# A corpus of regional Dutch speech

*John Nerbonne, Sandrien van Ommen, Charlotte Gooskens, Leen Impe & Sebastian Kürschner*

**Abstract**[1]

We present in this paper a phonetically transcribed corpus of regional Dutch speech from the Netherlands and Belgian Flanders and some example analyses using it. The corpus consists not only of 200 common words, but also of 200 nonsense words as these were pronounced by radio announcers from regional radio stations. The announcers regularly speak on radio programs with an explicitly regional remit, and they agreed to our recording them using regional speech but also using the standard language (standard Netherlandic Dutch in the Netherlands and standard Belgian Dutch in Flanders), which they also spoke professionally. The corpus is publicly available for analyses concerning, e.g., the relative differences between Dutch and Belgian regional speech, the relative similarity of regional speech to standard speech in the two countries, and the probity of models such as Auer's & Hinskens' (1996) "cone model", within which regional speech should be found. Nerbonne, Van Ommen, Wieling & Gooskens (to appear) have indeed used the corpus material to test whether regional speech adheres to the predictions of the cone model.

## 1. Introduction

### *1.1 The new dialectic between standard languages and dialects*

Many, indeed most, European languages gave rise to standard languages during and following the Renaissance, and those standard languages were the medium of communication for printed material, which became available in ever growing quantities. For several centuries standard languages led lives of peaceful coexistence with local varieties – DIALECTS – each limited to certain spheres of interaction, but the two sorts of varieties – standard and dialect – might be indifferent to each other. Linguists have used the term DIGLOSSIA to describe situations where two languages (or two varieties of the same language) are used through a population, and where the situation determines which language is used (Ferguson 1959).

During the past century the relative independence of dialects and standard languages has given way to a situation in which researchers suspect massive influence of standard languages on local dialects due to the

1 We are pleased to acknowledge NWO and FWO's support within the VNC-programme (Belgian Dutch Netherlandic Committee of Dutch language and culture). Principle investigators: Dirk Geeraerts, Roeland van Hout and John Nerbonne.

growing dominance of the standard varieties. There are several reasons for the growing importance of standard languages. The mobility of language users has increased a great deal, meaning not just short-term mobility due to tourism or commerce, but the long-term mobility due to people moving house from one part of a country to another. Compulsory education, normally involving the standard language, or some (mildly) accented version of it, has become the norm all over Europe. Finally, radio and television have become pervasive and were for many years broadcast exclusively in the standard language (and indeed, not all countries broadcast in non-standard varieties).

The reasoning behind the sociolinguistic suspicion that standard languages are probably influencing the dialects seems unassailable, and we shall present the dominant model of how dialects and standards interact in Section 3 (below). As empirical scientists we wish to examine regional speech in detail and test hypotheses about the putative influence of the standard empirically. This was the motivation for undertaking the research that resulted in our compiling the corpus.

### *1.2 The project*

The corpus of standard and regional Dutch language material consists of transcriptions of isolated words, pronounced in the Belgian and Netherlandic Dutch standard, as well as in eight regional varieties from both the Netherlands and Flanders. These samples of regional speech are regionally accented, are intended to be comprehensible in an entire region (and not just in a single village), and are thus differentiated from the formal standard and the base dialects. The current paper motivates the collection of the corpus material, provides a description of it, sketches some of the phonetic relations among the different language varieties in Belgian and Netherlandic Dutch, and suggests directions for further research.

The material presented in this paper was gathered for the project 'The mutual intelligibility of language varieties in the Low Countries'. The recordings (without the transcriptions) have already been presented in part in Impe, Geeraerts & Speelman (2008) and in Impe (2010), and the corpus of transcriptions (in X-SAMPA IPA) is made freely available at www.let.rug.nl/nerbonne/papers (search for 'A corpus of regional Dutch speech').

The intelligibility project required that a large number of words common to the Dutch of the Netherlands and the Dutch of Belgium be pronounced and transcribed as they would occur in the standard and in

regional speech. This was necessary because the project wished to compare different factors which might contribute to the (lack of) mutual intelligibility of Dutch language varieties in the Low Countries (i.e. Flanders and the Netherlands). The different factors under investigation were e.g. attitudinal determinants, familiarity and linguistic differences (Impe 2010). In the current paper we focus on the last potential determinant, as measuring pronunciation differences requires that material be available in phonetic transcription. Studies on Scandinavian languages have shown that some linguistic differences between closely related language varieties correlate strongly with mutual intelligibility (Gooskens 2007, Gooskens, Heeringa & Beijering 2009). The term LINGUISTIC DISTANCE indicates a measurable difference between language varieties, and, in intelligibility research, it often refers to the phonetic and/or lexical differences among them. Other linguistic differences, such as syntax or morphology, have not yet been proven to significantly influence mutual intelligibility in related varieties. Of all linguistic factors the phonetic distance, e.g. measured using the Levenshtein distance (see Nerbonne & Heeringa 2010, Heeringa 2004), has been found to correlate most strongly with mutual intelligibility (Gooskens 2007).[2] In Section 4 (below) we sketch the result of measuring phonetic distances between several Dutch language varieties. In the future we plan to compare mutual intelligibility scores of subjects from the same regions the material originates from.

Although the project that funded the data collection and corpus creation sought to investigate whether the phonetic distances between Netherlandic and Belgian Dutch language varieties are indeed a determinant of intelligibility, the data also provides insight into the phonetic relations between regional speech and standard varieties in the Low Countries as well as insight into the differences between the Belgian and Netherlandic Dutch language multidimensional speech continuum. To illustrate what we mean by this we introduce Auer's & Hinskens' (1996) conical model, which is presented in Figure 1.

### *1.3 Related work*

Auer and Hinskens (1996) elaborate a *cone model* to illustrate the modern dynamic between base dialects and standard languages. There is a base of local dialects, some of which have existed from centuries, as well as an

2 A weaker correlation has been found for lexical distance, whereas a direct relationship between extra-linguistic factors such as contact and attitude and mutual intelligibility (see Gooskens (2006) for a discussion) has been difficult to prove.

apex, representing the standard language. Some speakers address an audience larger than their local village, but cannot or will not use the standard. It may be difficult for them to use the standard, but they may also wish to display loyalty to their region. Such speakers naturally adopt an intermediate form of speech.

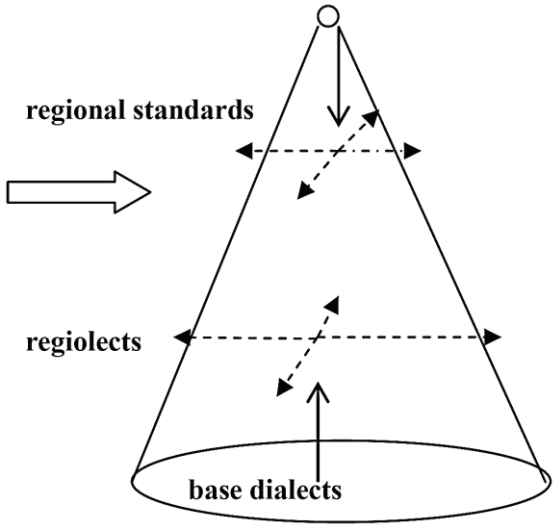



*Figure 1: Auer's (2005) cone-shaped speech continuum, with base-dialects, regiolects, regional standards and spoken/written standards. Advergence of base dialects to each other and the standard leads to intermediate, regional, varieties.*

Figure 1 differentiates between horizontal and vertical relations: horizontal relations are e.g. relations among different geographical dialects, whereas vertical relations are relations among types of speech, e.g. standard, regional and dialectal speech. The diagonal arrows in the model depict the idea that regional speech may vary not only on a horizontal level, but also on a vertical level. The arrows pointing towards the "regiolects",[3] originating from the standard and the base dialects, symbolize the convergence of varieties, the influence of the standard on regional speech which we discussed in the introduction. Although the peak of the cone might

3 The term 'regiolect' is theory laden and might be restricted to types of speech that conform to the cone model. If we used the term to refer to our samples, we would seem to assume that the samples conform to the cone model, exactly what we test in Section 4.1, which would be a straightforward case of the petititio principia fallacy, something which would dismay Prof. Van Heuven. So we shall avoid referring to our samples as samples of 'regiolects' here.

suggest that standard varieties themselves admit no variation, this is of course not the case, as even carefully produced standard speech often contains traces of regional "color". Since we are interested in the functions of standard speech in regions with differing dialects, we deliberately collected samples of standard speech from speakers of different regions (see below).

