Computers and the Humanities **37:** 257–271, 2003.

The Use of the Almeida-Braun System in the Measurement of Dutch Dialect Distances

WILBERT HEERINGA and ANGELIKA BRAUN
Faculty of Arts, Humanities Computing, University of Groningen, the Netherlands; Institute of Germanic Linguistics, Phonetics, Philips-University of Marburg, Germany

Abstract. Measuring dialect distances can be based on the comparison of words, and the comparison words should be based on the comparison of sounds. In this research we used an adjusted version of an articulation-based system, developed by Almeida and Braun (1986) for finding sound distances, using the IPA system. For comparison of two pronunciations of a word corresponding with two different varieties, we used the Levenshtein algorithm, which finds the easiest way in which one word can be changed into the other by inserting, deleting or substituting sounds. As operations weights of these three operations we used distances as found with the Almeida & Braun system. The dialect distance is now equal to the average of a range of word distances. We applied the technique to 360 Dutch dialects. The transcriptions of 125 words for each dialect are taken from the *Reeks Nederlandse Dialectatlassen* (Blancquaert and Peé, 1925–1982). We get a division with clear similarities to traditional dialect maps when classifying dialects. Using logarithmic sound distances improves results compared to results based on constant sound distances.

Key words: dialect, dialectology, dialectometry, phonologic (dis)similarity

1. Introduction

Kessler (1995) introduced the use of the Levenshtein distance as tool for measuring dialect distances. The Levenshtein distance is a string edit distance measure and Kessler applied this algorithm to the comparison of Irish dialects. Later on this approach was taken up by Nerbonne *et al.* (1996) and applied to Dutch dialects. When sounds are compared in the algorithm, there are only two possibilities: they are equal or they are not equal.

Nerbonne and Heeringa (1997, 1998, 2001), and Nerbonne *et al.* (1999a, b), took steps forward by using a more refined comparison of segments. Sound distances are found by the comparison of features. Feature systems of Hoppenbrouwers and Hoppenbrouwers (2001) and Vieregge *et al.* (1984) are used.

In this paper we present the use of an adjusted version of a system which was developed by Almeida and Braun (1986). Any attempt to quantify distances in pronunciation between dialects of a given language and/or the degree of dialectality (i.e. the distance between a certain dialect and what is considered the "standard") presupposes the consideration of the phonetic domain which one intends to study. This could be either the articulatory or the acoustic or the perceptual level. The

Almeida & Braun system is an articulatory system in which sound distances are derived from the IPA vowel quadrilateral and the IPA consonant table. Just like the perception-based system of Vieregge *et al*. (1984), this system originally was developed in order to evaluate the reliability and validity of phonetic transcriptions.[1] In the case of transcriptions, the dependent variable is the transcription, whereas the independent variable is the speech utterance. The transcriber skill may be derived from the similarity of two transcriptions.

The comparison of the standard with the dialectal utterances refers to the measurement of dialectality, whereas the comparison between different phonetic realizations of a given utterance in various dialects relates to the question of between-dialect differences, which is of interest here. At any rate, the distance measures which were developed to assess transcriptions can be just as well used to quantify dialect distances. This, however, presupposes that the variable "transcriber" is kept constant by either having only one transcriber, who also undergoes reliability testing, or working with high-quality consensus transcriptions. Otherwise, there is the danger of creating so-called *Exploratorendialekte* ("explorer dialects"), i.e. "dialects" created not by differences in pronunciation but by different people transcribing them. Unfortunately this phenomenon is found in the *Reeks Nederlandse Dialectatlassen* (cf. Goossens, 1965), although the goal was that transcribers should work using Blancquaert's guidelines. On the other hand, the use of a good feature system will reduce the effect of these differences maximally.

The goal of this paper is to investigate whether sound distances derived from the IPA tables can be used in finding dialect distances. In Section 2 we describe the adjusted Almeida & Braun system and the way in which sound distance are found using this system. In Section 3 we describe the Levenshtein distance and explain how the sound distances found with the Almeida & Braun system are used in this algorithm. We apply the methodology to material of the *Reeks Nederlandse Dialectatlassen* (RND), a series of Dutch dialect atlasses (Blancquaert and Peé, 1925–1982). In Section 4 we end with conclusions.

