Orthographic encoding of the Viennese dialect for machine translation

Tina Hildenbrandt [1], Sylvia Moosmüller [1], Friedrich Neubarth [2]

[1] Acoustics Research Institute, Austrian Academy of Sciences
[2] Austrian Research Institute for Artificial Intelligence (OFAI)
tina@kfs.oeaw.ac.at, sylvia.moosmueller@oeaw.ac.at, friedrich.neubarth@ofai.at

Abstract

Language technology concerned with dialects is confronted with a situation where the target language variety is generally used in spoken form, and, due to a lack of standardisation initiatives, educational reinforcement and usage in printed media, written texts often follow an impromptu orthography, resulting in great variation of spelling. A standardised orthographic encoding is, however, a necessary precondition in order to apply methods of language technology, most prominently machine translation. Scripting dialects usually mediates between similarity of a given standard orthography and precision of representing the phonology of the dialect. The generation of uniform resources for language processing necessitates considering additional requirements, such as lexical unambiguousness, consistency, morphological and phonological transparency, which are of higher importance than readability. In the current contribution, we propose a set of orthographic conventions for the dialect/sociolect spoken in Vienna. This orthography is primarily based on a thorough phonological analysis, whereas deviations from this principle can be attributed to disambiguation of otherwise homographic forms.

Keywords: machine translation, phonology, orthography, Viennese dialect

1. Introduction

Applying methods of language technology to dialectal varieties of major languages is a rather new field of research that poses specific challenges. State-of-the art human-machine interfaces deal with standard varieties, relying on a given, standardised orthography. In the light of personalised or regionalised services, the use of dialects or sociolects may provide a strong incentive to use such a system. Taking speech synthesis as a prominent example, it is not so difficult to provide localised pronunciation, but dialectal language varieties involve substantial deviations from standard language on various linguistic levels. For that reason, it is necessary to capture and reproduce all major idiosyncrasies displayed by a certain language variety, be they syntactic, lexical or phonological in nature.

Apart from the fact that written corpora of dialectal language varieties are typically rare, we face the serious problem that orthographic conventions can vary greatly among existing resources. For methods of language technology, but especially for machine translation (MT), this is intolerable. What is needed is a consistent and comprehensive way to identify words in a language variety on the basis of their orthographic representation. The key focus of this paper will be a specification of how to obtain such an orthographic representation for a given variety, optimised for the purposes outlined above. The variety targeted in this work is the Viennese dialect.

The reflection of regional characteristics always stands against attempts to arrive at a supra-regional standard as a compromise but also as a driving force for language standardisation. Designing an orthographic system for dialects, that almost by definition manifest mainly in an oral tradition, faces similar problems. People, when they write in a dialect, freely oscillate between adopting standard orthography and attempting to represent their phonetic knowledge of their dialect with graphemes. Using ‘familiar’ spelling clearly helps comprehension, while the latter strategy stresses the originality, and sometimes even has to be applied to words or word-forms that are not existent in the standard language.

As a first step, it seems indispensable to thoroughly analyze the phonological properties of a particular variety, and in a second step to find an optimal way to represent these properties with orthographic symbols. When producing literary texts, readability might be a prominent issue, so the orientation towards the standard orthography may have a higher priority. In the light of applicability in speech technology (i.e. speech synthesis, MT), accuracy towards phonology constitutes the most important criterion, on a par with the necessity to minimise lexical ambiguities. Another concern is the question whether the use of non-standard symbols and diacritics should be favoured or disfavoured; maintenance issues regarding lexical resources and text corpora point to the latter.

The work presented in this paper stems from the project ‘Machine Learning Techniques for Modeling of Language Varieties’ (MLT4MLV) that deals with statistical machine translation (SMT) between standard and dialectal varieties. In the following section, we discuss sociolinguistic aspects of the Viennese dialect, focusing on salient features that contribute to its uniqueness. Section 3 is dedicated to phonological phenomena and their representation within the proposed orthographic system, while section 4 outlines the formal characteristics that lead to specific decisions, aiming at a higher degree of generalisability.

2. Sociolinguistic aspects of the Viennese dialect

Urban dialects are primarily social dialects, usually spoken by the lower social classes of a city or metropolis (Labov 2001). This observation also holds for the Viennese dialect (VD). Moreover, it follows naturally from population density that a higher degree of contact between varieties can be observed. In particular, interaction between VD and Standard Austrian German (SAG) takes place. Therefore, the different social groups make use of specific features of the other group, mainly for pragmatic reasons. Hence, an authentic VD speaker also performs switches into the standard variety.