We agree that this cone is well motivated in situations such as the one in the modern Netherlands and Belgium, where the vast majority of dialect speakers also use standard Dutch regularly, albeit with different degrees of proficiency. It is only natural to see intermediate speech forms arise where some speakers are motivated to sound regional but nonetheless remain comprehensible to a larger group. But we also note that it is difficult to remain within the cone in a natural way. We return to this in the conclusion.

### *1.4 Belgian and Netherlandic Dutch*

Dutch is particularly complex in its interactions between "standard" and local languages because in fact there is not one, but two standards, one for each national community.

Dutch has been described as a pluricentric language (Geerts 1992, Deprez 1997). The formal (written) standard does not differ much between Belgian and Netherlandic Dutch, but the spoken standards have phonetically diverged (Van de Velde 1996), resulting in two separate (but closely related) standard varieties. Figure 2 is a schematic representation of such a situation.

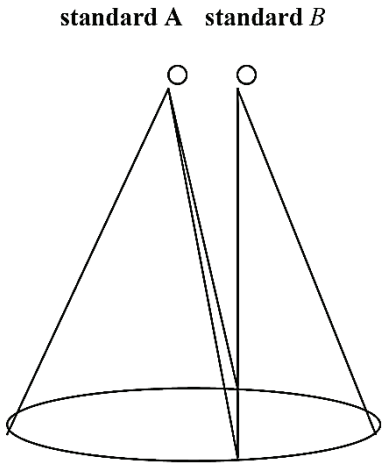



*Figure 2: Auer's (2005) model of a two-standard diaglossic situation, where separate standards have evolved.*

Both in the Netherlands and in Flanders, standardization is taking place, leading to a loss of (base) dialects and shift in the function of these dialects as markers of local identity, i.e., as allegiance to a particular region. As, e.g., Grondelaers, Van Aken, Speelman & Geeraerts (2001) note, the standardization in Belgian Dutch set in at a much later stage than in most other European languages (French was the language of the government and administration until 1898), and Belgian standardization has not yet been completed according to many linguists. The Belgian situation is complicated further by the existence of a less formal spoken standard together with a more formal one used both in speech and in writing. This less formal standard is used only in speech, never in writing and is referred to variously as COLLOQUIAL BELGIAN DUTCH (CBD), or *tussentaal* 'in-between language' (Taeldeman 1993) and otherwise (Impe 2010: 27).

### *1.5 This study*

In this study we shall seek empirical information about regional speech that might illuminate some of the issues above. In particular we shall address the following research questions.

(1) a. Are the various forms of regional speech phonetically equally similar to the standard variety?
    b. Does a regional speech form's similarity to the standard correlate with the region's social prestige?

(2) a. Are phonetic distances between the regional speech of the different Belgian regions and the Belgian standard larger than the corresponding distances in the Netherlands?
    b. Is regional Belgian speech more varied than regional Netherlandic speech? Are phonetic distances in the Belgian speech from different regions larger than the distances among the Netherlandic Dutch regions?

As we shall need a corpus of regional speech to do this, we shall also take care that the corpus is rich enough to support other investigations as well. Section 2 presents the corpus material, Section 3 the data analysis technique, and Section 4 examines four issues concerning regional speech using this corpus. We discuss our results in a concluding Section 5.

## 2. Material

### *2.1 Regions*

Ten varieties of the Dutch language, namely the Netherlandic and Belgian Dutch standard varieties as well as four regional varieties in both countries, were selected for the corpus. In each country, one of the regional varieties was selected from a central area, two from peripheral areas and one from an intermediate area (cf. Figure 3).

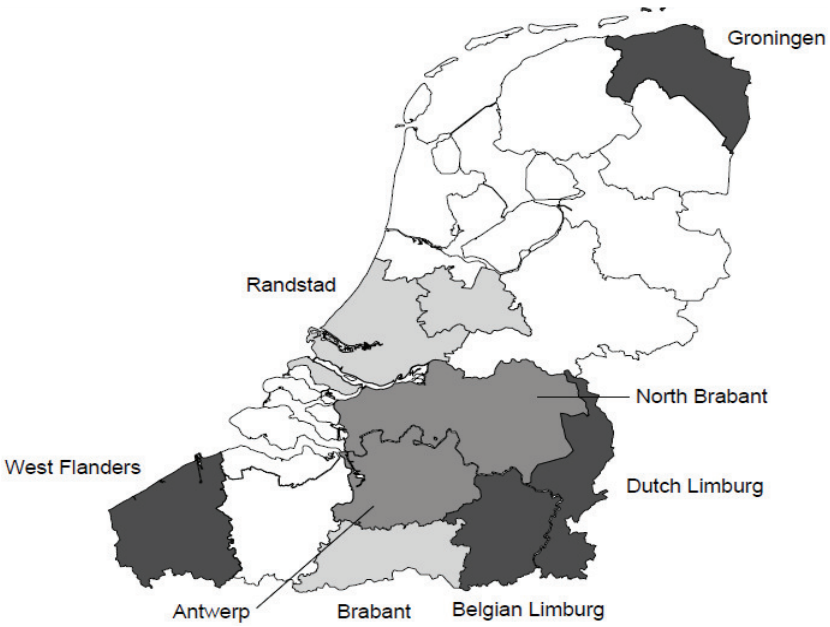



*Figure 3: Map of Flanders and the Netherlands*

The areas chosen differ with respect to their political and economic importance in their respective countries. The regions Brabant and Randstad are the most central areas, e.g., containing the capital cities[4] in Flanders and the Netherlands, respectively. Besides the regions' economic and cultural importance both regions have dominant positions in the media in their respective countries. The regions West Flanders, Belgian Limburg, Groningen and Dutch Limburg, on the other hand, are peripheral areas, where dialectal language use is allegedly better preserved than in the other areas. The regions Antwerp and North Brabant are considered intermediate areas: they are closer to the central region than the peripheral areas.

4 The working assumption is that the prestige of a region increases when the capital city of a country is situated in or near the region.

*2.2 Items*

The corpus consists of different pronunciations of a list of 200 words and 200 non-words. Half of the existing words are bi-national words, i.e. words that are spoken frequently both in the Netherlands and in Flanders. The other half consists of national words, of which 50 are typically Netherlandic Dutch and 50 are typically Belgian Dutch.
Table 1 contains some examples of bi-national and national words. See the appendix for a complete list. The typicality of the national words was checked using a Stable Lexical Marker Analysis (Speelman, Grondelaers & Geeraerts 2008), a statistical test that selects typical elements in a corpus. This test was done on two large corpora (viz. an online football forum corpus and the corpus of spoken Dutch (CGN)). To check if the words are familiar to language users, a pilot study was conducted with 50 subjects, confirming the classification (Impe 2010). The selected words are suitable for regional pronunciation, as they contain sounds speakers can produce with typical regional cues. These sounds are e.g. short vowels (for which openness and advancement are known to differ between varieties), long vowels (some of which in standard Netherlandic Dutch and some Belgian Dutch varieties are pronounced as diphthongs), diphthongs (which are pronounced as monophthongs in some varieties), trills and fricatives (differing in place of articulation and voicing), and the [ən]-endings of verbs (in which either [ə], [n] or both are apocoped while preserving syllabicity, or both are deleted).