2. Comparison of Sounds

Vieregge *et al*. (1984) presented a feature system which was developed for checking the quality of phonetic transcriptions. This involves comparison of consensus transcriptions. The system consists of 4 multi-valued features only for vowels, and 10 multi-valued features only for consonants. For the vowels features were chosen which reflect the traditional vowel scheme. Therefore data from the literature (which have been tested experimentally to a great extent) was consulted. For a subset of Dutch consonants (viz., the [p], [b], [t], [d], [k], [f], [v], [s], [z], [x], [m], [n], [ŋ], [l], [ʀ], [w], [j] and [h]) perceptual distances were found. In a perception experiment subjects were asked to give the distance between two consonants on a scale from 1 (minimal dissimilarity) to 10 (maximal dissimilarity).

Next consonant features were chosen and weighed so that sound distances on the basis of features approach the perceptual distances maximally.

Although the system was originally developed for Dutch vowels and consonants, with some extensions it may also be used for other languages as Cucchiarini (1993) showed. She extended the system so as to accommodate consonants of Limburg and Czech that were not included, as well as other sounds that probably could crop up in the transcriptions which she used. However, when expanding the system to other languages, one should be aware of the fact that different languages have different sound systems, the phonological spaces may be different filled. Cucchiarini realizes this and writes (p. 97): “So, as it was clear that a theoretically satisfactory evaluation system was not possible, we tried to obtain a system that would at least be satisfactory from a practical point of view”. Probably the use of a Dutch perceptual Vieregge system which is extended and applied to e.g. Czech will reflect the perception of Dutch people listening to Czech, rather than the perception of the Czech speakers themselves.

In the original Vieregge system only *length* is processed. However, other IPA suprasegmentals and diacritics such as *nasality* can be processed, if needed by incorporating other features.

At the same time the Vieregge system was developed, Almeida and Braun developed an alternative based on the IPA tables. The system was initiated in the phonetics department of the research institute for German linguistics “Deutscher Sprachatlas” (Marburg, Germany) in 1980 and was further developed and formalized later. In contrast to the Vieregge system the Almeida & Braun system is articulation-based. The system relies on the assumption that transcription is a process which first consists in an imitation of the relevant utterance, furthermore in an inference on the part of the transcriber on the articulatory gestures of the speaker, and finally in a phonetic description thereof (Almeida, 1984; Almeida and Braun, 1985). The description is carried out in terms of the criteria used by the International Phonetic Alphabet (the version revised to 1993)[2] which essentially consists in an abbreviation for a combination of articulatory features. From the beginning the system covers the complete IPA vowel and pulmonic consonant set. Furthermore, in the original system a large number of suprasegmentals and diacritics can be processed.

In Section 2.1 we explain how vowel distances are calculated and in Section 2.2 consonant distances. In Section 2.3 we describe how diphthongs can be defined in this system. In Section 2.4 and Section 2.5 it is explained how subsets of IPA suprasegmentals and IPA diacritics are processed. Once the sound distances are derived from the IPA tables, eventually the logarithmic values are used in dialect comparison. This will be examined in Section 2.6.

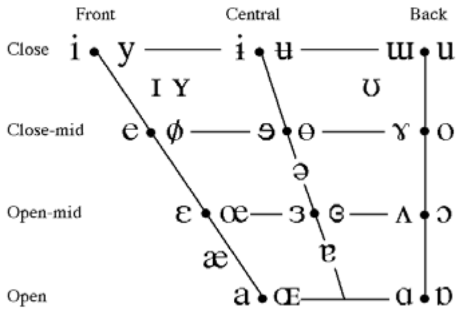

Figure 1. The IPA vowel quadrilateral.

Table I. Vowel features and their possible values

vowel		advancement		height		rounded	
no	0	front	1	close	1	no	0
yes	1	central	2	near-close	2	yes	1
		back	3	close-mid	3		
				central	4		
				open-mid	5		
				near-open	6		
				open	7		

2.1. Vowel distances

The basis for finding vowel distances is the IPA vowel quadrilateral as given in Figure 1. The quadrilateral reflects three features: *advancement*, *height* and *rounding*. The possible values for the features are listed in Table I. A feature *vowel* is added. Usually for vowels this feature is set to 1. The use of this feature will be explained later.

