# A Distributional Concept for Modeling Dialectal Variation in TTS

Friedrich Neubarth[1] and Christian Kranzler[2,3]

[1] OFAI, Austrian Research Institute for Artificial Intelligence, Vienna, Austria
friedrich.neubarth@ofai.at
[2] ftw., Telecommunications Research Center, Vienna, Austria
[3] SPSC, Signal Processing and Speech Communication Lab, TU Graz, Graz, Austria
kranzler@ftw.at

**Abstract.** Current TTS systems usually represent a certain standard of a given language, regional or social variation is barely reflected. In this paper, we describe certain strategies for modeling language varieties on the basis of a common language resource, in particular Austrian varieties from German sources. The goal is to find optimal procedures in order to represent these differences with minimal efforts in annotation and processing. We delimit the discussion to the lower levels of the transformation of linguistic information – phonetic encoding. One question is if it is necessary or desirable to aim at maximal accurateness of the phonetic transcriptions. We will show that while certain differences could in principle be captured by the context within the speech data, other differences definitely have to be re-modelled, since they either involve ambiguous correspondences, or the string of phones is different in such a way that automatic procedures such as alignment or unit selection would be negatively affected, hence degrade the overall quality of the synthesized speech.

**Keywords:** TTS, language varieties, dialects, Austrian German, regionalization.

## 1 Introduction

With human-computer interfaces regionalization and user adaptation gain increasingly more importance. For TTS systems, this means we have to deal with new challenges since language processing as well as speech synthesis components are available and confined to certain standard varieties of many languages in the world, but regional or socially determined varieties often deviate from these standard varieties to an extent that would suggest treating them as entirely new languages. This would be an uninteresting solution, and it would be (and at present actually is) unattractive to implement these varieties from an economic perspective.

In this paper, we will describe certain strategies one may adopt in order to deal with the problem of modeling different varieties of a given language within a TTS system. We will explicate these strategies based on work done in a project which aims at developing sources and speech-synthesis voices for several Austrian German varieties. The key problem is the trade-off between re-using existing language resources and adaptation of these sources for better accurateness. We will show that while

A. Esposito et al. (Eds.): Multimodal Signals, LNAI 5398, pp. 208–215, 2009.

certain differences between varieties and the given standard could be neglected in principle, because these differences are covered systematically by the speech data, other differences do not display unambiguous correlations, which makes a thorough re-modeling indispensable. However, exploiting linguistic knowledge about the source language and its varieties makes this process feasible in terms of automatic and semi-automatic adaptation.

In the mentioned project, we develop four synthetic voices within the Festival platform [1], one representing Austrian German, and three representing different Viennese varieties, which have to be regarded rather as sociolects than as dialects. Only a few adaptations have to be made from the German standard to Austrian German, and most of them can be done automatically. On the other hand, dialects and sociolects pose more severe problems, and it turns out that adaptations have to be implemented on all levels of linguistic representation, from syntactic processes over lexical specifics to the encoding of the sound inventory.

The German standard is generally defined as the (normative) standard for the language varieties in Germany, Austria and Switzerland. Publicly available resources usually refer to such a standard – for example CELEX for phonetic transcriptions of German words. In reality, German, as many other languages, has to be regarded a pluricentric language, with an Austrian and Swiss variety, as well as a large set of locally (and/or socially) determined varieties [2]. Furthermore, one has to take into account that the boundaries between more standard-like varieties such as Austrian German and more specific varieties are not always as clear-cut as one would wish. Neither is it straightforward to assign a speaker's idiom to a specific variety, nor do speakers restrict themselves to just one idiom, but very often, they display a more or less continuous shift between various levels [3]. Therefore, it is important to exact a strict control over these variables, since the material to be recorded as speech data must be both consistent and characteristic for the language variety to be represented.

## 2 Levels of Representation

In order to deal with potential shifts between standard and dialectal varieties, but also to be able to represent different dialectal varieties exhaustively it is indispensable to establish methods for the transfer of linguistic information. The two constraints guiding such methods are minimization of efforts and full coverage of ambiguities. So, for example, if one variant differs from the other systematically only in vowel qualities, nothing has to be done beyond recording.

However, this is rarely the case. Rather, we expect differences on all levels of representation. In the example at hand, the set of phones[1] cannot be transformed in a one-to-one fashion. Certain phonologically active processes like l-vocalization, while being quite regular, affect neighboring segments and also involve deletion. Certain morphemes are different between standard and varieties, but not always in a systematic

[1] We rather use the term 'phone' here, not only because it is not theoretically biased, but also because it is far more precise. The term 'phoneme' has its origin in structuralist phonology and does not refer to sounds of speech. Moreover, one should ask, whether its alleged status as a cognitive unit is not entirely mistaken.