*Table 1: Bi-national (grey) and typically Belgian and Netherlandic Dutch words (white)*

| Belgian Dutch | Netherlandic Dutch | translation |
|---|---|---|
| boom | boom | tree |
| verhaal | verhaal | story |
| aandacht | aandacht | attention |
| kuisen | schoonmaken | to clean |
| ambetant | vervelend | annoying |
| verschieten | schrikken | to be scared/frightened |

We include not only existing words but also made-up words in our data collection in order to facilitate comparisons involving the importance of

lexical recognition for perceptions of differences. The made-up material cannot be recognized lexically, so perceptions of its unnaturalness or "foreignness" cannot depend on its being recognized. The non-words are based on a set of 20 existing, bi-national, words. All non-words are constructed by making a small number of changes to the original word, rendering a word that is either phonotactically, morphotactically or semantically plausible in Dutch. A complete list can be found in the appendix of this paper (below).

Examples of phonotactically plausible non-words are: *sleem, mils* and *bafoor*. They are strings of sounds that sound like a Dutch word but do not consist of morphemes carrying meaning. Since some morphotactically plausible non-words consist of a phonotactically plausible syllable, combined with a Dutch morpheme, one could say they are inflected non-words. Examples are *hoelig* (this might be seen to correspond with the English non-word 'hul-ish') and *deparatie* (corresponding to the English non-word 'deparation').

Semantically plausible non-words are strings deriving from two existing lexical morphemes which in combination do not form existing words. Examples of semantically plausible words: *vaasgeur* ('vase smell'), *bosknecht* ('forest servant'), *rookbeek* ('smoke creek').

We verified that all non-words indeed did not exist, using both dictionaries and the internet, and using a pilot study. Subjects were asked to fill in a 7-point scale on how certain they were whether a target was or was not an existing word (Impe 2010).

### *2.3 Recordings*

The eight different regional varieties are obtained by having the 400-item list pronounced by one professional speaker from each of the different regions and recorded in sound-proof radio studios with high-quality audio equipment. The eight speakers were all male radio announcers/commentators on regional radio stations. They were all accustomed to using both regional speech and standard Dutch in their work as radio announcers, and it was unproblematic for them to switch from regional to standard speech. They were all born and raised in the region they worked in at the time of this study, i.e. the area in which they used regional speech professionally, and their age was between 27 and 34 years. The speakers were selected for voice quality to control for individual speaker variation as much as possible. Recognizing the existence of speech with a regional accent does not imply that a regional variety is a consistent and

distinguishable linguistic system (Auer 2005). How distant a regional accent is from the standard language may differ among speakers. In the instructions, the speakers were asked to use "informal regionally accented speech" comprehensible in the speaker's entire region. The speaker was thus referred to the situational context with respect to regional markers, as speakers adapt the degree of regionality of their speech to the communicative situation (Auer 2005). In this way the regional variety is differentiated from both the formal standard and the local dialect. The speakers use regional speech on a daily basis for professional purposes, which means they have a delineated concept of regionally accented speech. Even though one speaker per region does not allow us to draw inferences on specific regiolects as varieties (which, as noted above, is already problematic), it does allow us to investigate the general positioning of regional speech in the sample of eight cases. Each of the speakers pronounced all 400 items in their own regional variety. Furthermore, each speaker pronounced 50 words and 50 non-words from the list in their standard language (i.e. either standard Netherlandic Dutch or standard Belgian Dutch), resulting in the total of 400 "standard-language" pronunciations. We asked the speakers to use the standard language in eliciting the standard pronunciation, meaning the sort of speech they would use professionally (as radio announcers) when they were not working in specifically regional broadcasts. The distinction between standard speech and *tussentaal* (CBD) was not mentioned. The recordings referred to as standard language thus contain a mixture of four different speakers, of which all speakers pronounced 100 different items. We deliberately collected samples of "standard speech" from all the radio announcers in order not to compare regional speech to a single version of the standard, but to that version of the standard which would be heard locally in a given region. As we shall see below (Figure 4) the local standards differed among themselves.

## 3. Method

### *3.1 Transcriptions*

The fact that the same words are pronounced in each variety makes it possible to compare the language varieties using the Levenshtein distance. The Levenshtein distance (see Nerbonne & Heeringa (2010) for further elaboration) is a string distance measure that we shall use to compute phonetic distances between all pairs of the ten language varieties. This dis-

tance can be based on various comparisons (even using spectrograms), but in this case the comparison will be made using phonetic transcriptions. Levenshtein distance has been shown to correlate strongly with human judgments of pronunciation differences (cf. Heeringa 2004, Gooskens & Heeringa 2004) and with behavioral measures of intelligibility (cf. e.g. Beijering, Gooskens & Heeringa 2008, Kürschner, Gooskens & Van Bezooijen 2008, Gooskens 2007). We used the web application **Gabmap** for our calculations (Nerbonne, Colen, Gooskens, Kleiweg & Leinonen 2011). We refer the reader to earlier publications for an explanation of how the measurement is calculated, its variations and properties and how its results may be analyzed and visualized (Nerbonne & Heeringa 2010).

The Levenshtein distance is calculated for each pair of words and normalized on the basis of the word lengths in a set that is compared. The cumulative distance between two varieties is the mean normalized distance of all corresponding word pairs from the two varieties in question.

Due to time constraints not all items could be transcribed. 300 out of 400 items were transcribed in XSAMPA, a computer-readable transcription standard based on the IPA (International Phonetic Alphabet). The 300 items consist of 200 words and 100 non-words in 10 varieties, making a total of 3000 transcriptions, consisting of 2000 words and 1000 non-words.

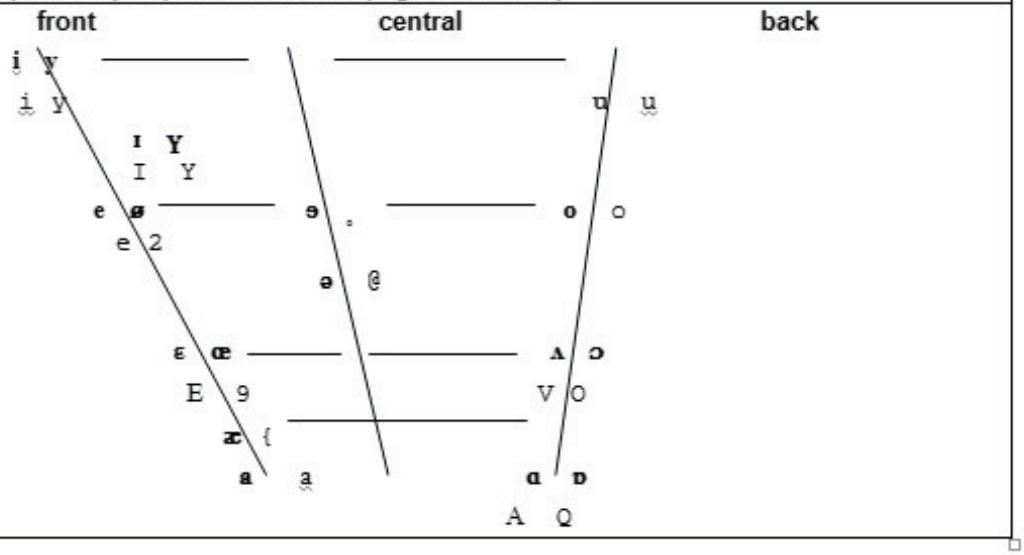
Transcriptions are influenced by the linguistic background of the transcriber, the amount of detail used and the choices transcribers make in "borderline" cases. To avoid such variation in the data induced by various transcribers, all transcriptions were made by one Netherlandic Dutch native speaker. The first 400 words were also individually transcribed by a second, non-native transcriber, after which the transcribers evaluated the transcriptions together and decided on the rules to be followed and symbols to be used during the task. The transcriptions of the Belgian Dutch varieties were checked by a Belgian Dutch native speaker

Only the transcriptions of the first transcriber (after the checks) are part of the material. As a reference system for the description of Dutch sounds, the sound system of the Goeman-Taeldeman-Van Reenen project (GTRP) was used, as this was applied in the Belgian part of that project and as described by Wieling, Heeringa & Nerbonne (2007). This was important as we also wished to compare the regional speech with base dialects from the region. See Table 2 for the symbols used in the current research.

