part of this collection and currently comprises more than 20,000 records.

This paper is structured as follows. In Section 2, we refer to related work by other authors. Section 3 introduces the problem at hand and links it to the challenges identified in Section 2. In Section 4, we describe our solution and the main lines of action taken so far. Finally, in Section 5 we discuss the impact of this work in the context of a broader body of research and outline outlooks and enhancements that we are currently planning to incorporate in our research.

2 Previous Work

In the context of our particular research, past projects have had a long and close relationship with geo-humanities and spatial thinking, which still continues today. For example, in the period 1924 to 1970, there was an ongoing effort to develop a dialect geography, which resulted in the generation of several paper-based maps and other geo-linguistic artefacts, some of which are currently being adapted to the web (Glauninger & Braun, 2017). The rise of web technologies allowed some authors to propose a system to navigate the data via maps (Wandl-Vogt, 2010; Wandl-Vogt et al., 2008), which was in line with important research at the time (Crampton, 2009). More recently, different models have been proposed to capture the intrinsic spatio-temporal nature of the data (Scholz, Hrastnig & Wandl-Vogt, 2017; Scholz, Lampoltshammer, Bartelme & Wandl-Vogt, 2016); in turn, other tools have been created to allow not only the display but also the critical analysis of the dataset by means of data visualization and web cartography techniques (Therón & Wandl-Vogt, 2014).

Within the context of *exploreAT!*, we developed a series of prototypes following a user-centered design approach (Benito & Goikhman, 2017). This paper elaborates in particular on the feedback and the problems revealed during the testing phases of the first two prototypes (Benito, Dorn, Therón, Wandl-Vogt & Losada, 2018; Benito et al., 2016), and lays the first stone for more complex prototypes to follow. The initial prototype[1] introduced a spatiotemporal visual analysis tool that enabled a headword-based exploration of the dataset. This tool displayed the spatial and temporal distribution of textual queries that were run by a search engine. Considerable work was put into creating network visualizations representing relationships between lemmas found in the headwords.

3 Problem description

The user-centered design process employed in the development of the prototypes helped to reveal issues that were used as core input for a further iteration, resulting in the new prototype we introduce in this paper. The exploration task could not be started from the network graph and required some sort of prefiltering, given the high number of nodes and edges present in the data (close to a million). Not only this, but users did not have the possibility of generating graphs for areas with more than 4,000 entries (see Figure 1). Although this prototype was good enough as a first approach to the problem, it failed to reveal important cultural insights, except in some specific situations. Moreover, it failed to

[1] https://github.com/acdh-oeaw/exploreAT-collectionexplorer

promote critical cartography 'effectively' (Gordon, Elwood & Mitchell, 2016) and was not able to 'tell a clear story' of the culture of the time, as recommended by experts from different fields (Venturini, Bounegru, Jacomy & Gray, 2015).