In the vowel quadrilateral we regard the distance between e.g. ɛ vs. ɜ (advancement: front vs. central), ɛ vs. æ (height: open-mid vs. open), and ɛ vs. œ (rounding: no vs. yes) as one step. So when simply subtracting the corresponding feature values from each other (see Table I) and taking the absolute value, we get a distance of one for each of these three pairs.

Besides the basic features derived from the vowel quadrilateral, some features are added to process some IPA suprasegmental and diacritics. We added the features *long*, *nasal*, *diphthong*, *breathy*, *creaky*, *toneme 1*, *toneme 2*, *circumflex*. Usually they are 0 (absent) or 1 (present). The feature *long* may also have the value 0.5 to represent "half long".

For the use of the Levenshtein distance we need also a definition of 'silence' in terms of the vowel features (see Section 3.2). We define it equal to the schwa, but now the type feature *vowel* is set to 0.

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d		ʈ ɖ	c ɟ	k ɡ	q ɢ		ʔ
Nasal	m	ɱ		n		ɳ	ɲ	ŋ	ɴ		
Trill	ʙ			r					ʀ		
Tap or Flap				ɾ		ɽ					
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant	w	ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Figure 2. The IPA consonant table.

Table II. Consonant features and their possible values

consonant		place		manner		voice	
no	0	bilabial	1	plosive	1	no	0
yes	1	labiodental	2	nasal	2	yes	1
		dental	3	trill	3		
		alveolar	4	tap or flap	4		
		postalveolar	5	fricative	5		
		retroflex	6	lateral fricative	6		
		palatal	7	approximant	7		
		velar	8	lateral approximant	8		
		uvular	9				
		pharyngeal	10				
		glottal	11				

2.2. CONSONANT DISTANCES

In our system we only use the pulmonic consonants, the non-pulmonic ones are not included. The basis for finding consonant distances is the IPA table for pulmonic consonants as given in Figure 2. In this figure it can be seen that in our system the voiced labial-velar approximant [w] is regarded as and will be treated like a bilabial approximant.

The table reflects three features: *place*, *manner* and *voice*. We regard both *place* and *manner* as a scale. The feature *place* gives the *location* of closure and ranges from front to back. The feature *manner* gives the *degree* of closure with roughly the following degrees: complete closure (plosives), oral closure (nasals), intermittant closure (trills, tap and flap), friction (fricatives) and frictionless approximation (approximants). The possible values for the features are listed in Table II. A type feature *consonant* is added. Usually for consonants this feature is set to 1. Its use will be explained hereafter.

In the consonant table we regard the distance between e.g. [z] vs. [ɾ] (manner: fricative vs. tap or flap), [z] vs. [ʒ] (place: alveolar vs. postalveolar) and [z] vs. [s] (voice: voiced vs. voiceless) as one step. So when simply subtracting the corresponding feature values from each other (see Table II) and taking the absolute value, we get a distance of one for each of these three pairs. We regard the distance between e.g. [ɱ] and [v] (manner: fricative vs. approximant) and [ʙ] and [r] (place: bilabial vs. alveolar) as two steps, although they may be regarded as neighbors.

Besides the basic features derived from the consonant table, some features are added to process some IPA diacritics. We added the features *syllabic* and *apical*. Usually they are 0 (absent) or 1 (present).

As for the vowels, we also need a definition of 'silence' in terms of consonant features, which will be used in the Levenshtein algorithm (Section 3.2). We defined it to be equal to the glottal stop, however, the type feature *consonant* is set to 0.

2.3. DIPHTHONGS

In the original IPA system only monophthongs are described. To be able to deal with diphthongs nonetheless, there are two possibilities. In the first approach a diphthong is processed as the sequence of two monophthongs. The results in this article are based on this approach. In the second approach a diphthong is processed as one sound with a changing color. We are also able to make calculations using this approach, where the feature bundles of the diphthongs are defined by analogy of the feature system of Hoppenbrouwers and Hoppenbrouwers (2001). For *closing diphthongs* the feature values are found as follows:

advancement	:	mean of both segments
high	:	value of first segment
round	:	mean of both segments
long	:	always *long* = 1
diphthong	:	always *diphthong* = 1

For *centring diphthongs* the feature values can be estimated as follows:

advancement	:	mean of both segments
high	:	value of first segment
round	:	mean of both segments
long	:	always *long* = 1
diphthong	:	always *diphthong* = 0

In Dutch dialects a long vowel in the transcription of the pronunciation of the one dialect may often correspond with a centring diphthong in another. To avoid exaggerating the difference between both pronunciations, for centring diphthongs the feature *diphthong* is set to 0, just as for (long) monophthongs.