This observation led Dressler and Wodak (1982) to develop a new model of interaction between varieties, in particular between SAG and VD. Evaluating the differences between the varieties within the framework of Natural Phonology (Dressler, 1984), the prevailing vertical relationship between varieties was given up in favour of a two-competence model. Within this model, dialects and the standard variety are conceived on equal terms, both linguistically and socially. Differences that do not meet the requirements for a phonological process, i.e. differences that lack phonetic motivation and gradualness, are connected by so-called input-switch-rules. These different forms can be traced back to different historical developments of these varieties. They comprise alternations as, e.g., /uː/ [g̊uːt] (Standard) alternating with /uɐ/ [g̊uɐd̥] (dialect) *gut* (‘good’), /iː/ [liːb̥] alternating with /iɐ/ [liɐb̥] *lieb* (‘dear’), or /ɑ(ː)/ [vɑsːɐ] alternating with /ɔ(ː)/ [vɔsːɐ] *Wasser* (‘water’). These sounds represent highly salient features and are easily recognised as belonging to either the standard or the dialect variety. Phonological processes, on the other hand, can belong either to the dialect or to the standard variety or to both. Nasal assimilation or the vocalisation of /r/ represents an example for a phonological process belonging to both varieties. The vocalisation of the lateral, on the other hand, constitutes a phonological process belonging to the Middle Bavarian dialects. It represents an attribute for social categorisation, which may lead to social stereotyping. As a consequence, this process becomes socially diagnostic, i.e. socially distinctive (Kristiansen, 2001).

From this evaluation of differences, a hierarchy of saliency can be derived: input-switch rules are more salient than dialectal phonological processes. These, in turn, are more salient than phonological processes belonging to both varieties. This hierarchical model of saliency has been corroborated by the results on attribution tests performed by Soukup (2009).

Expectations of listeners clearly point towards a higher acceptance of a stereotypical representation of dialects, which is equated with authenticity (cf. Moosmüller 2012). A stereotypical representation of VD excludes switching into SAG and concentrates on highly salient, diagnostic features of the dialect. These theoretical groundings and empirical results have consequences on speech technology applications and need to be considered.

3. A concise phonological description of the Viennese dialect and its orthographic conventions

Viennese dialect belongs to the Central Middle Bavarian dialects. However, being the dialect of a metropolis, it displays specific, characteristic features of its own.

In the following section, we will describe the most prominent phonological properties that lead us to propose specific orthographic conventions differing from Standard German. In particular, these are:

- Delabialisation of front vowels
- /ɑ/ ↔ /ɔ/
- Viennese Monophthongisation
- Merger of /e/ and /ɛ/
- Neutralisation of plosives
- Compensatory lengthening of V+C sequences
- Vocalisation of the liquids /r/ and /l/
- Word final nasals
- Reduction of prefixes

3.1. Delabialisation of front vowels

The vowel system of the Middle Bavarian dialects comprises 14 vowels /i(ː), e(ː), ɛ(ː), u(ː), o(ː), ɔ(ː), ɑ(ː)/. Historically, front rounded vowels were de-labialised in the Bavarian dialects, rendering e.g., /g̊lig̊ː/ vs. /glyk/ *Glück* (‘luck’). Delabialisation also affected the diphthong /ɔɛ/, which merged with /aɛ/. Accordingly, the vowels are orthographically represented as such:

(1)	VD	IPA	SAG	gloss
	‹glik›	[g̊lig̊ː]	*Glück*	'luck'
	‹dia›	[d̥iːɐ]	*Tür*	'door'
	‹kepf›	[keb̥ːf]	*Köpfe*	'heads'
	‹efn›	['eːfm̩]	*Öfen*	'furnaces'
	‹eich›	[æːç]	*euch*	'you'
	‹heid›	[hæːd̥]	*heute*	'today'

3.2. /ɑ/ ↔ /ɔ/

SAG /ɑ(ː)/ alternates with dialectal /ɔ(ː)/ (see input-switch rules). However, preceding a nasal consonant, /ɔ(ː)/ is realised as /ɒ̃(ː)/, e.g., [ˈhɒ̃mːɐ] *Hammer* (‘hammer’). Before word or morpheme boundaries, the nasal consonant is deleted, and the vowel undergoes compensatory lengthening, e.g., [mɒ̃ː] *Mann* (‘man’) or [ˈɒ̃ːfɒ̃ŋɐ] *anfangen* (‘to begin’).

