Orthographic Transcription Systems for Dialects – A Case Study on Viennese Dialect

Friedrich Neubarth

Abstract. For this paper, I am not in the position to present original work – it is about an orthographic transcription system for Viennese dialect that was developed by Sylvia Moosmüller in the course of a project that aimed at machine translation from Standard German into this dialect. Rather, I want to recapitulate the making of this special-purpose orthography and discuss a few issues that automatically come up with such an enterprise. Two sources of information will be of special concern: H.C. Artmann's collection of poems, *med ana schwoazzn dintn* and Maria Hornung's *Wörterbuch der Wiener Mundart*. Both of them provide highly consistent ways of transliterating a dialect that is primarily spoken, but in very different ways – for different purposes. The orthography described here has yet another purpose: machine translation, hence language technology. Reviewing many of the questions that came up during the numerous discussions we had in that project may well be interesting to readers who face a similar situation. The conclusion may seem somewhat disappointing, but should be read as an encouragement: much work has already been done, but there is no end to it. Each time we face a new target (and language technology is an ever emerging, multiple target), we have to rethink our ways of doing or encoding things. Meanwhile, it is also inevitable that we must continue rethinking our linguistic knowledge base.

INTRODUCTION

When designing orthographies for particular dialects, it is worthwhile to contemplate a few issues before beginning such an enterprise. Some of these will be of rather general nature, others more specific to the actual use of such a writing system. The case I will present and discuss here is an orthography for Viennese dialect (VD), that mainly targets machine translation as a concrete application (cf. Haddow 2013, Neubarth et al. 2013, Neubarth & Trost 2017). There are many different purposes for such writing systems: literature, song-texts, poetry, lexica etc., but especially for applications in language technology, which strategies to adopt depends on the target application. Writing language with a fixed set of symbols (letters) is always a sophisticated compromise that has occupied scholars and practitioners since the outset of the Phoenician alphabet, the first one in history that provides symbols for individual sounds.

In the case at hand, there was a need to provide an orthography that would be easy to work with, both for humans and machines – meaning

that it should not be too hard to decipher by humans reading and editing text (machines don't care much about that), and hence does not depart too far from the standard orthography (if possible), but on the other hand should not pose too much burden on machine readability – meaning it should adopt a set of characters with as few diacritic symbols as possible and no other graphic means. On the other hand, such an orthographic coding should represent as much as possible the phonetic properties of the dialect, thus giving compatibility with or resemblance to the orthography of the standard variety (Standard German – SG henceforth) a lower priority.

What are these issues? We can formulate them as three principal questions: 1) What exactly is the dialect we are working with? 2) What set of characters should we use and how much does that enable us to represent the phonological properties of that dialect? And 3) in which ways does the orthographic encoding of the dialect relate to the orthography of the standard variety?

First it is important to note that dialects and writing systems seem to be intrinsically incommensurable. (Not meaning that there cannot or should not be writing systems for dialects.) But in essence, the effect of an orthography is fostering normalization, while the defining property of a dialect is that it is a language for a confined group of people sharing it, and that there are many other dialects that differ from that particular dialect (while there may be shared commonalities to be exploited by extended groups). This issue of differentiation (in language) seems to be at the core of the development of human languages (and a fortiori, of dialects). Any linguist working on dialects might object that this is an oversimplification – I would fully agree – but the multitude of ways to speak is still a fact.

On the other hand, the writing of many languages (e.g., High German) was not standardized for a long time; attempts at standardization rather arose when it became societally and politically opportune (or necessary) to pursue such an enterprise. Accepting this thought, the free use of orthography in social media seems to support the idea that writing systems are not only taken to be a means for standardization. To put more weight on this: there are many so-called non-standard languages: dealing with these has developed into a branch of language technology of its own.

Given that there are so many dialects and none of them has its own committed standard way of writing it, how should we delineate any of these language varieties in order to gain a background against which we can start to develop an orthography? The answer is as easy as it is problematic: one has to decide. One has to develop an image of the particular

dialect that may have regional or social defining properties. And we have to keep in mind that this image will always be an abstraction.

It was Sylvia Moosmüller, drawing from her longstanding experience of working with dialects in many manifestations, who took on the burden of deciding upon a Viennese dialect as a working standard. A standard which in reality may perhaps never exist, but represents an abstraction from very closely related varieties that can be perceived and accepted by a majority of people who may still have their own image of a Viennese dialect. Someone needed to decide, and Sylvia was in the position to do so, having internalized the manifold properties of this idiom – a term that may even better reflect what we were trying to grasp here.