*Table 2: All symbols used in the transcriptions for both Belgian and Netherlandic Dutch, given in IPA (above) and XSAMPA (below)*

| CONSONANTS (when in pairs, left is voiceless, right is voiced) | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | bilab | labdent | dental | alv | palv | retr | pal | vel | uv | phar | glot |
| plosive | p b<br>p b | | | t d<br>t d | | | | k<br>k | | | ʔ<br>? |
| nasal | m<br>m | ɱ<br>F | | n<br>n | | | | ŋ<br>N | | | |
| trill | | | | r<br>r | | | | | R<br>R\ | | |
| tap or flap | | | | ɾ<br>4 | | | | | | | |
| fricative | | f v<br>f v | | s z<br>s z | ʃ ʒ<br>S Z | | ɕ<br>C | X ɣ<br>x G | χ ʁ<br>X R | | |
| lat. fricative | | | | | | | | | | | |
| approx. | | ʋ<br>v\ | | ɹ<br>r\ | | | j<br>j | | | | |
| lat. approx. | | | | l<br>l | | | | | | | |
| other symbols: | velarized /l/ | | | | | | ɬ 15 | | | | |
| | labial-velar approximant | | | | | | W w | | | | |

VOWELS (when in pairs, left is unrounded, right is rounded)




A small set of diacritics was used to describe stress-assignment, devoicing, nasalization, syllabification and aspiration. The diacritics used in our transcriptions are given in Table 3.

*Table 3: Diacritics and suprasegmentals used in transcriptions*

| IPA | XSAMPA | Meaning |
|---|---|---|
| ◌̥ | _0 | Voiceless |
| ◌ʰ | _h | Aspirated |
| ◌̩ | = | Syllabic |
| ◌̃ | ~ | Nasal |
| ː | : | Long |
| . | . | Syllable-break |
| ˈ | " | Primary stress |
| ˌ | % | Secondary stress |

## 4. Research questions

The corpus is suitable for addressing a number of questions concerning the relation of regional speech to standard varieties on the one hand and to the region's base dialects on the other.

### *4.1 Regional differences in speech*

As Auer (2005) notes, the space between standard(s) and dialects can be seen as a continuum, where regional speech does not necessarily constitute a separate variety, but is perhaps better seen as a by-product of base-dialect leveling resulting in a kind of convergence toward the standard. The degree of convergence toward the standard may thus vary among different sorts of regional speech. It would be conceivable that all the instances of regional speech might take a common position in the continuum, at a common horizontal level, so that the various common forms of regional speech are equidistant to the standard. But regional speech patterns might vary and scatter to different position in the continuum, so that some are closer to the standard than others. As Auer (2005) and Videnov (1999) further note, the social prestige of a speech group influences the effect this group has on various varieties, including the standard language. Following this reasoning, we might expect that the higher the prestige of a region, the closer the speech variety of this region will be to the standard language. This proposed relation leads to two further research questions that the current corpus equips us to ask, as noted in Section 1 and repeated here for convenience.

(1) a. Are the various forms of regional speech phonetically equally similar to the standard variety?
b. Does a regional speech form's similarity to the standard correlate with social prestige?

We hypothesize that regional speech is not all equally close to the standard variety, but rather that the proximity to the standard correlates with the prestige of the region. We therefore conjecture that the answer is negative to the first question and positive to the second.

To answer these questions, a graph is given in Figure 4, where the mean distance of each sample of regional speech to its standard variety is given. The varieties are ordered by social prestige of the region (more peripheral regions on the side, more 'central' regions in the middle).

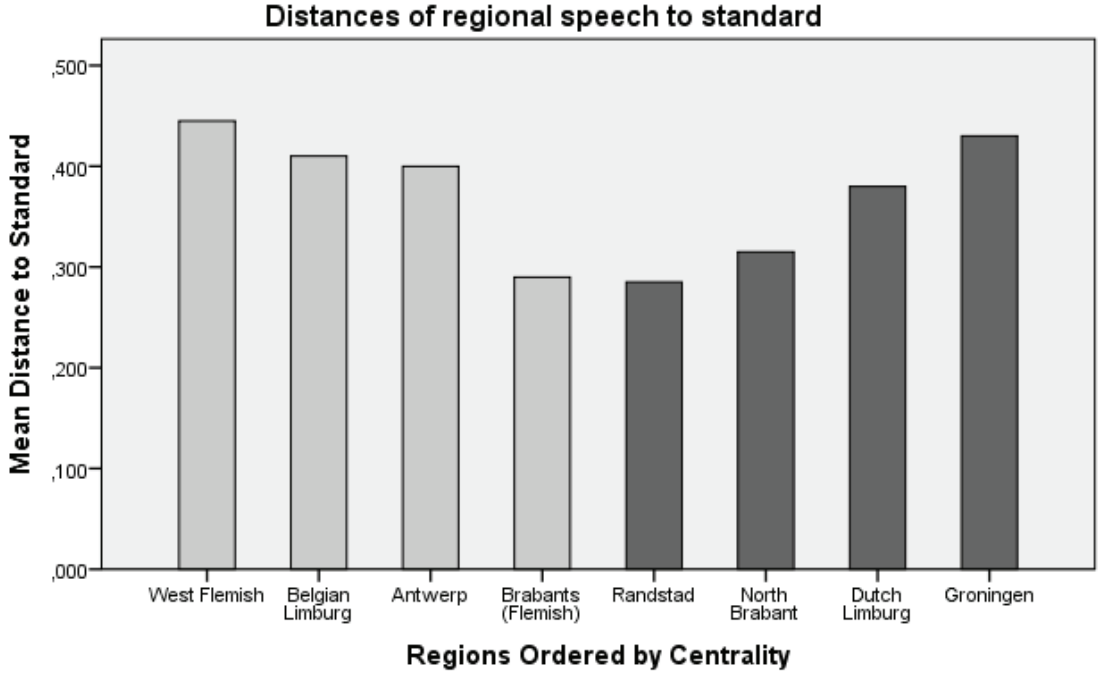



*Figure 4: Mean distances of regional speech from the local standard. On the left (lighter bars): mean distances of Belgian Dutch regional speech samples to standard Belgian Dutch. On the right (darker bars): mean distance of Netherlandic Dutch regional speech to standard Netherlandic Dutch. Central varieties (regions containing the capital city) are in the middle, peripheral varieties on the side.*

As can be seen in Figure 4, the distances between word pronunciations in regional speech and their respective standard pronunciations are not equal, confirming our negative hypothesis with respect to the first research question in this section. Regional pronunciations in central areas

are closer to standard pronunciations than regional pronunciations in more peripheral areas. A one-way ANOVA shows that the mean distance of regional to standard pronunciations differs significantly among regions ($F(2382.7) = 36.7$, $p < 0.05$).[5] A closer look at the various regional-standard distances (using the Bonferroni post hoc test) shows that in Belgian Dutch all differences are significant, except for the difference between Belgian Limburg and Antwerp (in the graph, too, the difference is visually small). In Netherlandic Dutch, Groningen differs significantly from all other varieties with respect to its distance to the standard, and Netherlandic Limburg differs significantly from Groningen and Randstad, but the differences between Randstad and North Brabant on the one hand and North Brabant and Netherlandic Limburg are not significant.

The regional speech of the various regions is therefore not equidistant from the standard. Referring back to the cone model, this means that regional speech does not occupy a horizontal plane in Auer and Hinskens' cone (at equal distances to the standards) but rather differ vertically. At the same time, the distances to the standard do follow the relative social importance of the regions, confirming our hypothesis with respect to the second research question in this section.