In order to represent this alternation, one character not present in SAG orthography was introduced into the set of graphemes: ‹å›, representing the rounded counterpart /ɔ/ to SAG /ɑ/. This is the only way to prevent graphemic ambiguities between /o/ and /ɔ/, as exemplified in (2).

(2)	VD	IPA	SAG	gloss
	‹offn›	[ofːm̩]	*offen*	‘open’
	‹åffn›	[ɔfːm̩]	*Affen*	‘monkeys’
	‹hosn›	[hoːsn̩]	*Hosen*	‘pants’
	‹håsn›	[hɔːsn̩]	*Hasen*	‘bunnies’
	‹hoid›	[hoed̥]	*holt*	‘(s/he) fetches’
	‹håid›	[hɔed̥]	*halte*	‘hold’ (IMP)

3.3. Viennese Monophthongisation

In VD, the diphthongs /aɛ/ (including delabialised /ɔɛ/) and /ɑɔ/ developed into monophthongised /æː/ and /ɒː/, respectively, being of equal length as the diphthongs. As a result, the basic vowel system of VD has to be extended by these two (long) monophthongs: /æː/ and /ɒː/ (cf. Schikola, 1954). Even though, monophthongisation is a highly salient feature of VD, we use the spelling of the corresponding SAG diphthongs, in order to avoid the introduction of special characters:

(3)	VD	IPA	SAG	gloss
	‹haus›	[hɒːs]	*Haus*	‘house’
	‹auto›	[ˈɒː d̥ːo]	*Auto*	‘car’
	‹weich›	[væːç]	*weich*	‘soft’
	‹eisn›	[æːsn̩]	*Eisen*	‘iron’
	‹heid›	[hæːd̥]	*heute*	‘today’

SAG /aɛ/ that evolved from Middle High German (MHG) *ei* historically developed into /ɔɐ/ in Bavarian dialects, realised as /ɑː/ in VD and some dialects of Southern Carinthia. Similarly, MHG /ou/, preceding nasals, also changed to /ɑː/ in the Bavarian dialects (e.g., ‹baam› /b̥ɑːm/ *Baum* ‘tree’). As a monophthong, it is significantly longer in stressed positions. For this reason, but also to discern it from other instances of ‹a› that correlate to /a/ in SAG, we represent these monophthongs with double characters: ‹aa›. As an example, VD distinguishes between, ‹weis› /væːs/ *weiß* (‘white’) and ‹waas› /vɑːs/ *weiß* ‘(I/she) know(s)’, homophones in SAG.

A further instance of monophthongisation is the rounded variant of [æː] occurring in the context of /l/-vocalisation. Its phonetic realisation is [œː], and orthographically we represent it with the string ‹äu› (the only case where the character ‘ä’ is used).

Because length in consonants is encoded throughout and vowel length is predictable (see below) there is no need to encode vowel length orthographically. This opens up the possibility to use doubling of characters for other purposes.

3.4. The merger of /e/ and /ɛ/

In VD, /e(ː)/ and /ɛ(ː)/ coincide. The merger of the vowels /e(ː)/ and /ɛ(ː)/ (and its rounded variant /ø(ː)/ and /œ(ː)/, respectively, see section 3.7) is a change that has long been observed in Middle Bavarian dialects. In the central Middle-Bavarian regions to which Vienna belongs, MHG /ɛ/ changed to /eː/, in combination with a lengthening process, whereas the quality of /ɛ/ was retained in words with a short vowel. Therefore, MHG *‹rëgen›* /rɛgen/ *Regen* (‘rain’) was changed to Middle Bavarian [reːŋ], realised in the same way as MHG ‹legen›, Middle Bavarian [leːŋ] *legen* (‘to lay’), whereas /ɛ/-vowels in words like MHG ‹lëcken› *lecken* (‘to lick’) retained their quality (see Moosmüller and Scheutz, 2013 for a discussion). In Vienna, an expansion of this merger is observed since the 1970ies, which renders an arbitrary usage of /e(ː)/ and /ɛ(ː)/, with a preference to realise [e(ː)] (Seidelmann, 1971). This tendency is still prevalent in VD (Seidelmann, 1971, Moosmüller and Scheutz, 2013). Note that /ɛ/ corresponding to SAG orthographic ‹ä› is encoded with the letter ‹e› throughout.