That seems to be a strange response to the question, which dialect are we working with. However, the term Viennese by itself reveals the imprecision to be expected under a dialectological perspective. Vienna has 23 districts and as many social classes as one wishes to define, and this is not the end of it – migrant groups have contributed to the idioms of the city for hundreds of years, most strongly since the end of the 19th century. Youth language is always a special issue that I here just want to mention but not comment on. It has to be emphasised that when creating a machine translation (or a speech synthesis) system for Viennese dialect, one automatically strives for a prototypical variety rather than trying to oversee the abundant variation.

The second issue revolves around the question, what goals should the design of an orthography for a particular dialect follow? There are three sub-issues: how many characters should be added to our alphabet (in German, the Umlaut characters ä, ö, ü and ß go beyond the set of basic ascii characters)? The answer: as few as possible, but if 'necessary', yes. How close should the writing system be to the phonetic properties of the dialect? This is an ill-posed question because writing systems never directly reflect what we do with our mouth and what enters our ears. They rather attempt to reflect systematic distinctions between sounds, and thus relate much more to the phonology of a particular language (variety). Moreover, it may not only be the representation of sounds, but also perhaps morphological issues that play a role in orthography, and certainly, orthographies are always conservative towards earlier stages of a language (English and French are good examples).

Phonology targets the systematic setup of sounds in a given language and certain processes that – also systematically – may alter these sounds. In classical, structuralist terms, two or more phonetic realisations of a speech sound may relate to the same phoneme – that stands for a qua-

si-symbolic, abstract sound concept of a given language. A well-known example is the two realisations of 'back' fricatives in German: *ich*- [ç] versus *ach*-Laut [x], that relate to the same phoneme (in phonetic terms: palatal versus velar, whatever symbol we take for it), but the occurrence of which is triggered by phonological context.

However, there are intricacies beyond the mere identification of sounds. In many German varieties there are length contrasts that may (or may not) correlate with melodic contrasts (different sounds). This 'or may not' statement shows the real challenge: length contrasts are not always coupled with melodic contrasts. Some of them are, some are not, and it always has to be questioned if we wrongly infer such a contrast because we know that it exists in other varieties, or if it is manifest distinction. Length and melody interact with each other in a sophisticated way, generating differences between dialectal varieties that will be heard by an experienced ear but that are still hard to grasp phonologically, even by an experienced linguistic mind. The write-as-you-hear strategy may be an initial attempt to get on with the task, but drawing from just a little fieldwork experience, I can bet that there will be a moment where one hears the same sound in two ways (or two similar sounds as the same). There is no ad-hoc phonetic solution, and even if I lost my best, the orthography would be just a mere phonetic transcription, missing certain necessary features that make it systemic and are necessary to qualify it as an orthography.

The core problem is still that we have an alphabet that for more than three millennia has served the purpose of coding how we actually speak – in contrast to symbolic writing systems such as Chinese characters where each character represents a morphological unit with its own meaning and a phonetic realisation that is intrinsically not defined. How should we adapt this system in order to reflect the way of speaking in a particular language or dialect? Writing systems have been developed, and orthographies have changed and evolved according to these needs. In electronic communication the ASCII standard, comprising the English alphabet, still imposes a strong bias, despite the fact that most languages other than English employ extended character sets (as mentioned before, in German there are 4 more characters, still found in the extended ASCII set). While the character sets vary from language to language, each language imposes its own conventions that themselves produce a bias against non-standard varieties.

This brings us to the third issue: how should we relate an orthography for a dialect to the conventions of the standard variety? Obviously we

need to rely on the conventions of the standard to a certain extent. Adherence to the standard conventions surely facilitates intelligibility and readability for a broader audience. On the other hand, such a trait may hinder the identification of the dialect as an independent language variety in its own right. So, designing the orthography towards the phonological properties of that variety has high merits, though it comes with the risk of declined readability.

All that follows basically reflects the work of Sylvia Moosmüller, as a project partner of the project MLT4MLV. The project had as its goal the development of a machine translation system between Viennese dialect and Standard German. It was up to Sylvia to come up with a first proposal for an orthography suitable to our needs. Needless to say that almost all of her suggestions directly entered the final version of encoding – after months of discussions that mostly served the purpose of making us all understand what considerations led her to make each of the many decisions. For an extensive overview, see Hildenbrandt et al. (2013).