### *4.2 Belgium-Netherlands differences*

In the graph in Section 4.1 (above), the distances between regional pronunciations and standard pronunciations are slightly smaller in Netherlandic Dutch than they are in Belgian Dutch. This was indeed to be expected as a consequence of the late onset of standardization in Belgian Dutch (discussed above). Because the Belgian standard arose comparatively late, the relative differences between regionally accented speech and the standard variety are relatively large when compared to the differences between standard and regionally accented speech in Netherlandic Dutch (Grondelaers et al. 2001). Grondelaers et al.'s results are based on a lexical comparison. The current corpus, based on phonetic differences, may shed new light on this matter.

5 We should note that we measured the pairwise differences on the sets of words from the standard on the one hand and the regional speech on the other. If we had measured the mean differences between the standards and the regional speech samples, then we would have compared four mean differences from Belgium to four mean differences from the Netherlands, too few for a statistically meaningful comparison.

In Figure 5 the hypotheses regarding these questions are translated to the conical model. The larger area of variety within the speech continuum of regional Belgian Dutch implies that distances between standard and regional varieties are larger, resulting in a greater vertical variety. Even though there is no certainty about the organization of regional speech in the model (either horizontally or vertically), the Belgian Dutch area has more room for variation: more vertical variety corresponds to more horizontal variety in regional speech, due to the conical shape of the regional speech space. This reasoning leads naturally to further questions:

(2) a. Are phonetic distances between the Belgian Dutch regional speech of the different regions and the Belgian standard larger than the corresponding distances in the Netherlands?

b. Is regional Belgian speech more varied than regional Netherlandic speech? Are phonetic distances in Belgian Dutch regional speech from different regions larger than the distances among the Netherlandic Dutch regions?

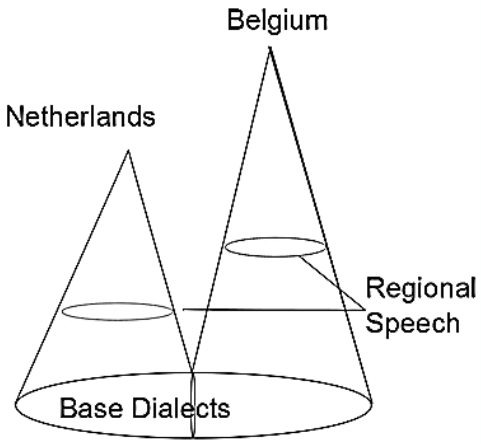



*Figure 5: A geometrical rendering of the hypothesis that Belgian regional speech will be less standard-like (than the regional speech in the Netherlands) due to the greater differences between the standard and the base dialects.*

We may answer the first question using a *t*-test, resulting in a small, but significant ($t(2388) = 4.446$, $p < 0.05$) difference between Belgian Dutch ($M = 0.404$, $SD=0.229$) and Netherlandic Dutch ($M = 0.362$, $SD = 0.226$), where m is the mean distance from regional speech to sample.[6]

6 See note 5 above.

The first question can thus be answered affirmatively: distances between the regional speech forms and the standard are indeed slightly smaller in Netherlandic Dutch than in Belgian Dutch. The difference is small, just 0,19 standard deviations (Cohen's d).

For the second question (does regional speech in Belgian Dutch differ more than it does in Netherlandic Dutch?) the regional speech forms of the one country should be compared to the regional speech forms of the other, so standard varieties are taken out of the equation. Figure 6 shows the difference between Belgian and Netherlandic Dutch. The figure shows the distances among the regional speech forms, ordered by centrality of all regions. On the left of the figure the distance of the peripheral variety to the intermediate peripheral variety is given for both Belgian and Netherlandic Dutch, while on the right of the figure the distance of the intermediate central variety to the most central variety is given.

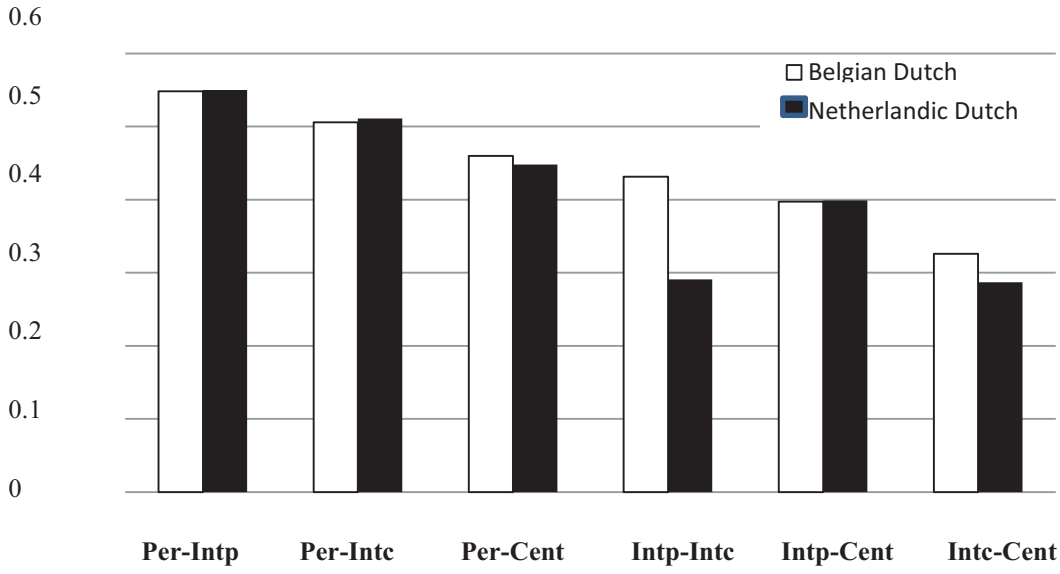



*Figure 6: Mean distances of each regiolect to all other regiolects in the same country. Per = peripheral, Intp = intermediate peripheral, Intc = intermediate central and Cent = Central*

Figure 6 suggests that Belgian Dutch and Netherlandic Dutch differ only slightly in the distances among regional speech forms. Almost all distances are the same, and only the intermediate central variety of Netherlandic Dutch (i.e. North Brabant) has a significantly smaller distance to the relevant intermediate peripheral variety and the central variety. Note that the distances between peripheral and intermediate peripheral

varieties are larger than between peripheral and central varieties. A *t*-test reveals that the mean difference in the distance between Belgian Dutch regional pronunciations ($M$ = 0.445, $SD$ = 0.227) and Netherlandic Dutch regional pronunciations ($M$ = 0.414, $SD$ = 0.242), although small (Cohen's $d$ = 0.14 $SD$), is significant ($t(3594) = 3.877$, $p < 0.05$),[7] but as we can see in the graph, this is due almost exclusively to the very standard-like North Brabant variety.

Since Belgian Dutch and Netherlandic Dutch differ significantly in regional-standard, but not in inter-regional distance we can conclude that the hypotheses presented above are only partly confirmed. The vertical area of phonetic variety in Belgian Dutch is slightly larger than that in Netherlandic Dutch, but no such conclusions can be drawn for the horizontal dimension of variety.

### *4.3 Words vs. non-words*

The data presented here and transcribed in the corpus are intended for use *inter alia* in psycholinguistic experiments, more specifically and most immediately in a lexical decision task, which means that we should verify that the phonetic relations between regional speech and standards are commensurable for words and non-words. Since we wish to use the non-words in a lexical decision task, we wish to verify that the subjects in lexical decision tasks are judging whether words are genuine in the absence of phonetic cues.

Figure 7 shows the mean distances to other varieties. The height of the line represents the mean distances of words (solid dark line) and non-words (dashed line) in this variety to all other language varieties (in Belgian and Netherlandic Dutch). We note that the lines representing words and non-words track each other quite well. Again, the varieties are ordered with Belgian Dutch varieties on the left and Netherlandic Dutch varieties on the right, so that the most peripheral varieties can be found on the sides and the most central varieties in the middle of the graph.

Do non-words created for use in this corpus differ from words in any way other than the fact that non-words do not carry meaning? The lines in the graph (words and non-words) deviate slightly from each other in some language varieties, such as Belgian Limburg, but overall, they follow the same line. A *t*-test confirms that, overall, words ($M$ = 0.431, SE = 0.238) and non-words ($M$ = 0.428, SE = 0.218) do not differ significantly ($t$ = 0.905, p

7 ee note five above.