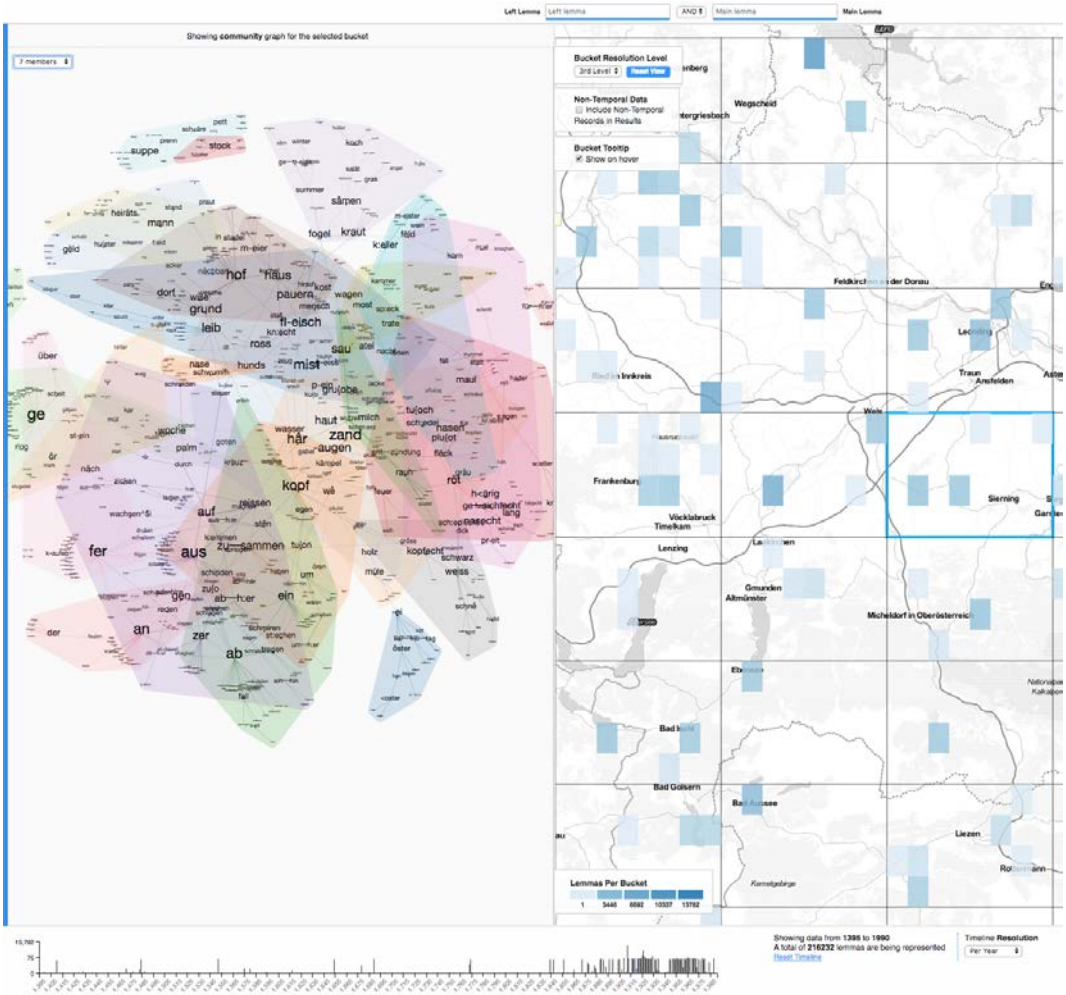

Figure 1: The first exploration tool displaying a large network. Although the graph was partitioned employing a community-detection algorithm in order to produce a more meaningful representation, the analyses of these highly-populated structures obtained very low success rates in our tests

One flaw was in the use of the geohash as a partition of space for aggregating results. Although the geohash was initially considered computationally optimal and allowed a rapid spatial search and representation of the data, it also concealed many interesting features (e.g. position, shape and hierarchy of historical regions) that need to be taken into consideration in the context of our problem. One particularly important issue we found in the collection-explorer was related to the use of headwords and lemmas: their significance in a cultural

study of lexical data, although straightforward and easy to grasp, was found to be very marginal. The discovery of these problems motivated us to semantically enrich the data and translate it to a newer format (TEI), which allowed a more systematic ingestion of the information (Bowers & Stöckle, 2018). This conversion permitted a subset of the data to be explored in a semasiological (form-based) and, to a lesser degree, in an onomasiological (concept-based) manner, and it served as primary input for our research. Currently the concepts are formed from the distinct words that conform to the different questions in the corpus, although we are aiming to apply topic-modelling and other NLP algorithms when further advances are achieved in this area. Given the dialectal complexity and variety of Austria, one of the richest in the world, this task is not trivial. Most of the existing algorithms (e.g. stemming) fail when the input is not given in modern standard German, which compels us to re-implement such algorithms on a per-dialect basis.

Using the first results of this conversion work, we developed the second prototype mentioned at the beginning of this section, named *Concept Lights*[2]. The tool offered interactive exploration via questions based on the display of concept co-occurrence in heatmap views. Since at the time the prototype was created the questionnaire data had not been linked to the sources, this second prototype did not present a spatial or temporal view, as had been the case in the previous prototype. However, we were able to apply distant-reading techniques that allowed the team to realize the major importance of employing semantically-enriched data in our visualizations. These visualizations proved to be much more successful in the study of folklore and traditions than the simpler headword-based exploration. As a consequence, the co-occurrence of concepts in different questionnaires is now key in our new third prototype.

4 The Tool

The analysis of the problems outlined above brought to light a series of questions that we have tried to address in this paper: How can we offer scholars the bigger picture of the problem without presenting them with an unfathomable network visualization with thousands of nodes and edges? How can we keep the subtleties found in the data without introducing unnecessary noise in the research process? How can we design responsive user-interfaces that allow the exploration of large lexicographical data without compromising efficiency? In order to answer these and other questions, our proposal looked at work by authors in the fields of cartography (Harley, 1989, 2009), visual storytelling (Venturini et al., 2015), and data visualization (Bostock & Davies, 2013; MacEachren & Kraak, 1997).