2.4. SUPRASEGMENTALS

The only IPA suprasegmental we process for the RND data is *length*. When processing length and syllabicity in feature systems, there are two possibilities. First, they can be processed by adapting the values of the features *length* and *syllabic*. Second, length can be processed by modifying the transcription. In that case short sounds are kept unchanged, sounds with no length indication are doubled, half long sounds are trebled, and long sounds are quadrupled. Syllabic sounds are treated as long sounds, so they are quadrupled.

In this research we want to keep as close as possible to the original approach of Almeida and Braun. In their system, 'extra-short' is not processed. We process it by keeping extra short sounds unchanged and doubling all other sounds. For vowels 'half long' and 'long' are processed by using the feature *length*: the values 0.5 and 1 respectively are usually assigned. However, for the RND data it is better to process 'half long' as 'long' since these length marks are not consistently used in this data source (see, Heeringa (2001) for more details). For consonants only 'long' is processed for only the nasals. In these cases the added feature *syllabic* is set on 1.

2.5. DIACRITICS

The Almeida & Braun system allows one to process a wide range of diacritics, such as *advanced*, *retracted*, *raised*, *lowered* (tendentials), *breathy* and *creaky* voice or *apical*.

In the RND not all IPA diacritics are used. Furthermore, a more detailed transcription will be more transcriber-dependent. Hoppenbrouwers and Hoppenbrouwers (2001), who apply their *feature frequency method* to the RND data for finding dialect similarities, use only a restricted set of diacritics. Following Hoppenbrouwers and Hoppenbrouwers we only process *voiced*, *voiceless*, *half nasalized*, *nasalized* and *syllabic*.

For normally voiced sounds noted as voiceless as well as normally voiceless sounds which are noted to be voiced the feature *voice* is set to 0.5. In contrast to the IPA system, sounds may be marked as half-nasalized in the RND. Half-nasalized sounds may be conceived of being produced with the velum in an intermediate position between fully raised and fully lowered. For half nasalized sounds the feature *nasal* is set to 0.5, and for nasalized sounds it is set to 1. Nasality is processed for vowels only.

Syllabicity is only processed for the [m], [n], [ŋ], [r] and [l]. This can be processed by changing the transcription or by adapting a feature. In the first case syllabic sounds are treated as long sounds (see Section 2.4). To keep close to the original Almeida & Braun system we process syllabicity by setting the feature *syllabic* to 1. We are aware of the fact that there is no agreed phonetic definition for syllabicity. However, syllabicity forms part of the descriptive framework of the IPA and thus needs to be dealt with since it occurs in both sets of transcriptions.

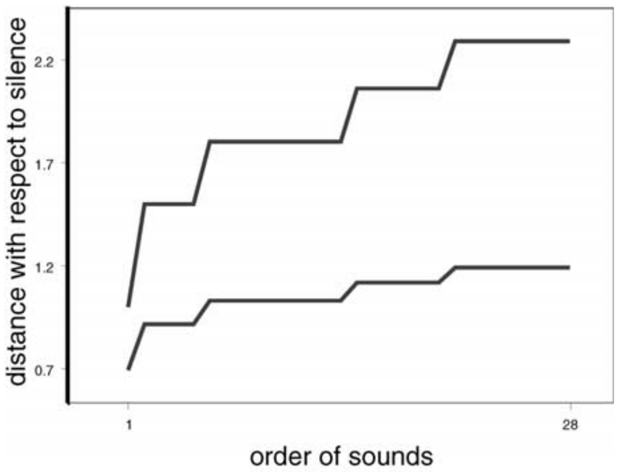

Figure 3. Linear (upper) and logarithmic (lower) distances of 28 IPA vowels with respect to silence. Distances are calculated as the sum of the differences between corresponding features. The graph shows the distances sorted from low (left) to high (right). Greater distances are reduced more than smaller ones by using the logarithm.