>> 0.05). Why the two word types differ in their mean distance to other language varieties in Belgian Limburg cannot be answered using the data at hand. It is important to keep in mind, though, that each language variety is pronounced by only one speaker, and in this case one speaker shows different behavior from all other speakers. In general it can be concluded that in this corpus, words and non-words do not differ in their phonetic relations among language varieties. The non-words are therefore suitable material for use in lexical decision tasks that wish to concentrate on lexical identity only, and that wish to avoid bias via phonetic cues.

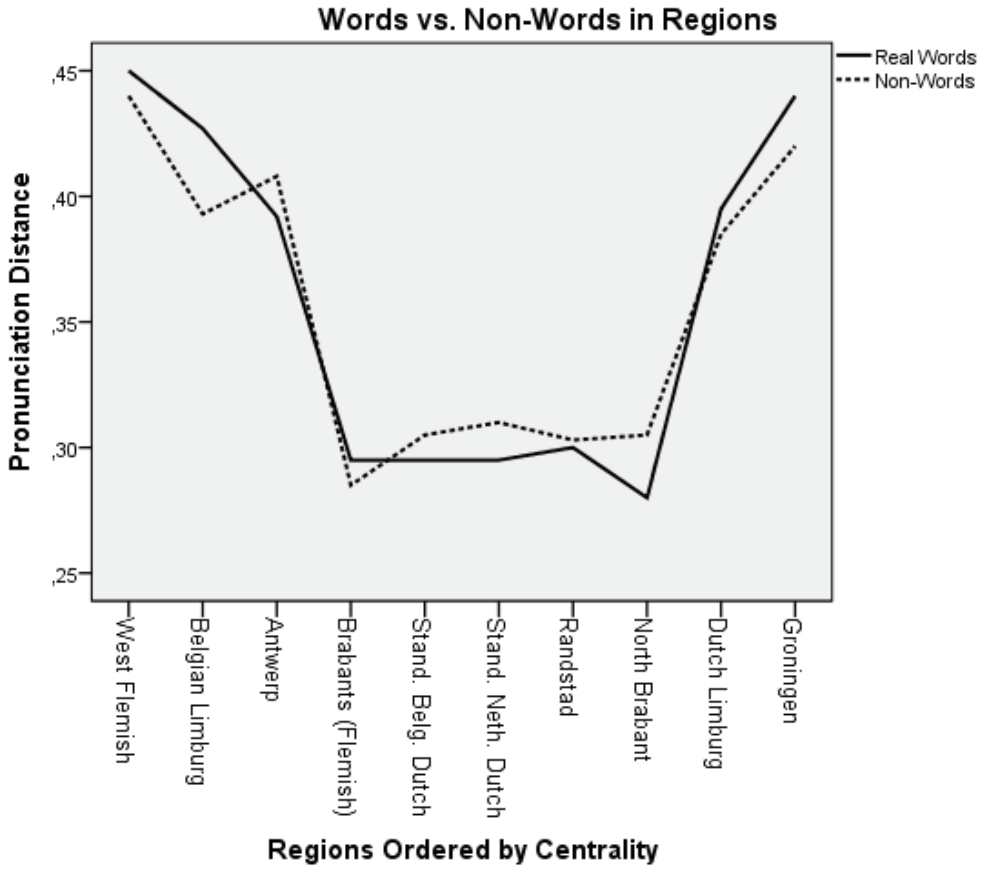



*Figure 7: Mean pronunciation distances of items in variety (on x-axis) to corresponding items in all other varieties for words and non-words. The least standard varieties are on the extreme left (Belgian Dutch) and extreme right (Netherlandic Dutch), the standard varieties are in the center.*

### *4.4 Regional speech and the Auer-Hinskens Cone*

A final question we wish to address concerns the probity of the Auer-Hinskens' cone model (see Figure 1, above). This section summarizes the

research in Nerbonne et al. (to appear), but we add methodological remarks and a discussion of statistical significance. We turn then to the research question: Does the speech of the various Dutch regional speakers conform to the Auer-Hinskens cone model?

The question is tricky to answer given the usual computational techniques available to dialectologists, and Nerbonne, Ommen, Wieling & Gooskens (to appear) do not discuss this methodologically. It is tricky to answer because although we can measure the pronunciation differences in the samples we collect, we obtain a distance from one variety to another, but not a direction (not a vector). So if we ask whether the regional speech is closer to the base dialects than the standard is, we effectively place regional speech within a circle whose center is the base of the cone (with the base dialects) and whose diameter is the distance from the base dialects to the standard. Figure 8 illustrates this (on the left). Fortunately, we can also ask whether the regional speech is closer to the standard than the base dialects are (Figure 8, on the right), and the conjunction of those two propositions requires that the regional speech be intermediate between base dialects and standards is (see Figure 8, center).

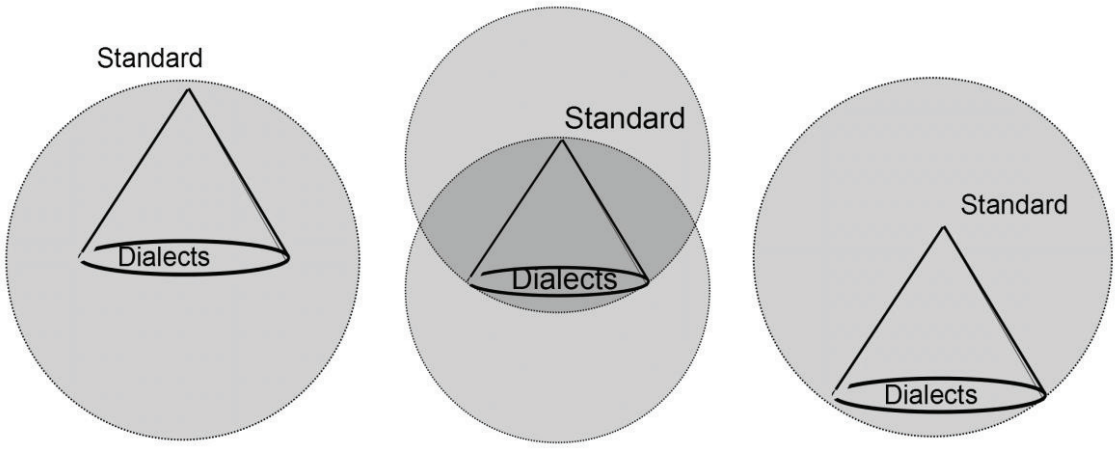



*Figure 8: Because our measurements are distances without direction, we test for the intermediacy of regional speech (between base dialects and standards) in two steps, first asking whether regional speech is closer to the base dialects than the standard is (left), where we have drawn a circle with the base dialects as center and the standard-base distance as radius. We then ask whether the regional speech is closer to the standard than the base dialects are (on the right), where we have drawn a circle of the same radius as on the left, with the apex of the cone as center. Both of these conditions must be met for the regional speech to be intermediate (center).*

As Nerbonne et al. (to appear) show, the various samples of regional speech in the Netherlands do not conform well to the cone model. In addition to the data introduced above, data from the Goeman-Taeldeman-Van Reenen project was used to represent base dialectal speech. More specifically, there were 37 overlapping words between the two data sets on which the measurements were based. Figure 9 depicts the results of our measurements. It turns out that by and large, our regional radio announcers speak in a manner that is different from the base dialects and the standard, but which is not properly intermediate. We drew the non-intermediate dialects outside the circle which they failed to appear in, but we caution that one should not interpret the vertical dimension too hastily. For example, Groningen falls outside the circle of speech forms which are closer to the standard than the base dialects are, so we placed it outside that circle, in fact a bit lower than the base dialects. Figure 9 may indeed suggest that the Groningen radio announcers spoke more "dialectally" than the base dialects speakers, but strictly speaking, we did not measure this, but only that their speech is more different from the standard than the base dialects are. We might have drawn Groningen above and further to the left to discourage the interpretation that it is more dialectal. We add, however, that the announcer may also be mixing non-standard elements from various base dialects and might, in this way, indeed speak more "dialectally" than the dialect speakers if he consistently favored non-standard elements from a range of base dialects, which however, are never found together in a single base dialect. Further analysis of the specific differences would be needed to be certain.