3.5. Neutralisation of plosives

An important feature of Bavarian dialects concerns the neutralisation of fortis and lenis plosives. This neutralisation is most prevalent in word-initial position, e.g., SAG *Torte* (‘cake’) and *dort* (‘there’) are pronounced identically in VD, namely ‹duatn› [ˈd̥ʊɐd̥ːn̩]. Due to historical reasons (Kranzmayer, 1956), neutralisation does not affect velar plosives in simplex onsets. Words such as, ‹gåatn› [ˈg̊ɔɐd̥ːn̩] *Garten* (‘garden’) and ‹kåatn› [ˈkɔɐd̥ːn̩] *Karten* (‘cards’) are distinct. Before sonorants, velar plosives are neutralised, leading to homophony of *klauben* (‘to pick up’) and *glauben* (‘to believe’): ‹glaubn› [g̊lɒːb̥m̩].

This fortis/lenis distinction of consonants is encoded as it appears in the dialect, and not adopted from Standard German orthography. Word-medial and word-final OHG fortis plosives are retained in the contexts of preceding nasals and vocalised liquids, as well as fortis plosives that had emerged from OHG geminates.

(4)	VD	IPA	SAG	gloss
	‹dant›	[d̥ɑnd̥ː]	*Tante*	‘aunt’
	‹bupn›	[ˈb̥ub̥ːm̩]	*Puppe*	‘doll’
	‹båikn›	[ˈb̥ɔeg̊ːŋ̩]	*Balken*	‘beam’

3.6. Compensatory lengthening of V+C sequences

In many Bavarian dialects, compensatory lengthening, or phonological isochrony (Ronneberger-Sibold, 1999), triggers an alternation between two patterns: either a (stressed) long vowel is followed by an intervocalic or domain final lenis consonant, or a short vowel is followed by a fortis consonant. As concerns plosives, we decided to encode this relationship with the respective lenis and fortis graphemes, as concerns fricatives, we decided to encode this relationship with double characters: ‹ofn› *Ofen* (‘oven’) vs. ‹offn› *offen* (‘open’), which corresponds to distinctions also made in Standard German orthography, whereas ‹lauffn› *laufen* (‘run’) reflects the phonological shape of that word, both in VD and SAG, which is (no longer) encoded in standard orthography. Such alternations can be due to morphological processes. One example is verb inflection, where the ending –d is assimilated to the stem consonant /d/: ‹i red› *ich rede* (‘I say’) vs. ‹si ret› (= /red+d/) *sie redet* (‘she says’). Other examples are plural formation of nouns having an /ɛ/ plural ending in SAG, but none in VD. They display compensatory lengthening effects: ‹disch› *Tisch* vs. ‹dischsch› *Tische* (‘table’ vs. ‘tables’), sometimes going along with stem alternations involving Umlaut ‹kobf› *Kopf* vs. ‹kepf› *Köpfe* (‘head’ vs. ‘heads’).

3.7. The vocalisation of /r/

Liquids are vocalised in word-final position and before consonants in the Middle Bavarian dialects. Preceding /r/, the quality of [+constricted][1] vowels changes to [-constricted]. Subsequently, /r/ is vocalised to [ɐ], e.g., /fiːr/ → [fiːɐ] *vier* (‘four’). Preceding the vowel /ɑ/, [ɐ] is absorbed, resulting in [ɑː], e.g., /kɑrbfn/ → [kɑːb̥fm̩] *Karpfen* (‘carp’). In word-final position, [ɛɐ]-sequences resulting from /r/-vocalisation usually undergo total assimilation. The default realisation is [ɐ], e.g., [ˈleːdɐ] *Leder* (‘leather’). However, this vowel, being unstressed, is subject to massive context dependent assimilation. We represent all instances of /r/-vocalisation with ‹a›:

(5)	VD	IPA	SAG	gloss
	‹schmåan›	[ʃmɔɐn]	*Schmarren*	‘cut-up pancake’
	‹mamelad›	[mɑmɛˈlɑːd̥]	*Marmelade*	‘jam’
	‹kafioi›	[kɑfiˈʲoe]	*Karfiol*	‘cauliflower’