In the RND syllabicity of nasals is sometimes marked with a "ˌ" under the sound, and sometimes with a "ː".

2.6. LOGARITHMIC DISTANCES

Using the feature bundles the distance between two sounds can be calculated. In the research presented in this paper this is done by taking the sum of the absolute differences of each pair of corresponding feature values. When the distance exceeds a certain ceiling, in the original Almeida & Braun system that distance is set to the value of a ceiling. However, the question arises as to what value the ceiling should be set to. Therefore, instead of using a ceiling we take the logarithm of the feature bundle distances. Because the distance between identical sounds is 0, and the logarithm of 0 is not defined, we first increase the distance with 1 and next calculate the logarithm of the distance. In this way, the distance between equal sounds still remains 0 since the logarithm of 1 is equal to 0. In general we calculate *ln(distance+1)*.

In Figure 3 the effect of taking the logarithm of the IPA vowel distances is shown. For each of the 28 IPA vowels the distance with respect to silence is calculated. Next the distances are sorted from low to high. In both cases, linear and logarithmic, [ə] is most like silence and [i], [y], [ɯ], [u], [a], [ɶ], [ɑ], [ɒ] are all most unlike silence. The graph shows the sorted distances. The points corresponding with distances are connected by lines to get a clearer picture. The effect of taking the logarithm is that higher distances are relatively more decreased than lower distances.

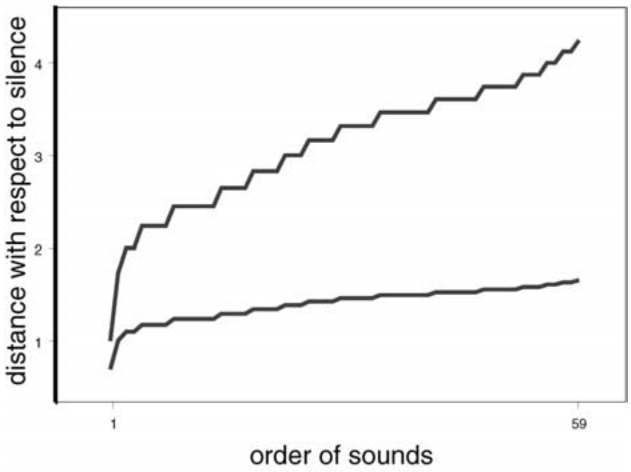

Figure 4. Linear (upper) and logarithmic (lower) distances of 59 IPA consonants with respect to silence. Distances are calculated as the sum of the differences between corresponding features. The graph shows the distances sorted from low (left) to high (right). Greater distances are reduced more than smaller ones by using the logarithm.

In Figure 4 the effect of taking the logarithm of the IPA consonant distances is shown. For each of the 59 IPA consonants the distance with respect to silence is calculated. Next the distances are sorted from low to high. In both cases, linear and logarithmic, [ʔ] is most like and [w] is most unlike silence. Just as for the vowels the graph shows the sorted distances. The points corresponding with distances are connected by lines again. Of course the same effect as for the vowels is seen here: higher distances are relatively more decreased than lower distances when taking the logarithm.

3. Comparison of Dialects

Once we have found distances between the IPA sounds, we can use them in the finding of word distances. Subsequently word distances are used to find dialect distances.

In this section we describe our calculation of dialect distances among 360 Dutch varieties. We give a brief description of the data set in Section 3.1. The way we find word distances, and next dialect distances is explained in Section 3.2. On the basis of the dialect distances, dialect areas can be found. This is described in Section 3.3.

3.1. Data source

The data on which the comparison of the Dutch dialects is based comes from the *Reeks Nederlandse Dialectatlassen* (RND) which was compiled by Blancquaert and Peé (1925–1982). From these atlases we chose 360 dialects. The dialects are roughly evenly scattered over the Dutch language area. In the RND for each dialect the same 141 sentences are translated and transcribed in phonetic script. From these sentences we chose 125 words, which we think are representative for the range of sounds in the varieties. More about the selection and digitalization of words from the RND can be found in Heeringa (2001).