We speculated that the speakers may be performing in a regional manner, and that this sort of performance is difficult. We note that it also turned out that the announcers were able to discriminate among the varieties quite well, so that they consistently sound more like the base dialects of their own region than like the base dialects of any other region. Nerbonne et al. (to appear) do not assess the statistical significance of their results, but we add here that, if one examines the results from the perspective of a binomial distribution, and further suppose that the chance of a regional sample falling between the standard and the base dialects is *0.5* as a null hypothesis, then the chance of finding one or fewer instances falling within the interval in a sample of eight is statistically significant ($p < 0.05$).

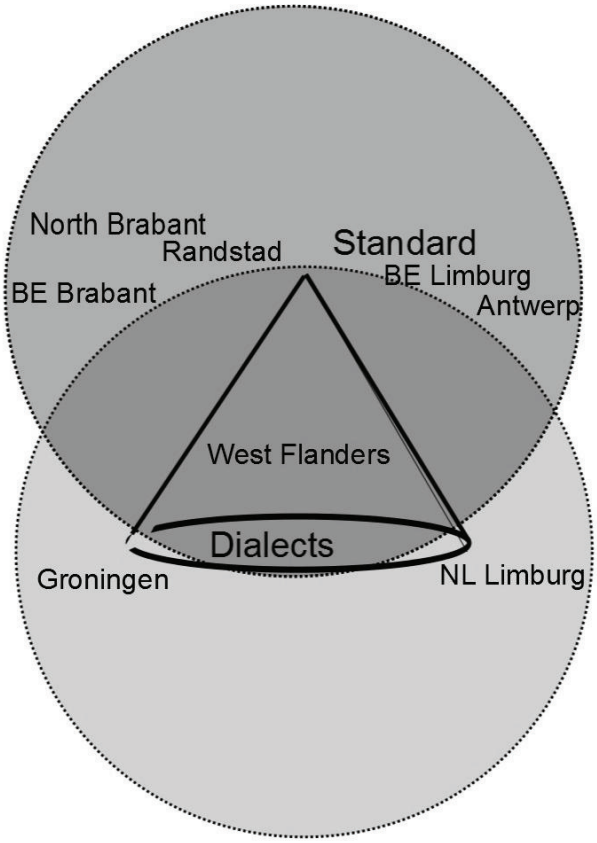



*Figure 9: Samples of regional speech with respect to the Auer-Hinskens regiolectal cone. Five of the eight samples were more different from the basilects of their region than the standard, and two differed more from the standard than the basilects did. Only West Flanders regional speech succeeded in striking a compromise between its standard language (Belgian Dutch) and its base dialects. From Nerbonne et al. (to appear). See text for further discussion.*

## 5. Conclusion

This paper describes the corpus now available at www.let.rug.nl/nerbonne/papers (search for 'A corpus of regional Dutch speech'). Research currently conducted on the data gives insight into the relation of regional to standard speech in Belgian and Netherlandic Dutch, and the corpus is suitable for further investigation as well.

In a first investigation, we showed that regional speech samples are not at equal distances from the standard. Rather, regional speech differs in its level of similarity to the standard, and the proximity to the standard is related to the general importance of a region in the nation. This relation suggests a bidirectional convergence of regional speech and standard: regional speech should move closer to the standard (standardization), and

the standard may assume some features of the regional speech of the more prestigious regions. In a second study we corroborated Grondelaers et al.'s (2001) finding that standardization has progressed further in the Netherlands than in Belgium, attending to pronunciation rather than lexis, as Grondelaers et al. did, and focusing on regional speech rather than base dialects. We noted further that the differences between the Dutch and Belgian regional pronunciations were significant, but quite small. In a third brief examination we showed that the non-words that Impe (2010: 17ff) derived from existing words were quite similar in their pronunciation distances (to genuine words), confirming their appropriateness in for use in psycholinguistics.

Fourth, the results of the phonetic distance computations show that regional speech does not conform to the Auer-Hinskens model for regiolects, and we speculated that it may be too difficult, even for professional speakers, to remain within the prescribed areas. We explicitly did not take issue with the developmental dynamic of regional speech implicit in the conical model, which is well motivated.

Standardization has progressed further in Netherlandic Dutch than in Belgian Dutch. In the current study phonetic differences both among regional speech forms and between the standard and the regional speech forms are slightly larger in Belgian Dutch than in Netherlandic Dutch. These relations suggest a larger vertical range of variety, but no conclusions can be drawn about horizontal variety yet. For more insight into the horizontal area of variety in Belgian and Netherlandic Dutch, the research has to be complemented with more dialectal language data.

Finally we suggest that the material might serve to provide further insights into the development of regional speech with further analysis and if used together with other material. With respect to the development of regional speech, one might wish to examine the material to determine which words tend to carry the strongest regional signals, i.e. identify the speaker as from a particular region. Once these "shibboleths" (Prokić, Çöltekin & Nerbonne 2012) are known, one might further investigate their properties, including their token frequency, their distribution with respect to regions, and the (token) frequency with which they are used in regional speech of the sort our speakers specialized in, i.e. radio broadcasts intended for entire regions. We should like to understand how well the professional regional speakers exploit such words. Note that the final topic just suggested would require that one collect the speech of regional broadcasts directly, meaning we have crossed the line in topics to those which require supplementary material. Once that supplementary material was available

in transcribed form, we should wish to examine the degree to which our word lists are representative of the regional speech heard in the broadcasts.

## *Bibliography*

Auer, P. (2005). Europe's sociolinguistic unity, or: A typology of European dialect/standard constellations. In N. Delbecque, J. van der Auwera & D. Geeraerts (eds.). *Perspectives on variation. Sociolinguistic, historical, comparative*. Berlin: De Gruyter Mouton, 7-42.

Auer, P. & Hinskens, F. (1996). The convergence and divergence of dialects in Europe. New and not so new developments in an old area. *Sociolinguistica*, 10, 1-30.

Beijering, K., Gooskens, C. & Heeringa, W. (2008). Predicting intelligibility and perceived linguistic distances by means of the Levenshtein algorithm. In M. van Koppen & B. Botma (eds.). *Linguistics in the Netherlands.* Amsterdam: John Benjamins, 13-24.

Deprez, K. (1997). Diets, Nederlands, Nederduits, Hollands, Vlaams, Belgisch-Nederlands. In M. Clyne (ed.). *Undoing and redoing corpus planning*. Berlin: Mouton de Gruyter, 249-312.

Ferguson, C. (1959). Diglossia. *Word*, 15, 325–340.

Geeraerts, D. (2001). Een zondagspak? Het Nederlands in Vlaanderen: Gedrag, beleid, attitudes. *Ons Erfdeel*, 44(3), 337-343.

Geerts, G. (1992). Is Dutch a pluricentric language? Pluricentric languages: Differing norms in different nations. In M. Clyne (ed.). *Undoing and redoing corpus planning*. Berlin: Mouton de Gruyter, 71-91.

Gooskens, C. (2006). Linguistic and extra-linguistic predictors of inter-Scandinavian intelligibility. In J. van de Weijer & B. Los (eds.). *Linguistics in the Netherlands*. Amsterdam: John Benjamins, 101-113.

Gooskens, C. (2007). The contribution of linguistic factors to the intelligibility of closely related languages. *Journal of Multilingual and Multicultural Development*, 28(6), 445-467.

Gooskens, C. & Heeringa, W. (2004). Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data. *Language Variation and Change*, 16(3), 189-207.

Gooskens, C., Heeringa W. & Beijering, K. (2009). Phonetic and lexical predictors of intelligibility. *International Journal of Humanities and Arts Computing*, 2(1-2), 63–81.