3.8. The vocalisation of /l/

Preceding a lateral, unrounded vowels are rounded (with the exception of rarely occurring /ɑ/). When the lateral is neither followed by a vowel nor syllabic, it becomes vocalised, e.g., [ˈhoed̥s] *Holz* (‘wood’), [ˈkɔed̥] *kalt* (‘cold’). Following a front vowel, the vowel resulting from vocalisation is absorbed, e.g., [ˈfyː] *viel* (‘much’), [ˈmœː] *Mehl* (‘flour’). Word-initially, after alveolar and

[1] Instead of the highly problematic feature [±tense], the feature [±constricted] is proposed to distinguish the relevant vowel qualities, see Moosmüller (2007) for a discussion.

post-alveolar obstruents, between back vowels, and in cases where the vocalisation of the lateral is suppressed (due to interaction with SAG), the lateral is velarised in VD. After bilabials, either the lateral remains clear or it is realised as a retroflex, after velars, it is palatalised.

/l/-vocalisation is a very characteristic phonological process of the Middle Bavarian dialects. Examples (6a-b) show /l/-vocalisation after front vowels spawning a series of front-round vowels (and one monophthongised diphthong [œː]) that are represented by ‹ü›, ‹ö› and ‹äu›. In intervocalic positions, the lateral is not vocalised, as exemplified in (6b’). (6c), finally, gives examples of /l/-vocalisation after round, back vowels. We encode the unabsorbed off-glide with ‹i›, leading to the graphemic representations ‹ui›, ‹oi›, and ‹åi›.

(6)	VD	IPA	SAG	gloss
a.	‹müch›	[myːç]	*Milch*	‘milk’
	‹deifö›	[ˈd̥æːfœ]	*Teufel*	‘devil’
b.	‹wäu›	[vœː]	*weil*	‘because’
b’.	‹äulich›	[ˈœːliç]	*eilig*	‘hurried’
c.	‹duipn›	[ˈd̥ueb̥ːm̩]	*Tulpe*	‘tulip’
	‹soidåd›	[soeˈd̥ɔːd̥]	*Soldat*	‘soldier’
	‹schdåi›	[ʃd̥ɔe]	*Stall*	‘stable’

Liquid vocalisation is not necessarily accompanied by compensatory lengthening of its preceding vowel. Thus, although the trigger /l/ may not be orthographically represented, length on the vowel is not encoded.

3.9. Word-final nasals

Word-final nasals, on the other hand, which are often absorbed while spreading the feature [+nasal] onto the preceding vowel, are fully retained in the orthography. An alternative option would have been to use a diacritic (e.g., ‘~’) or to not encode nasality at all in these contexts. Both options were dismissed for keeping the grapheme inventory as minimal as possible and saving the orthographic encodings from ambiguity, e.g., ‹wein› [væ̃ː] *Wein* (‘wine’). There is one exception concerning determiners involving ‹-aa› (without ‹n›), corresponding to SAG ‹-ein› (e.g., *ein, kein, irgendein*). Here, nasality was lost during historical development, but also it is the only way to distinguish these forms from forms involving an additional –n suffix: ‹aa› (NOM.MASC) *ein* ‘a/one’ vs. ‹aan› (ACC.MASC) *einen* ‘a/one’). This pattern carries over to possessive pronouns with a similar form (‹-ei›): ‹mei auto› *mein Auto* (‘my car’) vs. ‹in mein auto› *in meinem Auto* (‘in my car’).

3.10. Reduction of prefixes

A further trait of Bavarian dialects regards the prefixes *ge–* and *be–* which are reduced, i.e. the vowel, which is always unstressed in this position, is not realised. In VD, the deletion of the vowel of the prefix *ge–* occurs before any following segment, as shown in (7a). However, preceding plosives, the velar plosive representing the prefix is absorbed, as shown in (7b).

(7)	VD	IPA	SAG	gloss
a.	‹gåabeit›	[ˈg̊ɔɐßæd̥ː]	*gearbeitet*	‘worked’
	‹gmåchd›	[g̊mɔxd̥]	*gemacht*	‘made/done’
	‹gfångan›	[ˈg̊fɒ̃ŋɛ̃]	*gefangen*	‘captured’
b.	‹båkd›	[b̥ɔg̊ːd̥]	*gepackt*	‘packed’
	‹dån›	[d̥ɒ̃ː]	*getan*	‘done’
	‹kaufd›	[kɒːfd̥]	*gekauft*	‘bought’