3.2. Comparing dialects

Using the Levenshtein distance dialects are compared by comparing the pronunciation of a word in the one with the pronunciation of the corresponding word in the other. The algorithm determines how one pronunciation is changed into the other by inserting, deleting or substituting sounds. Weights are assigned to this three operations. In the simplest form of the algorithm, all operations have the same cost, e.g., 1. Assume *saw a girl* is pronounced as [sɔːəgʌrl] in Standard American, and as [sɔːrəgɜːl] in one of the dialects of Boston. Changing the one pronunciation into the other can be done as follows:

sɔəgʌrl	delete r	1
sɔəgʌl	replace ʌ/ɜ	1
sɔəgɜl	insert r	1
sɔrəgɜl		
		3

In fact many sequence operations map [sɔːəgʌrl] to [sɔːrəgɜːl]. In the most long-winded way first all segments of the first pronunciation are deleted and subsequently all segments of the second pronunciation are inserted. This gives a cost of 7 deletions + 7 insertions = 14 operations. However, the Levenshtein algorithm finds always the minimum number of operations needed to change the one pronunciation into the other. We suppose that this reflects the way in which differences in pronunciation are perceived.

Comparing words in this way, in general the distance between longer words will be greater than the distance between shorter words. The longer the word, the greater the chance of differences with respect to the corresponding word in an other variety. Because this does not accord with the idea that words are linguistic units, the sum of the operations is divided by the length of the longest alignment which gives the minimum cost. The longest alignment has the greatest number of matches. In our example we have the following alignment:

s	ɔ		ə	g	ʌ	r	l
s	ɔ	r	ə	g	ɜ		l
0	0	1	0	0	1	1	0

The total cost of 3 is now divided by the length of 8. This gives a word distance of 0.38 or 38%.

In Section 2 we explained how distances between sounds can be found using the IPA vowel quadrilateral and the IPA consonant table. This makes it possible to refine our Levenshtein algorithm by using the segment distances as operations weights. Now the cost of insertions, deletions and substitutions is not always equal to 1, but varies, i.e. it is equal to the IPA feature distance between the sound and 'silence' (insertions and deletions) or between both sounds (substitution).

Because in the IPA system vowels and consonants have different features, and because we want to allow for syllabification, the Levenshtein algorithm is adapted so that only a vowel may match with a vowel, a consonant with a consonant, the [j] or [w] with a vowel (or opposite) and the [i] or [u] with a consonant (or opposite). When the [i] or [u] is compared to a consonant, it is treated like the [j] or [w]. When the [j] or [w] is compared to a vowel, it is treated like the [i] or [u].

In our research we used 125 words. So when comparing two dialects we get 125 Levenshtein distances. Now the dialect distance is equal to the sum of the 125 Levenshtein distances divided by 125. All distances between the 360 dialects are arranged in a $n \times n$ matrix. Correlating the distances with the geographic distances it appears that $r = 0.65$, a significant correlation.

3.3. Classifying dialects

On the basis of the matrix which contains the distances between the 360 dialects, we applied cluster analysis. The goal of clustering is to identify the main groups in complex data. Clustering is a iterative procedure. At each step of the procedure we select the shortest distance in the matrix, and then fuse the two data points which gave rise to it. Since we wish to iterate the procedure, we have to assign a distance from the newly formed cluster to all remaining points. The iterations are repeated until no elements are left which can be fused to a new cluster. The result is an hierarchically structured tree in which the dialects are the leaves (Jain and Dubes, 1988). For calculating the distance from a newly formed cluster to all remaining points, we used the Ward's method. This method appears to give the most well-balanced tree from all common alternatives. From the dendrogram the eight most significant groups were identified. On the map in Figure 5 the areas corresponding with these groups can be found. The different grey values have no meaning, but are chosen so that neighboring areas have always different colors and so that borders thus become visible. Dialect islands are marked with a diamond. Especially in Friesland (the northwestern area) in a number of places a town Frisian dialect is spoken, which are dialect islands in the Frisian dialect continuum.