Grondelaers S., Van Aken, H., Speelman, D. & Geeraerts, D. (2001). Inhoudswoorden en preposities als standaardiseringsindicatoren. De diachrone en synchrone status van het Belgische Nederlands. *Nederlandse Taalkunde*, 6, 179-202.

Heeringa, W. (2004). *Measuring dialect pronunciation differences using Levenshtein distance*. Ph.D. Dissertation Groningen: University of Groningen.

Impe, L. (2010). Mutual intelligibility of language varieties in the Low Countries: linguistic and attitudinal determinants. Ph.D. Diss. Leuven: University of Leuven.

Impe, L., Geeraerts, D. & Speelman, D. (2008). Mutual intelligibility of standard and regional Dutch language varieties. *International Journal of Humanities and Arts Computing*, 2(1–2), 101-117.

Kürschner, S., Gooskens, C. & Van Bezooijen, R. (2008). Linguistic determinants of the intelligibility of Swedish words among Danes. *International Journal of Humanities and Arts Computing*, 2(1-2). 83-100.

Nerbonne, J. & Heeringa, W. (2010). Measuring dialect differences. In J.-E. Schmidt & P. Auer (eds.). *Language and Space: Theories and Methods.* Chap. 31. In series Handbooks of Linguistics and Communication Science. Berlin: Mouton de Gruyter, 550-567.

Nerbonne, J., Colen, R., Gooskens, C., Kleiweg, P. & Leinonen, T. (2011). Gabmap - A web application for dialectology. *Dialectologia*, 65-89.

Nerbonne, J., Van Ommen, S., Wieling, M. & Gooskens C. (to appear). Measuring socially motivated pronunciation differences. In L. Borin & A. Saxena (eds.). (Comparing) Approaches to Measuring Linguistic Differences (tentative title). Berlin: Mouton De Gruyter.

Prokić, J., Çöltekin, Ç. & Nerbonne, J. (2012). Detecting shibboleths. In M. Butt & J. Prokić (eds.). Visualization of language patterns and uncovering language history from multilingual resources. Workshop at the 13th Conference of the European Chapter of the Association for Computational Linguistics. Avignon: Association for Computational Linguistics, 72-80.

Speelman, D., Grondelaers, S. & Geeraerts, D. (2008). Variation in the choice of adjectives in the two main national varieties of Dutch. In G. Kristiansen & R. Dirven (eds.). *Cognitive sociolinguistics: Language variation, cultural models, social systems*. Berlin: De Gruyter, 205-233.

Taeldeman, J. (1993). Welk Nederlands voor Vlamingen? In L. De Grauwe & J. De Vos (eds.). *Van sneeuwpoppen tot tasmuurtje: Aspecten van Nederlandse taal- en literatuurstudie.* (Spieghel historiae 33), 9-28.

Van de Velde, H. (1996). Variatie en verandering in het gesproken Standaard-Nederlands (1935-1993). Ph.D. Dissertation Nijmegen: University of Nijmegen.

Videnov, M. (1999). The present-day Bulgarian language situation: Trends and prospects. *International Journal of the Sociology of Language*, 135, 11-36.

Wieling, W., Heeringa, W. & Nerbonne, J. (2007). An aggregate analysis of pronunciation in the Goeman-Taeldeman-Van Reenen-Project data. *Taal en Tongval*, 59, 84–116.

## Appendix 1 item list - words

Binational
aandacht
afhangen
afvallen
afwezig
antwoord
armoede
arrogant
bang
bed
begrijpen
behalen
benauwd
besparen
betalen
bezetten
bezoek
bloot
boom
brullen
brutal
diet
dierentuin
dreiging
droog
duur
eerlijk
eeuwigheid
eigenwijs
fotograaf
gebeuren
gebouw
gebruik
gebruiken
gestoord
gevoel
goed
goedkoop
groot
haastig
hard
helm
hoog
inzitten
kapstok
keuken
klinker
kopen
kort
lachen
leunen
levend
levendig
liggen
maaien
maken
midden
mogen
nemen
noemen
nummer
onveilig
oorlog
oprit
opvolgen
opwinden
pakken
passagier
persoon
piekeren
prachtig
puin
rijden
rinkelen
rustig
scheren
schok
schrijven
schutting
simpel
slecht
spreken
stabiel
stampen
steeg
vakantie
vallen
verhaal
verstaanbaar
vlieg
vreemd
water
weten
wijf
wijze
winkel
wrijven
zegen
zeggen
ziek
zwaar
Belgian Dutch
aanvaarden
ambetant
autostrade
babbelen
blokken
constant
content
deftig
dikwijls
droevig
eenvoudig
eigenaardig
enorm
fameus
flik
fuif
gebuur
gerust
gezaag
goesting
hesp
juist
klucht
kotmadam
kuisen
leerkracht
living
ma
ogenblik
onnozel
opdoen
peizen
pint
plezant
proper
schoon
smijten
spijtig
tamelijk
uitslag
vake
verband
verlof
verschieten
verstaan
vlug
vrijen
wenen
zalig
zeveren
Netherlandic Dutch
aardig
balen
bedrijf
buurman
eikel
gaaf
gauw
geinig
gezeik
gillen
goor
gozer
hartstikke
heerlijk
hoeven
huilen
huiskamer
jammer
jatten
jemig
jus
kegelen
kroeg
kweekschool
lullen
medicijn
meid
microfoon
moment
ome
onwijs
onzin
ouwehoeren
pinnen
pissig
ranzig
salaris
schoonmaken
schrikken
snappen
sneu
toesturen
triest
vent
verkering
vervelend
vrolijk
zakken
zeiken
ziekenhuis

# Appendix 2 Item list non-words

afdas
afdoek
afdok
afkaat
astig
baarsel
bafoor
barrecht
beboeken
bedraggen
bedrazen
bedregen
bedreizen
bedrijken
begieren
belekeren
belen
bemasseren
besaald
beziegen
boekig
bontrijk
bosknecht
boustig
deparatie
dozing
eenzig
eepratie
eervaas
eikbeek
fadijk
fantaar
fantig
fantuin
fijntig
fontaar
fonter
gaapdak
geulaar
glem
greilen
gropen
hakkig
hakspecht
hapig
heisig
hekkig
heleren
herig
hijlen
hijten
hoelig
hoepig
hoffen
hokig
hongstig
houkeren
huikig
husen
huurzaam
ipdak
kaakdoek
kalderen
kamees
kameit
kantaar
kanteer
kantijs
kastoor
katen
keekbaar
keerint
kegen
keuking
keuzing
kijling
klaten
koerding
kokken
komaat
komeeuw
kommoet
komoot
konter
kooling
koordoek
kotmeet
kratzaal
laam
lasig
leenzaak
leestig
leiderig
leten
lijfzacht
lijmmacht
lokmeet
lontijn
lontlijn
lookbak
lookhoek
loren
lorpen
luiren
luum
makaat
mankoor
manter
massaar
massig
mazeen
meelderen
meelzaam
mekig
mepen
mils
moors
mostaal
noodijk
oogstig
oolbast
oopdek
oorzaam
operaat
operig
opheert
oprakig
opreftig
paarkool
paarzaam
pantoer
pastaar
paten
pekzak
pellatie
perater
potrecht
pralig
prateren
pratiek
pratig
prieten
ralen
rankig
ratzaag
roelen
roethol
roffen
roking
rolen
rolken
rookbeek
rulen
sangig
saparatie
schalferen
schaperen
schidderen
schikkeren
schomeet
seeuw
semer
semig
sleem
slep
solm
spaulderen
stafboor
stangig
stezen
stolpig
tagen
taspen
teupen
teuzen
tijgen
tijzen
tonrecht
torven
trapen
trarken
treisten
treppen
vaargeil
vaarlaar
vaasgeur
vakrecht
veergel
veerzaal
verdriegen
verdrugen
verklogen
voorgoel
vrazen
vulzaam
vuurmaag
zaking
zauwig
zeering
zelderen