As a result, we propose an interactive dashboard[3], which can be seen in Figure 2. It presents a questionnaire-based semasiological and onomasiological exploration tool with three linked views. As in previous prototypes, the search area (on the left) allows users to run queries resulting from the combination of three fields: headwords, concepts (meaning in standard German assigned by an expert at the time of the record's creation), and questionnaire

[2] https://github.com/acdh-oeaw/exploreAT-conceptlights
[3] https://github.com/acdh-oeaw/exploreAT-concepttopography

number. The bubble graph of the questionnaires also shows the distribution of the current result set and is connected to the map and graph (explained below). At top-right, a thematic map is presented; at bottom-right, a network visualization shows concepts shared between questionnaires. The dashboard sits on three main pillars that combine approaches taken from the two earlier prototypes.

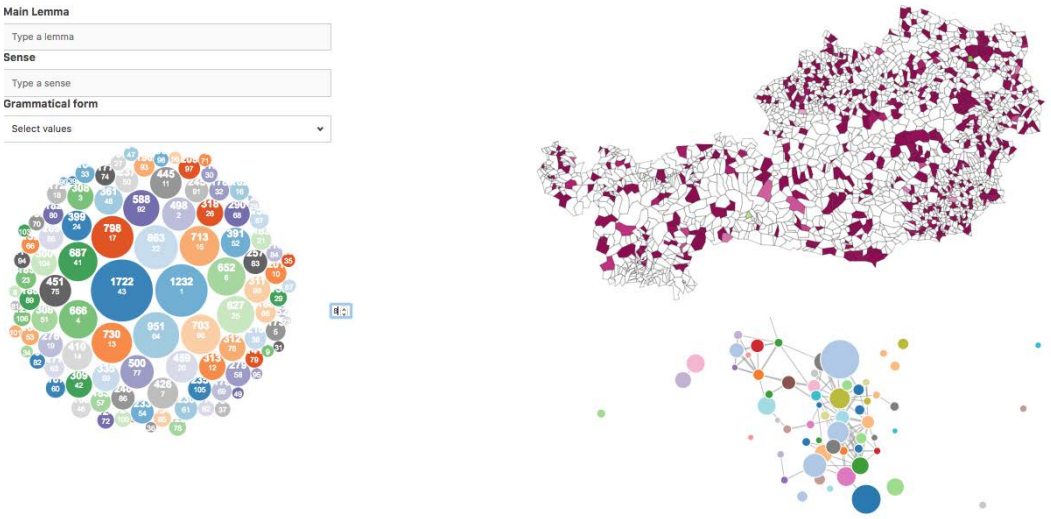

Figure 2: Proposed linked-view exploration dashboard showing the three main components of the prototype

Revised Spatial Approach

The first step in creating a cohesive visual narrative is related to the preparation of tailored thematic maps that are able to faithfully display present and past work on the dataset. One of the main issues with creating the maps is that the geocoding of historical toponyms is less straightforward than it is for current place names. We depart from work by other authors (Scholz et al., 2016) who during their research prepared a spatial MYSQL database from which we extracted the geocoding information. The extracted spatial shapes are presented in Figure 3. Employing common tools like geopandas, matplotlib, gdal and other command-line tools, we were able to produce a TopoJSON file suitable for data visualization in the browser using the d3 data visualization library and related best practices (Andrienko et al., 2010; Bostock, 2016; Bostock & Davies, 2013; MacEachren & Kraak, 1997). The resulting map from the TopoJSON file is shown at the top-right of Figure 2. Results coming in from textual queries in the search engine are also dynamically projected into the map according to the chosen resolution.

Smarter Network Analysis

Working with concepts instead of single-meaning words allows different sources to be captured according to the idea they refer to. It also allows the user to see where they

originated. As a consequence of the conversion process the data underwent (as explained in Section 2), we are now able to produce smaller but more meaningful graphs that respond synchronously to user-generated textual queries – that is, an immediate visual response is seen in the other elements of the dashboard (see bottom of Figure 2). Instead of producing large networks, the current prototype generates semantically-richer networks based on concept co-occurrence between questionnaires, a feature that fits into the project's final aims. We also enabled the ability to fine-tune the graph by the number of common concepts shared among nodes (questionnaires): this is common practice in these kinds of visualizations and allows the end-user to focus on the relevant portions of the graph, reducing the cognitive load involved in analysing the graph as a whole. However, in future developments exploring bigger networks, this approach will have to be verified, and other prototypes exist that are able to trigger a data-refinement cycle that is more closely aligned with the aims of our research. For example, a web-enabled semantic version of the data which will expose richer graphs to explore in the visualizations is currently being prepared (Abgaz, Dorn, Piringer & Wandl-Vogt, 2018) and will be integrated into our tool in the near future.