In traditional dialectology three main groups are distinguished in the Dutch language area: Frisian, Lower Saxon and Lower Franconian. These groups can obviously be found on our map. On the map the northwestern area containing Grouw forms the Frisian area. The town Frisian varieties (most diamonds and the

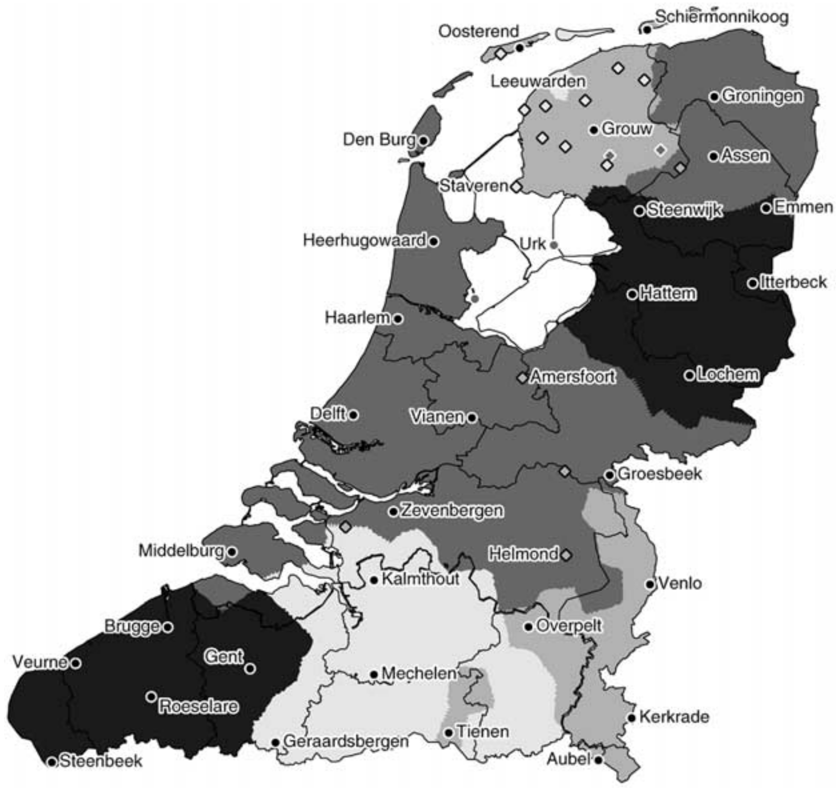

Figure 5. Using cluster analysis a dendrogram is derived from the 360 × 360 matrix. From the dendrogram the 8 most significant groups can be identified. The grey colors have no meaning, they are chosen so that borders between neighboring areas can be seen. Dialect islands (mainly town Frisian varieties) are marked with a diamond. Dialect islands with the same color are in the same group and form one group with dialects of one or more nearby areas with the same color.

island Ameland, north of Leeuwarden) are clearly distinct from 'pure' Frisian. The two most northeastern areas containing Groningen and Hattem form the Lower Saxon area. The remaining four areas form the Lower Franconian area.

The division has clear similarities with dialect maps of Te Winkel (1901) and Daan and Blok (1969). On all maps the Frisian area can clearly be found. Further on all maps a Groningen dialect area can be found. However, the maps do not agree about the southern boundary of the Groningen area, which lies (partly) in the northern part of Drenthe. On all maps the boundary between Lower Saxon and Lower Franconian is found at about the same place. The boundaries of the Franconian area on our map are more in accordance with the map of Daan and Blok than with the map of Te Winkel. On both the map of Daan and Blok and our map the northern boundary of the eastern group in Belgium coincides roughly with the state border between the Netherlands and Belgium. For the Limburg area (the most southeastern area containing Kerkrade) applies that both Te Winkel and Daan and Blok have similar boundaries. On our map the western boundary is more westwards, and the northern boundary is more northwards. However, it is striking that the dialects of Tienen and Diest (located in the area north of Tienen) belongs also to the Limburg group on our map. Both dialects are borderline cases in the Limburg area on traditional maps. Furthermore, it is striking that the dialects of Amersfoort, Ravenstein (diamond left from Groesbeek), Helmond and Steenbergen (diamond lower left of Zevenbergen) belong to the Limburg dialect group as well. On our map a Lower Franconian dialect belongs to the Limburg group if the uvular [ʀ] is used (in most cases). For other Franconian dialects the alveolar [r] is noted in the transcriptions.