VIENNESE ORTHOGRAPHIES: 2 EXAMPLES

There are numerous sources for written Viennese dialect (e.g., Schuster 1956, Schikola 1954), each of them following its own conventions, more or less consistently. I know of two examples that are consistent to an extent that is absolutely impressive, and which represent opposite positions in how they realize their goals. It is worthwhile to discuss both of them – they are ideal examples that give an introduction to particular problems of creating (yet another) writing system for Viennese dialect. The first one is H.C. Artmann's collection of poems in Viennese dialect, *med ana schwoazzn dintn*, and the other is Maria Hornung's Lexicon of Viennese, *Wörterbuch der Wiener Mundart*.

H.C. ARTMANN'S *MED ANA SCHWOAZZN DINTN*

Artmann's approach was to represent most phonological distinctions in as simple a form as possible, exploiting the commonly used writing system in an often quite surprising way and making text written by him in Viennese look like the transcription of some language spoken far away, perhaps in Africa. It was an artistic goal to make it look highly different from the standard language, but it was also an artistic enterprise to be as accurate as possible with the means (available characters on the typewriter) at hand.

Characters in use: Artmann's orthography uses the letters <a-z> (except for <v> and <y>), as well as the *Umlaut* characters <ä>, <ö> and <ü>. The letter <v> can be subsumed by <f> and <w>, whereas <y> occurs only in loanwords in SG and has no function in VD. The letter <c> occurs only in the two multi-character graphemes <ch> for velar/palatal fricatives [x]/[ç] and <sch> for the post-alveolar fricative [ʃ]. The letter <q> (without a following <u>) represents a combination of <gw>, <x> a combination of <ks>, and <z> basically stands for <ds>; when doubled it represents the geminated affricate that can be represented by <ts> as well (as in the title of the collection: *schwoazzn* – SG *schwarzen* 'black-Fem.Dat.Sg').

Let us review a few lines from one of his poems. It is titled *liad* ('song') and its first 4 lines are: *a bak / one bam / one gros / one wossa* ('a park / without trees / without grass / without water'). From this stanza alone, we can discuss several issues:

'a' and 'o': generally, in Bavarian dialects, where /a/ surfaces in SG, in VD we find a rounded, open mid-vowel [ɔ] – with many exceptions, one of which is found in the second word of the poem – *bak* (SG *Park* 'park'), where rounding of the /a/ does not take place (loanwords most often don't have it) and the r-vocalisation into [a] is void in this case, since it targets a sound identical to the preceding vowel. So <a> primarily stands for /a/, secondarily it marks the result of r-vocalisation when following a vowel other than /a/ (e.g., *deamometa* – SG *Thermometer* 'thermometer' – see below). Where in SG /o/ surfaces, we find a rounded, closed mid-vowel [o], e.g., *brod* (SG *Brot* 'bread'). So, we have a three-way contrast, whereas the alphabet provides two graphemes. In SG that contrast is solved between /a/ which is unambiguous and <o> for [o] vs. [ɔ] which is phonologically determined by a length contrast (oversimplifying), but both of them pertain to /o/ sounds. In VD, all /o/ sounds seem to be realized as [o], whereas /a/ sounds have the two variants [a] and [ɔ]. Artmann deliberately ignores this intricacy and uses <a> for [a] and <o> for [o] and [ɔ] – as in SG, phonologically adequate, but this decision creates ambiguities. (E.g., in his transcription it would be *brodwiaschtl med an brod* – SG *Bratwürstel mit einem Brot* 'fried sausage with a piece of bread'). In the text presented we find those transliterations of the words *gros, wossa* – SG *Gras, Wasser* 'grass, water'.

Nasal 'a': a special case is /a/ before nasals. Phonetically it will be realized with a nasal quality but also with an [u] quality. Sometimes the triggering nasal will be clearly realized, sometimes it can even be dropped. Artmann tries his best to reflect this instability of nasals. He uses the combination <au> to mark nasal /a/ throughout and only writes <n> when

the chances are high that it will be heard. E.g., *waun* – SG *wann* ‘when’ vs. *auschaud* – SG *anschaut* ‘look at’, where the second occurrence of <au> represents the regular diphthong /au/ that is phonetically realised as a monophthong in VD – [ɒː].