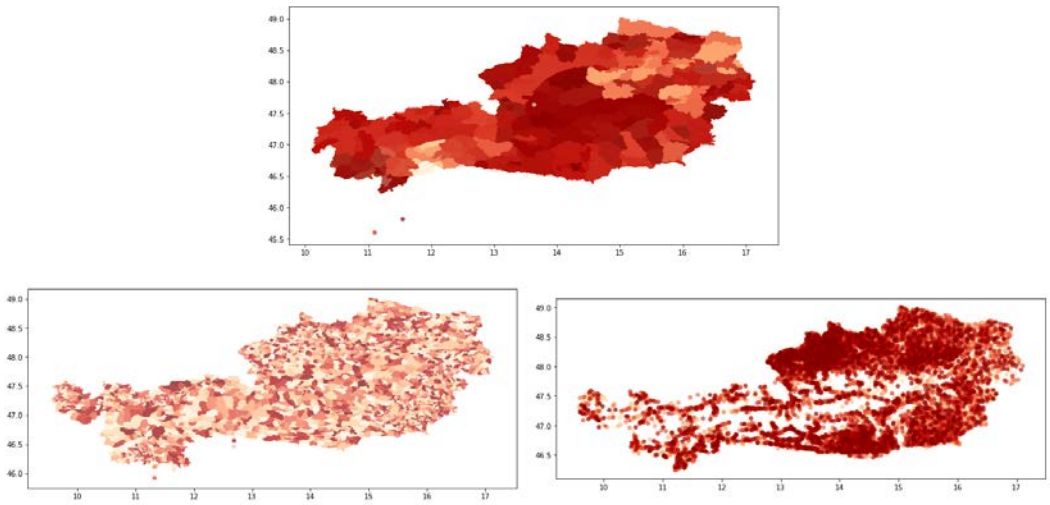

Figure 3: The administrative regions we extracted from MySQL presented using geopandas and matplotlib. Top: Regions – 544 shapes. Bottom left: Communities – 2,587 shapes. Bottom right: Settlements – 18,059 shapes

5 Discussion

The idea of narrative must be kept in all the stages of the software life-cycle in order to produce adequate results. Despite the fact that the prototype is still reduced in functionality for the end-user, it is based on solid foundations that in the future will allow us to bring more data into the analysis from different sources. As our iterative user-centered design process evolves, we are seeing how we are able to produce visualizations that not merely

report on the complexity of the problem at hand but also reveal themselves as useful tools for the lay user. This process is tedious and not short of difficulties: the conceptual distance between the Humanities and Computing is still great. By concentrating our efforts on visualizations like the one presented here, less based on bare computational power and more oriented towards capturing expert knowledge in maps and other visual artefacts, we believe we will be able to reduce the gap further in the future.

References

Abgaz, Y., Dorn, A., Piringer, B., & Wandl-Vogt, E. (2018). A semantic Model for Traditional Data Collection Questionnaires enabling Cultural Analysis. *6th Workshop on Linked Data in Linguistics LDL-2018: Towards Linguistic Data Science at LREC 2018*.

Andrienko, G., Andrienko, N., Demsar, U., Dransch, D., Dykes, J., Fabrikant, S. I., … Tominski, C. (2010). Space, time and visual analytics. *International Journal of Geographical Information Science*, *24*(10), 1577–1600.

Benito, A., Dorn, A., Therón, R., Wandl-Vogt, E., & Losada, A. G. (2018). Shedding Light on Indigenous Knowledge Concepts and World Perception through Visual Analysis. *Digital Humanities 2018*. Retrieved from https://dh2018.adho.org/

Benito, A., & Goikhman, A. (2017). Exposing Cultural Heritage through Computer Screens: The Role of User-centered Design in the DH. *WG4: Lexicography and Lexicology from a Pan-European Perspective. The final conference of the ENeL COST Action. European Network of e-Lexicography*. https://doi.org/10.13140/RG.2.2.13207.27040

Benito, A., Losada, A. G., Therón, R., Dorn, A., Seltmann, M., & Wandl-Vogt, E. (2016). A Spatio-temporal Visual Analysis Tool for Historical Dictionaries. *Proceedings of the Fourth International Conference on Technological Ecosystems for Enhancing Multiculturality* (pp. 985–990). ACM.