**Table 1.** Levels of representation encoding differences between Austrian German and Viennese

| Linguistic level | Austrian German | Viennese | Coding level |
|---|---|---|---|
| sound | [ɘ] | [ɛ] | sound |
| symbol set – phones | [a͡ɛ] [a] | [aː] / [æː] / [œː] [a] / [ɔ/ü] | lexicon setup |
| phonology | [fiːl] *viel* 'much' [va͡ɛl] *weil* 'because' | [fyː] *füü* [vœː] *wäu* | rules |
| morphological | *pass-te* 'would fit' *Gläs-chen* 'glass'diminutive | *pass-ert* *Glas-erl* | lexicon transfer |
| morpho-syntactic | *lesen können* 'can read' *ertrinken* 'drown' | *derlesen* *dersaufen* | |
| lexicon – open class – functional: – articles – pronouns | *fett, dick* 'fat' *Kopf* 'head' *der* 'the' *hinaus* 'to-out' *heraus* 'from-out' | *blad* *bluzer* *d' / da / der* *ausse* *aussa* | lexicon specific |
| phrasal | *weil du weggehen sollst!* 'because you should leave' *er ging* 'he went' | *wäusd di iba d'* *heisa haun soisd!* *ea is gangen.* | post-lexical transformation |

way. Regarding lexical specificities it is evident that one needs a cascaded structure of lists of lexicon entries for items that occur only in a certain dialectal variety (or a set of related varieties). Since it has to be assumed that these lists are rarely complete certain decisions regarding pronunciation have to be made on the basis of an estimation taking into account also a morphological analysis and the orthography used in the input. And finally, in Austrian dialects certain processes are active at the phrasal level which also affect the pronunciation of neighboring words: unstressed personal pronouns behave phonologically as clitica, complementizers show up with (cliticized) inflected forms etc. This means that post-lexical rules have to be taken into account as well. In addition, the morphological paradigms lack certain forms systematically: there is no genitive, and there is no indicative preterit (except for auxiliary *sein* 'to be'). With idiomatic differences on the phrasal level, there is almost no chance to provide automatic procedures, at best we can list and use them in semi-automatic translations between standard texts and their counterparts in one of the varieties. A summary of all these differences distributed over various levels of linguistic encoding is given in Table 1,

In the following, we will concentrate on those levels, which have the greatest potential for automatization: the set of symbols representing speech sounds and their correspondences between specific varieties and phonological processes that can be formalized in a rule based way.

## 3 Phone Substitution

The set of speech sounds of a given variety naturally will show differences to other varieties. If only some phones are produced differently, shifted in place of articulation

for example, no efforts would have to be spent to adapt the respective transcriptions, even if some of the used symbols would not depict the correct phonetic items. The retrieval of the correct units would be granted anyway. However, this is not the situation we find. There are n-to-n matchings between the sets, deletions and sometimes insertions as well. So-called neutralizations are a very good candidate for the automatic transfer of lexical information, because they can be formalized as n-to-1 transformations. (In many cases, the situation is not as trivial, since neutralizations are often conditioned by the phonological context.)

To give a few examples, Austrian dialectal varieties and most varieties of the Austrian standard lack voiced sibilants (*singen* /**z**I.N@n/ → /**s**I.N@n/ '*sing*')[2]. This is a case of total neutralization. Plosive consonants in onset position are neutralized toward a (voiceless) lenis variant (*Tante* /**t**an.t@/ → /**d**an.t@/ '*aunt*'), except for the velar /k/ (*Karten* /**k**a:.t@n/ → /**k**A6.tn=/ '*cards*'). As can be seen in the first example, post-nasal (and post-/r/, realized as /6/) contexts also block neutralization. This is a case where both, phonological structure and local information about neighboring segments have to be taken into account in order to formalize the context adequately.

To extend this example, between vowels (but also before syllabic nasals and /l/) the lenis stops tend to spirantize (*Leber* /le:.**b**6/ → /le:.**B**6/ '*liver*'), whereas geminated or fortis stops are not affected by this kind of lenition (*schnuppern* /SnU.p6n / → /Snu.p6n/ '*snuffle*'). Interesting is the context at the end of the syllable, normally targeted by final devoicing. If for example a clitic pronoun starting with a vowel follows within the same prosodic phrase, this position turns into an intervocalic context, which blocks final devoicing and enhances lenition. So for example the PRES.3P.SG ending /t/ turns out as a /d/ (or /D/) – (*kommt+er* /kOm**t** ?E6/ → /kUm.**d**6/ '*comes_he*'). Therefore, we have to assume that in Austrian varieties of German the morpheme representing the inflectional ending is underlyingly a /d/. However, there are certain endings that are /t/, and those are stable. We find them in the (preterit/subjunctive) forms of modals: (*könnte+er* /k9n**t**@ ?E6/ → /kEn.**t**6/ '*could he*'). The lesson we should learn from these data is that it is more than just useful to encode also certain aspects of morphological structure [4].