R-Vocalisation: the sound /r/ is stable in onset position, however, within the same phonological domain it may colour the preceding vowel (if there is one and if it is not in onset position, it generally is not realized as a consonant – hence the term vocalisation. Our initial example serves again: *bak*, relating to SG *Park*, is written without ‘r’ because /r/ is not pronounced in any way. Another line from our poem shows this effect: *nua r a blaus deamometa* – SG: *nur ein blaues Thermometer* ‘just a blue thermometer’. The ‘r’ in ‘Thermometer’ is pronounced and transcribed as ‘a’. Notice that despite the fact that /r/ is most often vocalized in VD, it may also be used to separate two adjacent vowels, as shown in the given example. Sometimes this phonological process is called ‘intrusive-r’, e.g., *nua r a blaus deamometa* – SG *nur ein blaues Thermometer* ‘just a blue thermometer’. What is striking is that the /r/ seems to adopt both roles – it lends its melody towards disappearance and takes up a special function to mark onsets in order to separate two (otherwise) adjacent vowels.

Diphthongs: what corresponds to a diphthong in SG is mostly realized as a monophthong in VD. Artmann’s transcription ignores this fact, employing a similar spelling as in SD. /au/ may be realized as [ɒː] in VD, but he pertains to write it as ‘au’ as in *blaus deamometa*. /ai/ is realized as [æː], but still written as ‘ei’. Some SG /ae/ diphthongs seem to be subject to the secondary Umlaut effect and are realized as [aː]. E.g., *drei* [dræː] – SG *drei* ‘three’ vs. *zwa* [dsvaː] – SG *zwei* ‘two’. Artmann uses just one letter <a> in such cases – including the indefinite article, but it is manifest twice in *a bak one bam* – the last word here corresponds to SG *Bäume*. There is a stem alternation between singular and plural, the latter having an Umlaut in its stem vowel. In Bavarian dialects, /a/-based stems (including diphthongs), that have rounding, hence being realized as either [ɔ] for /a/ or [ɒː] for /au/, cannot transfer the base vowel into a fronted vowel (such as /u/ to /ü/), but simply un-round them, resulting in a secondary [a]. For this reason, this phenomenon is also called secondary Umlaut. SG <eu/äu> corresponding to the diphthong /oe/ is generally unrounded in VD (e.g., *neich* – SG *neu* ‘new’) But in the context of /l/ (which, as a liquid vocalizes similar to /r/), we find secondary rounding of front vowels and diphthongs: the rounded alternative to [æː] is [œː] and graphemically represented as <äu> – the only occurrence of the character ‘ä’, e.g.: *wäu* – SG *weil* ‘because’.

Long vowels: It is an open issue whether we have length contrasts in vowels in VD. While there seem to be rather clear effects of isochrony in Upper- and Lower-Austrian dialects (long vowel – simplex consonant, short vowel – geminate consonant), e.g. [roːg] – SG Rock 'skirt' (sg.) vs. [regː] – SG Röcke 'skirts' (pl.), the situation is much less clear in VD (see Kühnhammer 2004 for a discussion of this example). Instead of isochrony effects we experience more a general length of vowels, and the burden of contrast seems to lie on the shoulders of consonants (see below). Artmann generally avoids marking long vowels; however, vowels in open, stressed syllables are sometimes reduplicated: *nii* – SG *nie* 'never', *schnee* – SG *Schnee* 'snow', *aa* – SG *auch* 'also', but even sometimes *zwaa* – SG *zwei* 'two' (where we have cited the same form with only one 'a' before).

Plosives – lenis / fortis: we are accustomed to a two-way distinction of plosives in many languages, and the terminology used to distinguish them by the pair lenis/fortis is at least neutral about what makes the difference. The phonetics of such contrasts reveals that at least two features may play a role here: aspiration and voicedness. This sounds a bit suspicious, and indeed there are languages that employ both of these features in phonological contrasts, giving rise not only to a two-way fortis-lenis contrast, but to a four-way distinction. One example is Hindi, with a neutral, an aspirated, a voiced and an aspirated voice variant of a plosive that is associated with the same place of articulation (i.e. 'p', 'ph', 'b', 'bh'). In VD syllable onset positions, lenis/fortis contrasts of that sort are clearly neutralized (therefore *bak* instead of *pak*, but there is one exception to be discussed below); in other positions, there are apparent lenis/fortis distinctions: intervocalically, after sonorants, but also due to concatenations we may experience different sounds, which, however, may reflect length rather than a melodic lenis/fortis contrast (or even both). Examples: *i red* – SG *ich rede* 'I speak', vs . *ea ret* – SG *er redet* 'he speaks'. The stem *red* has only a lenis/simplex /d/ that surfaces in 1P.Sg form, while the 3P.Sg form has a suffix *–d* that merges with the stem to show up as a fortis/geminate /t/. One could write geminate 'dd' in that case, but then other cases, where fortis consonants show up as fortis/geminates without any reason to assume gemination would become illogical – one of them is our very first example: *bak*. It is a fortis variant of the bilabial plosive, and there is no linguistic reason to assume it should be a geminate other than that its phonetic interpretation is identical to a configuration where we can identify a geminate structure upon morphological grounds. This may give rise to further theoretical discussions; on practical grounds matters