3.11. Generalising orthographic conventions

Apart from Hornung (1998), there are no lexical resources for VD with a consistent and extensive orthography. However, this proposal is not directly transferable to our purposes since it involves abundant usage of diacritics, favouring phonetic accuracy. From the perspective of language technology, special characters and diacritics are treated as a last resort when ambiguities cannot be avoided by other means. For our VD orthography only one additional character is introduced, namely ‹å›, in order to distinguish /o/ and /ɔ/ vs. /a/. On the other hand, a few characters of the German alphabet were replaced by other characters/ strings: ‹z› corresponds to ‹ds› or ‹ts› (depending on the fortis/lenis status of the sound), ‹x› corresponds to ‹gs› and ‹ks›, whereas ‹v› is pronounced as /v/ or /f/ even in SAG and therefore orthographically represented as ‹w› or ‹f› in our VD orthography. The letters ‹ä› and ‹c› are only used in the combinations ‹äu›, ‹ch› and ‹sch›. Certain characters are not used: ‹q›, ‘sharp s’ ‹ß› and ‹y›.

Orthography should represent phonology as close as possible, but distinctions that either are not decisive (e.g., /e/ vs. /ɛ/) or can be retrieved otherwise (e.g, vowel length) should be dispensed with. There is little room for exceptions: ‹eea› *eher* (‘rather’) would be indistinguishable from ‹ea› *er* (‘he’) if it were written according to the conventions. Regarding specific phonological peculiarities of VD, we ended up with the following orthographic conventions: lenis/fortis plosives with characters existing in the alphabet (e.g., ‹d› vs. ‹t›) and fricatives with double characters (e.g., ‹s› vs. ‹ss›) /r/ vocalization as falling diphthongs involving the character ‹a› (e.g., /i+r/ as ‹ia›), /l/-vocalization as it is pronounced (rounded front vowels or rising diphthongs, e.g., ‹oi›). On the other hand, vocalisation of final nasals cannot be reconstructed from the context alone or without using diacritics, hence we have to retain the nasal in the spelling although it is (usually) not pronounced.

4. Orthography for SMT

The development of an orthographic representation for a language variety would be a pointless exercise if it were not used within a practical application, such as SMT. This paper is not about a technical methodology per se, rather it analyses and describes the steps that have to be undertaken before methods of LT can be applied. Therefore, it is not possible to present results of an evaluation concerning the orthography itself. However, it might be worthwhile to ascribe the principles guiding the design of an orthographic system to the needs of LT.

SMT between closely related languages or varieties often deals with a target language with very little resources, whereas the strong similarities between the two can be useful to apply specific methods. In certain cases, the rich resourced variety may serve as a pivot for translating between the less resourced variety and some other major language (cf. Nakov and Ng 2012). Phrase-based MT on the word level is able to incorporate many idiosyncrasies of the target language, but due to the relative limitation of training data, it will not be very effective, given the great number of out-of-vocabulary words to be expected. To overcome such a shortcoming, a second layer of translation on level of characters (unigram, bigram) is introduced. It is trained on word pairs filtered by the phonetic proximity between the two varieties (cognates). In combination, these two levels provide promising results, for details, see Haddow et al. (2012). Nakov and Tiedemann (2012) describe a similar strategy for exploiting the similarity between Bulgarian and Macedonian.

Here, consistency and economy regarding the set of characters come into play. Mixing various orthographic conventions will decrease the chance that a character level model captures the dependencies between strings of characters. Additionally, extending the set of characters in favour (narrow) phonetic accuracy decreases the coverage of the training data with regard to possible combinations. The threshold for reduction of the set of characters obviously is unambiguousness: homographs are only allowed if they are also homophonous.

Why do we so strongly emphasise phonological transparency, when any consistent and unambiguous writing system would serve our purposes? Dialects mainly manifest in oral communication. This invites us to envisage the application of speech technologies to dialectal varieties. The closer the orthography is to the actual phonetic realization in a given language variety, the easier it will be to provide the necessary resources to directly access these technologies (pronunciation lexicon, letter-to-sound rules). Still, it is necessary to link dialectal varieties to a standard variety, since the textual resources for developing human-machine interfaces (dialogue systems, screen readers) will be available only in the standard variety in most cases. Hence, it seems necessary to provide both, a MT system to link these varieties on a textual base, and speech interfaces capable of dealing with dialectal varieties. The orthography proposed in this paper constitutes one common link between these divergent needs.