A case where Austrian German differs from the German standard is the palatal velar fricative /C/: orthographically "*ch*", in onset position, and as part of the derivative ending '*–ig*' it is pronounced /C/ in the German standard, but as /k/ in Austrian German. (The ending '*–ig*' is /IC/ again in dialectal varieties.) Here, we have to resort to orthographic information, since other contexts where the string /IC/ occurs (but written as '*–ich*') do not take part in that shift.

Much more problematic are cases where phones are not neutralized, but split up into two (or more) different correspondents. The vowel /a/ from the standard language is realized by default as rounded ([ɔ/ü]) in dialectal varieties. However, most (but not all) foreign and loan words, and some indigenous ones retain the /a/. The diphthong /aI/ shows up as a monophthong in the Viennese variety (3: [æː]), but

[2] Phonetic transcriptions are given in German-SAMPA notation, slightly adapted for the Austrian varieties. Wherever it deviates from the standard, IPA symbols are given in square brackets.

**Table 2.** Phone correspondences for /a/ and /aI/

| word | gloss | German standard | | Viennese | |
|---|---|---|---|---|---|
| *Pass* | passport | a | /pas:/ | A | /bAs:/ |
| *pass* | fit-IMPERATIVE | a | /pas:/. | a | /bas:/ |
| *zwei* | two | aI | /tsvaI/ | a: | /tsva:/ |
| *drei* | three | aI | /draI/ | 3: | /dr3:/ |
| *weil* | because | aI | /vaIl/ | &: | /v&:/ |

sometimes it is pronounced as an /a/, and in the context of l-vocalization the monophthong definitely is a different phone (& : [œ:]) which is not present in the German varieties. Examples are given below in the following table.

It is not possible to deal with such ambiguity in an automatic way, the only thing one can do is to find out which of them constitutes the more marked case (e.g., /aI/ → /a/), or whether there is a correlation between different strata of a language (foreign words). In the latter case an additional problem is whether these strata can be identified automatically, we suggest that in the case discussed it is possible only to a certain extent. At the present stage, we employ a semi-automatic procedure – automatic transformations with manual corrections.

## 4 Phone String Rules

There are certain rules that do not refer to correspondences between sets of phone-symbols, but reflect phonological processes that are sensitive to local context and/or syllable structure. They are normally easy to capture and we formalize them as regular expressions over strings of phones. There are two rules of that type that capture the differences between the German standard and Austrian German: r-vocalization and syllabic nasals (with assimilation of the nasal to the preceding onset consonant). In Table 3 there are some examples illustrating these rules.

**Table 3.** /r/-vocalization and syllabic nasals in Austrian German

| word | gloss | German standard | Austrian German |
|---|---|---|---|
| *Lehrer* | teacher | le:.r@r | lE:6.r6 |
| *werben* | solicit | vEr.b@n | vE6.bm= |
| *mehrere* | several | me:.r@.r@ | mE:6.r6.r@ |
| *beruhigen* | calm down | b@.ru:.I.g@n | b@.ru:.I.GN= |
| *Barbar* | Barbarian | bar.'ba:r | ba.'ba: |

In principle, these differences are already captured by the speech data, and the local context should be precise enough to identify the correct units. However, these rules also involve deletion and insertion within the string of phones, so the results of automatic alignment significantly improve if those rules are reflected in the lexicon that applies to the speech data. Notice that the process of r-vocalization is not solely determined by the local context but also by morphological information that determines the phonological domain where this process may apply. In the case of '*beruhigen*' r-vocalization is blocked although the /r/ follows a vowel (schwa), because that vowel is part of a prefix, hence invisible for the process.

**Table 4.** /l/-vocalization in Viennese varieties

| word | gloss | German standard | Viennese |
|---|---|---|---|
| *Wald* | forest | va**l**d | vA**I**d |
| *Holz* | wood | hO**l**ts | ho**I**ts |
| *voller* | full with | fO.l:6 | fo.l:6 |
| *Walter* | Walter | val.t6 | val.t6 |
| *viel* | much | f**i:l** | f**y:** |
| *viele* | many | f**i:**.l@ | f**y:**.l@ |

In Viennese varieties and many other Austrian dialects, another process is at work that has similar effects in terms of structure modification: l-vocalization. If the preceding vowel is a front vowel, /l/ affects the quality of this vowel (+round) and deletes if it is not in an onset position of a syllable with a phonetically realized vowel (pace syllabic nasals). This is how the phones +round/+front (i.e., /y/ and /2,9/) emerge again in the phone inventory, since the corresponding phones of the standard varieties are unrounded in the dialectal varieties (e.g., *Füße* /f**y**:.s:**@**/ → /f**I6**s:/ '*feet*', but *fühlt* /f**y**:l/ → /f**y:**/ '*feel*'). After non-front vowels and in non-onset position /l/ reduces to a front glide, phonetically speaking it forms a diphthong (e.g., /oI/) with the preceding vowel. Just in the case of /a/ (which is the non-default correspondent to /a/ in the standard variety) it remains unaffected.