seem quite clear: use b/d/g for lenis/simplex plosives and p/t/k for fortis/geminate ones, and disregard the melody/length intricacies.

Geminates: have already been discussed before in the context of plosives. Other consonants that do not have a fortis/lenis pair reflected in the alphabet need to be encoded by doubled letters. There is not much more to say about this – since the strategy is taken over from standard orthography, but Artmann goes a step further and assigns lenis and fortis affricates single and double occurrences of letters: whatever is pronounced [ds] will be encoded as 'z'; as a fortis/geminate [ts], it will logically be encoded as 'zz' (*med ana schwoazzn dintn*). Nevertheless, while alveolar /s/ and labio-dental /f/ fricatives seem to have a clear geminate structure, perceivable from their phonetic interpretation, nasals only sometimes do so, and liquids are perceived generally as simplex. Artmann's intuitions, reflected in his way of writing, seem to be very plausible here.

HORNUNG'S *WÖRTERBUCH DER WIENER MUNDART*

A lexicon clearly serves different purposes than using a dialect in poetry. When reading Maria Hornung's lexicon, the transcription looks as unfamiliar to SG than Artmann's texts, but its appearance reminds less of the transcription of an African language and gives more the impression of a scientifically motivated coding. Indeed, Hornung attempts to represent each and every phonetic differentiation, which makes the use of diacritics and other typographic means unavoidable. I will not try to give a comprehensive overview of her orthographic solution here, but rather sketch the important differences to Artmann's solution.

'a' and 'o': the variants of SG /a/ sounds are represented by three graphemes: <a> for [a], <ǫ> for the rounded variants [ɔ], and <å> for the nasal variant. This leaves <o> for the counterparts of /o/ in SG, realized as [o].

Liquid vocalisation: as mentioned before, the effect of vocalisation of /r/ and /l/ may trigger the deletion of the source consonant, meaning that the sound will not be perceivable phonetically. As a matter of fact, this is not entirely true – the (non-)realisation of the liquid may be subject to variation. In order to indicate this, Hornung uses superscript letters, a decision which is linguistically accurate, but which would be very problematic to import for an orthography for written texts, e.g., ǫschbęa^{r}*n* – SG *absperren* 'to lock'; ǫsåmme^{l}*n* – SG *absammeln* 'to collect'. Likewise, nasals and plosives that may or may not be pronounced are also set in

superscript, thus representing all kinds of gradually applied processes of phonetic reduction.

Geminates: while Artmann uses single letters corresponding to fortis consonants <p, t, k>, Hornung always indicates gemination by reduplicating the relevant graphemes. Such that in intervocalic contexts or in coda position we may find <pp, tt, ck>, whereas fortis consonants after sonorants will still be represented by simplex graphemes. However, her strategy also targets the multi-character graphemes <ch> and <sch>. There are very rare cases of ambiguity, but by reduplication of these character sequences, one can distinguish between *zechn* – SG *Zehe* 'toe' and *zechchn* – SG *zechen* 'to boose', or the classical example for isochrony *fisch* – SG *Fisch* 'fish' (sg.) and *fischsch* (pl.).

Affricates: Hornung refrains from using the letters <x> and <z>, as well as <q> in order to subsume complex graphemic strings – hence these letters are not used.

Accents: special accents are marked in Hornung's lexicon, e.g., *dahínta* – SG *dahinter* 'behind'. This is a gratifying feature for a lexicon, but it would not be an ideal one for an orthography.