Bostock, M. (2016). Command-Line Cartography, Part 1. Retrieved 30 May 2018, from https://medium.com/@mbostock/command-line-cartography-part-1-897aa8f8ca2c

Bostock, M., & Davies, J. (2013). Code as cartography. *The Cartographic Journal*, *50*(2), 129–135.

Bowers, J., & Stöckle, P. (2018). TEI and Bavarian dialect resources in Austria: updates from the DBÖ and WBÖ. *Proceedings of the Second Workshop on Corpus-Based Research in the Humanities*.

Champion, E. (2015). Seeing is Revealing: A Critical Discussion on Visualisation and the Digital Humanities, *Proceedings of the Digital Humanities Conference 2015*.

Crampton, J. W. (2009). Cartography: maps 2.0. *Progress in Human Geography*, *33*(1), 91–100.

Dorn, A., Wandl-Vogt, E., Bowers, J., Piringer, B., & Seltmann, M. (2016). exploreAT! – Perspectives of exploring a dialect language resource in a framework of European digital infrastructures. *1st International Congress on Sociolinguistics*. Retrieved from http://ics1.elte.hu/

Glauninger, M., & Braun, J. D. (2017). Austrian Dialect Cartography 1924-1956. Retrieved 30 May 2018, from https://www.oeaw.ac.at/acdh/projects/austrian-dialect-cartography/

Gordon, E., Elwood, S., & Mitchell, K. (2016). Critical spatial learning: participatory mapping, spatial histories, and youth civic engagement. *Children's Geographies*, *14*(5), 558–572. https://doi.org/10.1080/14733285.2015.1136736

Harley, J. B. (1989). Deconstructing the map. *Cartographica*, *26*(2), 1–20.

Harley, J. B. (2009). Maps, knowledge, and power. *Geographic Thought: A Praxis Perspective*, 129–148.

MacEachren, A. M., & Kraak, M.-J. (1997). *Exploratory cartographic visualization: advancing the agenda*. Elsevier.

Scholz, J., Hrastnig, E. & Wandl-Vogt, E. 2017. A Spatio-Temporal Linked Data Concept for Modeling Spatio-temporal Dialectal Data. *Lecture at SPHINX-Workshop, a pre-conference workshop at COSIT 2017*.

Scholz, J., Lampoltshammer, T. J., Bartelme, N., & Wandl-Vogt, E. (2016). Spatial-temporal modeling of linguistic regions and processes with combined indeterminate and crisp boundaries. *Progress in Cartography* (pp. 133–151). Springer.

Therón, R., & Wandl-Vogt, E. (2014). The fun of exploration: How to access a non-standard language corpus visually. *Vis-LR: Visualization as Added Value in the Development, Use and Evaluation of Language Resources. Workshop at LREC2014, Ninth International Conference on Language Resources and Evaluation*, 1239–1245.

Venturini, T., Bounegru, L., Jacomy, M., & Gray, J. (2015). How to Tell Stories with Networks: Exploring the Narrative Affordances of Graphs with the Iliad.

Wandl-Vogt, E. (2010). Datenbank der bairischen Mundarten in Österreich electronically mapped (dbo@ema) / Database of Bavarian Dialects in Austria electronically mapped (dbo@ema). Interactive, georeferenced resources for the Dictionary of Bavarian dialects in Austria (WBÖ, DBÖ). Retrieved 30 May 2018, from https://wboe.oeaw.ac.at/projekt/beschreibung/

Wandl-Vogt, E., Bartelme, N., Fliedl, G., Hassler, M., Kop, C., Mayr, H., … Vöhringer, J. (2008). dbo@ema. A system for archiving, handling and mapping heterogeneous dialect data for dialect dictionaries. *Proceedings of the XIII Euralex International Congress, Barcelona, Universitat Pompeu Fabra.* Documenta Universitaria.

Wandl-Vogt, E., Kieslinger, B., O'Connor, A., & Therón, R. (2015). exploreAT! - Perspektiven einer Transformation am Beispiel eines lexikographischen Jahrhundertprojekts. In *DHd (Digital Humanities Im Deutschsprachigen Raum) 2015*. Retrieved from https://dhd2015.uni-graz.at/