5. Conclusion

Focusing on (statistical) MT as the primary context, the orthographic system we developed for the Viennese dialect was designed to maximise the following criteria: consistency, unambiguousness, phonological transparency and mainly using characters that are present in the set of SAG orthography. While some of the decisions we had to take were straightforward, others reflect compromises between conflicting strategies. Phonological transparency is of special importance, since it guarantees that the encoding can serve as a reliable and easy-to-process input to systems dealing with dialectal speech.

Acknowledgements

The project ‘Machine Learning Techniques for Modeling of Language Varieties’ (MLT4MLV – ICT10-049, http://varieties.ofai.at/) is funded by the Vienna Science and Technology Fund (WWTF).

References

Dressler, W. U. and Wodak, R. (1982). Sociophonological methods in the study of sociolinguistic variation in Viennese German. In: *Language in Society*, 11, pp. 339-370.

Dressler, W. U. (1984). Explaining Natural Phonology. In: *Phonology Yearbook*, Vol. 1, pp. 29-51.

Haddow, B., Hernandez, A., Neubarth, F. and Trost, H. (2013). Corpus development for machine translation between standard and dialectal varieties. In: Proc. of the Adaptation of Language Resources and Tools for Closely Related Languages and Language Variants workshop, RANLP 2013, Hissar, Bulgaria, pp. 7-14.

Hornung, M. (1998). *Wörterbuch der Wiener Mundart.* In cooperation with L. Swossil, 1st edition, Vienna: ÖBV, Pädagogischer Verlag.

Kranzmayer, E. (1956). *Historische Lautgeographie des gesamtbairischen Dialektraumes*. Vol. 1. Österreichische Akademie der Wissenschaften.

Kristiansen, G. (2001). Social and linguistic stereotyping: A cognitive approach to accents. In: *Estudios Ingleses de la Universidad Complutense*, 9, pp. 129-145.

Labov, W. (2001). *Principles of Linguistic Change (II): social factors.* Massachusetts: Blackwell.

Moosmüller, S. (2007). *Vowels in Standard Austrian German. An acoustic-phonetic and phonological analysis*. Habilitationsschrift, Wien.

Mossmüller, S. (2011). Sound changes and variation in the Viennese dialect. In: Dziubalska-Kołaczyk, K. et al. (eds.), *On Words and Sounds: A selection of papers from the 40th PLM 2009*. Cambridge: Cambridge Scholars Publishing, pp. 134-147.

Moosmüller, S. (2012). The roles of stereotypes, phonetic knowledge, and phonological knowledge in the evaluation of dialect authenticity. In: Calamai, S. et al. (eds.), *Proceedings. of the Workshop “Sociophonetics at the cross-roads of speech variation, processing, and communication”*. Pisa: Le Edizioni della Normale, pp. 49-52.

Moosmüller, S. and Scheutz, H. (2013). Chain shifts revisited: The case of Monophthongisation and *E*-merger in the city dialects of Salzburg and Vienna. In: Auer, P. et al. (eds.), *Language Variation – European Perspectives IV*. Amsterdam: Benjamins, pp. 173-186.

Nakov, P. and Ng, H. T. (2012). Improving Statistical Machine Translation for a Resource-Poor Language Using Related Resource-Rich Languages. In: *Journal of Artificial Intelligence Research*, 44, pp. 179-222.

Nakov, P. and Tiedemann, J. (2012). Combining Word-Level and Character-Level Models for Machine Translation Between Closely-Related Languages. In: *Proc. of the 50th Annual Meeting of the Association for Computational Linguistics*, Jeju, Rep. of Korea, 8-14 July 2012, pp. 301-305.

Ronneberger-Sibold, E. (1999). Ambisyllabic consonants in German: Evidence from dialectal pronunciation of lexical creations. In: Rennison J. and Kühnhammer, K. (eds.), *Phonologica 1996*. The Hague: Thesus, pp. 247-271.

Schikola, H. (1954). *Schriftdeutsch und Wienerisch.* Wien: Österreichischer Bundesverlag für Unterricht, Wissenschaft and Kunst.

Seidelmann, E. (1971). Lautwandel und Systemwandel in der Wiener Stadtmundart. Ein strukturgeschichtlicher Abriß. In: *Zeitschrift für Dialektologie und Linguistik*, 38.2, pp. 145-166.

Soukup, B. (2009). *Dialect Use as Interaction Strategy: A Sociolinguistic Study of Contextualization, Speech Perception, and Language Attitudes in Austria.* Vienna: Braumüller.