Front mid vowels: both unrounded /e/ and rounded /ö/ have a tense/lax (closed/open) variant. In SG, this distinction is clearly correlated to length, while in VD, but in Middle-Bavarian dialects in general, the tense/closed variant can be assumed to be basic. Nevertheless, there are two complicating factors: for /e/ preceding vocalising /r/ the lax/open variant will be found throughout, whereas /ö/ only arises with vocalising /l/, and one might want to pose the question how length is calculated there. Many examples are quite clear; others may be subject to free variation (e.g., [øtan] vs. [œtan] – SG *Eltern* 'parents'. And finally, the influence of SG may exact a confusing force upon this intrinsically unstable distinction. Hornung marks the lax/open variants with an ogonek diacritic throughout, again, this is a trait more appropriate for a lexicon rather than for a writing system.

AN ORTHOGRAPHY FOR MT: 1 EXAMPLE

When Sylvia presented her proposal for an orthography of Viennese dialect, in her most unspectacular way almost a decade ago, she explicitly mentioned that she had studied both solutions presented above and had arrived at her own compromise, appropriate for the target application of machine translation. As it happened many times, I heard and remembered her words, but it took me years to understand the implications that came

along. In fact, while knowing about every single detail of the phonology and phonetics of Viennese dialect, she had distilled a multi-dimensional matrix of possibilities for encoding into one coherent set of rules that would guide us through the whole project, generating a comprehensive output of the MT-system that would be readable, but also quite ready for speech-synthesis (as I found out some time after the project ended – it needed one line of code to replace one character and it worked). It was a success story, maybe of little impact, yet, but hopefully an instructive case-study for future enterprises. Let us dive into details once more:

Characters used: for reasons of readability, but also fostering consistency, the set of characters was reduced following Hornung and abandoning Artmann's strategy of employing characters that would represent multi-phone strings: <x> is replaced by <gs/ks>, <q> by <gv> and <z> by <ds/ts>; <v> and <y> are not used in any of the three transliteration systems.

'a' and 'o': the three-way distinction is resolved by introducing the character <å>, but not as Hornung does only for nasal /a/, but generally for all rounded occurrences of the sounds corresponding to SG /a/. This is the only additional character in the set of characters for this orthography; no diacritics are used.

Plosives: here, Artmann's strategy is followed more or less without modifications: geminate/fortis plosives are represented by the letters <p, t, k>, others are represented by <b, d, g>. This strategy makes much sense; however, it results in a situation where two identical phonological constellations are transliterated in two different ways. It is well known that in Bavarian dialects onset plosives are neutralised towards the lenis variant, except for the velar plosive, which resists neutralisation and rather forms a velar affricate: *goatn* [g̊ɔatn] – SG *Garten* 'garden' vs. *koatn* [g̊^{h}ɔatn̩] – SG *Karten* 'cards'. That it is indeed an affricate can be shown by the identity of the two occurrences of /g/ and /h/ in the following example, one corresponding to a fortis /k/ in SG, the other being a lexical co-occurrence of the two sounds: *den koidn de kånsd da ghåidn* [den ghɔɪdn̩ deː ghɔnsd da ghɔɪdn̩] – SG *den kalten Tee kannst du dir behalten* 'you can keep that cold tea for yourself'. An alternative would be to totally dispense with letters encoding fortis consonants, thus having <b, d, g> and <gh> in onsets, and simplex vs. geminate forms in other contexts. This would, however, carry over to fortis plosives after sonorants, whereupon it would be quite questionable to analyse them as geminates. A fortiori, it is one of the few commitments to standard orthography that all three approaches adopt. Nevertheless, this is the only inconsequence – morphologically

conditioned geminates are well reflected in orthography, e.g., *i red* – SG *ich rede* 'I speak' vs. *ea ret* (=*redd*) – SG *er redet* 'he speaks'.

Geminates: vowel length seems to be subsumed under the phonological context of subsequent consonants. Phonetically, the distinction between simplex and geminate forms seems well-grounded, so what remains of vowel length should rather be encoded on the side of subsequent consonants. Plosive geminates are encoded by characters corresponding to fortis consonants in SG, others are marked by reduplication. I want to report here that we indeed had a hard time with sonorants (reflecting Artmann's ambivalent, but phonetically precise decisions to write one or two letters). Regarding fricatives, it was much more obvious when they would be geminates or not – still, we refrained from introducing duplication to the multi-character strings <ch> and <sch> – leaving those (very rare cases) in ambiguity.

Diphthongs: are encoded as suggested by Artmann (<au>, <ei>, <äu>); however, the secondary Umlaut diphthong resulting in /a/ is deliberately encoded by double <aa>, hence *baam* – SG *Bäume* 'trees'; *draam* – SG *Traum* 'dream', as opposed to *dram* – SG *Tram/Straßenbahn* 'tram'.