The implementation of this rule is rather straightforward, and, as indicated above, it improves the quality of automatic segmentation as well as the accuracy of unit selection.

It has been mentioned already that if certain processes also affect the phonological structure of lexical items (deletions and insertions) then a thorough remodeling of the whole phonetic lexicon is advisable, at least in order to guarantee the desired quality of automatic alignment of the speech database for speech synthesis. We have shown that many transformations can be formalized easily – provided the resources of the standard variety are rich enough in information (morphological information, word stress assignment etc.) a new version of the pronunciation dictionary can be generated automatically and only ambiguous cases have to be checked manually. One way to tackle the problem of ambiguous transformations would be to employ decision-making methods in the stage of automatic alignment. This, however, would only resolve items that are present in the recorded speech data.

What is striking is the amount of linguistic expert knowledge that is incorporated in the establishment of such a transformation component as opposed to machine learning techniques, such as existing letter-to-sound conversion methods. At present, it was easier to follow this path, but we suspect that in the future it will be possible to establish methods that incorporate both, some (much more abstract) linguistic knowledge about phonological systems and processes, and a statistically based component that automatically extracts the rule module. This would be a great achievement towards the applicability of such methods independently of specific languages or varieties.

We have deliberately shied away from commenting on a severe problem that immediately arises if we proceed higher in the hierarchy of linguistic modules

(morphology, lexicon, phrases): orthographic encoding of dialectal variants. The problem arises in the moment when words have to enter the lexicon that do not exist at all in the standard variety. Dialectal varieties in most cases have no conventionalized system of orthographic transcription, if there exist some, there exist many of them which oscillate between phonetic accuracy and similarity to the standard orthography, and in many cases, the available resources suffer from low consistency. In order to permit dialectal input for a TTS system, this problem has to be dealt with by some means or other. At present, we cannot report any results yet, but only give an outline of the strategy we want to adopt: Provided we know the processes at work that transform phonetic encodings of one language variant to the other – a either by linguistic analysis or by application of machine learning techniques – we can generate a list of potential candidates both in the standard lexicon and the dialect lexicon by applying the rule module in both directions, which potentially has multiple outputs. The selection of the hopefully right candidate must be triggered by factors such as which rules have applied in which order, word frequency, but perhaps also statistic language modeling can help to improve this task.

## 5 Conclusion

In this paper, we have described certain strategies one may adopt in order to establish linguistic resources for (dialectal/sociolectal) language varieties from available sources representing a given standard, in our case Austrian varieties from common resources for the German language. We concentrated on the adaptation and (re-) modeling on the level of phone strings, leading to the question how accurate the phonetic encoding of a given variety must be. Certain differences between the standard language and its varieties either involve ambiguous or non-predictable variants or, if they are regular, they alter the phone string in a way that would negatively affect the quality of automatic procedures such as segmentation or unit selection. The conclusion was to re-model the variety as a whole, exploiting the regularities as much as possible. Future research will have to focus on the question, whether non-regular or ambiguous derivations can also be resolved by statistically based methods, such as decision making during alignment, or with the help of machine-learning techniques.

## Acknowledgements

This paper reflects work done within the project “Vienna Sociolect and Dialect Synthesis” (VSDS), supported by the Vienna Science and Technology Fund (WWTF). The Telecommunications Research Center Vienna (ftw.) is supported by the Austrian Government and by the City of Vienna within the competence center program COMET. OFAI is supported by the Austrian Federal Ministry for Transport, Innovation and Technology and by the Austrian Federal Ministry for Science and Research.

## References

1. Black, A.W., Lenzo, K.A.: Building synthetic voices, http://festvox.org/festvox/festvox_toc.html
2. Muhr, R., Schrott, R.: Österreichisches Deutsch und andere nationale Varietäten plurizentrischer Sprachen in Europa. öbv&hpt, Wien (1997)
3. Moosmüller, S.: Die österreichische Variante der Standardaussprache. In: Morgan, E.M., Püschel, U. (eds.) Beiträge zur deutschen Standardaussprache, Werner Dausien, Hanau, Halle, pp. 204–214 (1996)
4. Fitt, S., Richmond, K.: Redundancy and productivity in the speech technology lexicon – can we do better? In: Proceedings of Interspeech (2006)