Nasals: in order to avoid a special diacritic (to encode it properly), nasal consonants are always retained, regardless of whether they are pronounced or not. Recall that Artmann's strategy was rather to assign <au> to both /ao/ as a diphthong and /an/ as a constellation where the vowel is clearly nasalized and also gets a diphthong-like interpretation with a rounded feature. Even when the nasal consonant is clearly facultative (where Hornung would use superscripts), the nasal is transcribed. This also facilitates morphological decoding: *schdaan* vs. *schdaana* – SG *Stein* vs. *Steine* 'stone / stones', the stem is identical, and the plural ending –a is easily identifiable as such.

Liquid-vocalisation: follows the outcome of a phonological process. /r/-vocalisation will be encoded by <a> where applicable, /l/-vocalisation comes in two variants: after non-front vowels, it changes into a front-glide, represented by <i> (e.g., *duipn* – SG *Tulpe* 'tulip'; note that place assimilations of nasals are deliberately not encoded). With front vowels, /l/-vocalisation results in secondary Umlaut, which will be represented with a single character, even in open syllables, e.g., *mö* – SG *Mehl* 'flour'. Epenthetic, or rather intrusive /r/s are not encoded in the text. Although intrusive /r/s may occur abundantly in original speech of VD, they are neither motivated lexically nor morphologically, but merely phonologically. For an MT-system this would only be confusing, for an appropriate output, one could easily think of a post-processing component that inserts

the intrusive /r/s in the right place. This is a gross simplification – it is indeed not so easy to identify the correct places to insert an intrusive /r/ – so we just spared it out.

Weak pronoun forms: most personal pronouns have a full form that is used when stressed, and a reduced form that is used when the pronoun is prosodically weak. Interestingly, the Nom/Acc weak forms for feminine and neuter 3P.Sg become homophonous: *si* → *s* (3P.Sg.Fem) and *es* → *s* (3P.Sg.Neut). In order to dissolve this ambiguity we marked with an apostrophe where the vowel deletion has taken place: *s'* vs. *'s* – a practice often found in the transliteration of vernacular texts. In retrospect, I am not all too happy with this decision: first, it introduces a new character – the apostrophe – with a specific function, but only occurring with pronouns. And there is one phonological effect where using apostrophes generates an impression (of active vowel deletion) that runs counter to intuition. This effect arises in the context of multiple clitic weak pronouns where two adjacent ones would be represented as /s/: in such a case, the two would be contracted into one /s/, dropping the information that there exist two distinct pronouns. Parallel to the phonological strategy of inserting /r/ intrusively between two adjacent vowels, the vowel /a/ is inserted between the two sibilants. Consider the following example: *dån håd s a si s* åndas *ibalegd.* – SG *dann hat sie sich es anders* überlegt 'then she changed her mind'. Since the first pronoun '*s*' represents the feminine, stemming from *'si'* (Artmann would write *se*, instead), one could argue that the intrusive /a/ replaces the vowel that was lost due to reduction. This idea, however, is flawed. What we have in VD are two identical lexical entries, both of them consisting just of the sound /s/. Take an example where the order of pronouns is reversed: *dån håd s a s auf aamåi intressiad.* – SG *dann hat es sie auf einmal interessiert* 'all at once it started to interest her'. Using apostrophes here looks quite confusing: *s' a 's.* Intuitively, one interprets the apostrophes as marking a missing vowel, but both apostrophes point towards the inserted vowel /a/. Nevertheless, for machine translation the adopted strategy with apostrophes does not pose any problems; indeed it fosters precision due to the lack of ambiguity. That the result is not always correct will hardly be noticeable, since such combinations are rather rare. (In fact, our system outputs: *Dån håd 's s' auf aamåi intressiad.*) In order to correct this problem, one would have to introduce a post-processing rule that targets exactly these pronouns, but for such a rule-based transformation, the apostrophes may also be advantageous.

CONCLUSIONS

In this paper, I have tried to review the make-up of an orthography for Viennese dialect that was developed by Sylvia Moosmüller in the course of a project that targeted machine translation between Standard German and Viennese dialect. In order not to repeat information published elsewhere (Hildenbrandt et al. 2013), I decided to first give a more general outlook on the problems one faces with such an enterprise, and second, to base the discussion on two influential predecessors: H.C. Artmann and Maria Hornung. Each of them follows his/her own targets (literature, lexicography) which provides a fantastic opportunity to illustrate, how a specific goal determines most decisions that have to be made in order to create an orthography. These two works were also taken as two diverging schemata on the basis of which it became possible to set up an orthography optimised for language technology.

The attentive reader may have noticed my ironic remark in brackets about 'yet another' orthography. As a matter of fact, most 'yet another' enterprises follow a specific purpose. Ours was machine translation, or language technology in general, where an orthography for a specific dialect should not encode too much phonetic variation, as Artmann's orthography does (for very good reasons); ideally none. In addition, the complete phonological differentiation provided in Hornung's lexicon, which forces her to adopt a set of additional characters plus typographic means in order to again encode variation, was not apt for the purposes we followed. I hope I could stress this point enough, namely that it is essential to review one's own goals first. The solution presented here is optimised for machine translation (and also works for speech synthesis), but it might not be ideal at all for documentation. Working with dialects means starting over and over again; however, one also needs to bear in mind that there are predecessors that one not only can but definitely should draw insights from. I have also tried to report on the multiple phonological and phonetic considerations that guided the decisions in each of the orthographic systems. In the contemporary paradigm of language technology, such issues may appear peripheral, but I hope to have made clear that encoding issues such as defining an orthography are at the core of language technology.

It is like working on a tunnel: from the one side phonology has to be understood better – it is definitely not enough to find a set of phones or phonemes for a given language (variety). Meanwhile, a close eye has to be kept on the concrete phonetic realisation. On the other hand, we need

to become able to deal with abundant variation in transliterations of particular language varieties. Social media provides ample sources for such variation, and variation is fun.

REFERENCES

Artmann, H.C. 1958. *med ana schwoazzn dintn. gedichta r aus bradnsee*. Salzburg: Otto Müller Verlag.

Haddow, Barry, Adolfo Hernández-Huerta, Friedrich Neubarth and Harald Trost. 2013. Corpus Development for Machine Translation between Standard and Dialectal Varieties. In: *Proceedings of the Workshop 'Adaptation of Language Resources and Tools for Closely Related Languages and Language Variants' of the 9th Int. Conf. on Recent Advances in Natural Language Processing* (RANLP 2013), Sept. 13th 2013, Hissar, Bulgaria, 7–14. [online: www.ofai.at/~friedrich.neubarth/papers/ranlp2013.pdf]

Hildenbrandt, Tina, Sylvia Moosmüller and Friedrich Neubarth. 2013. Orthographic encoding of the Viennese dialect for machine translation. In: Zygmunt Vetulani, Hans Uszkoreit (Hrsg.) *Human Language Technologies as a Challenge for Computer Science and Linguistics*, Proceedings of the 6th Language & Technology Conference (LTC'13), Dec. 7-9, 2013, Poznań, Poland, 399–403. [online: www.ofai.at/~friedrich.neubarth/papers/ltc2013-hildenbrandt.pdf]

Hornung, Maria. 2002. *Wörterbuch der Wiener Mundart*. In collaboration with Leopold Swossil. Wien: ÖBV, Pädagogischer Verlag, 2nd edition, 1st edition: 1998.

Kühnhammer, Klaus. 2004. *Isochrony in Austrian German*. M.A. thesis, University of Vienna.

Neubarth, Friedrich, Barry Haddow, Adolfo Hernández-Huerta and Harald Trost. 2013. A hybrid approach to statistical machine translation between standard and dialectal varieties. In: Zygmunt Vetulani & Hans Uszkoreit (Hrsg.) *Human Language Technologies as a Challenge for Computer Science and Linguistics, Proc. of the 6th Language & Technology Conference* (LTC'13), Dec. 7-9, 2013, Poznan, Poland, 414–418. [online: www.ofai.at/~friedrich.neubarth/papers/ltc2013-neubarth.pdf]

Neubarth Friedrich and Harald Trost. 2017. Statistische Maschinelle Übersetzung vom Standarddeutschen in den Wiener Dialekt. In: Claudia Resch & Wolfgang U. Dressler (eds.) *Digitale Methoden der Korpusforschung in Österreich. Österr. Akademie der Wissenschaften*, 180–203.

Schuster, Mauriz. 1956. *Sprachlehre der Wiener Mundart*. Edited by Hans Schikola, Wien, Reprint: Wien 1984.

Schikola, Hans. 1954. *Schriftdeutsch und Wienerisch*. Wien.