4. Conclusions

With the Almeida & Braun system distance between all IPA sounds can be easily found. We used this distances for finding dialect distances. Once having dialect distances, we made a dialect classification using cluster analysis. The main groups of the tree obtained by clustering shows a division with clear similarities to traditional dialect maps.

We compared results obtained on the basis of non-logarithmic Almeida & Braun sound distances with results obtained on the basis of a similar system in which two sounds are equal or not. In this simplified system the distance between e.g. [i] and [ɪ] is as large as the distance between [i] and [ɒ]. However, the results on the basis of the non-logarithmic Almeida & Braun sound distances were not obviously more similar to traditional results. However, when using the logarithmic distances, the similarity with the traditional division is improved.

We get a rather high correlation between dialect distances and geographic distances ($r = 0.65$) when using logarithmic feature distances. This is lower than distances obtained with the system in which no variable sound distances are used ($r = 0.68$), but higher than when using non-logarithmic sound distances ($r = 0.63$). The correlations show a rather strong relation between geography and variation in the dialect continuum.

In this article we focused on the Almeida & Braun system. This system is interesting because it uses the well-known IPA tables. However, many other feature systems can be used, e.g. the feature systems of Vieregge *et al.* (1984), Cucchiarini (1993) or Ladefoged and Maddieson (1996). In our research we could only regard a restricted number of feature systems. The intention is to perform a thorough comparison and validation of a set of different systems in future.

Acknowledgement

We thank Peter Kleiweg for the mapping software.

Notes

[1] The reliability of transcriptions is measured by determining the degree of similarity between transcriptions carried out either by the same transcriber at different times (this corresponds to the use of "reliability" in its strict sense; cf. Bürkle, 1986) or by different transcribers (an option also left Vieregge *et al*., 1984); the validity of transcriptions is measured by comparing individual transcriptions with master transcriptions (cf. Vieregge *et al*., 1984).
[2] The system can be found in the *Handbook of the International Phonetic Association* (1999) as well as via: *http://www2.arts.gla.ac.uk/IPA/ipachart.html*.

References

Almeida A. (1984) Zur Methodik der Datenaufbereitung in der Linguistik: Das Beispiel phonetischer Transkription. In Berger, L. (ed.), *Sprechausdruck*, Scriptor, Frankfurt am Main, pp. 111–122.

Almeida A., Braun A. (1985) What is Transcription? In Kürschner, W. and Vogt, R. (eds.), *Grammatik, Semantik, Textlinguistik. Akten des 19. Linguistischen Kolloquiums Vechta 1984*, Vol. 1, Tübingen, pp. 37–48.

Almeida A., Braun A. (1986) "Richtig" und "falsch" in phonetischer Transkription; Vorschläge zum Vergleich von Transkriptionen mit Beispielen aus deutschen Dialekten. *Zeitschrift für Dialektologie und Linguistik*, LIII(2), pp. 158–172.

Bürkle (1986) Zur Validität eines Maßes zur Reliabilitätsbestimmung Phonetisch-segmenteller Transkriptionen. *Zeitschrift für Dialektologie und Linguistik*, pp. 173–181.

Cucchiarini C. (1993) *Phonetic Transcription: A Methodological and Emperical Study*. Ph.D. thesis, Katholieke Universiteit Nijmegen, Nijmegen.

Daan J., Blok D. P. (1969) *Van Randstad tot Landrand; toelichting bij de kaart: Dialecten en Naamkunde*. Noord-Hollandsche Uitgevers Maatschappij, Amsterdam.

Goossens J. (1965) *Die niederländische Strukturgeographie und die "Reeks Nederlandse Dialectatlassen"*. Bijdragen en mededelingen der Dialectencommissie van de Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam. N.V. Noord-Hollandsche Uitgevers Maatschappij, Amsterdam.

Heeringa W. (2001) De selectie en digitalisatie van dialecten en woorden uit de Reeks Nederlandse Dialectatlassen. *TABU: Bulletin voor taalwetenschap*, 31(1/2), pp. 61–103.

Hoppenbrouwers C., Hoppenbrouwers G. (2001) *De indeling van de Nederlandse streektalen. Dialecten van 156 steden en dorpen geklasseerd volgens de FFM*. Koninklijke Van Gorcum B.V., Assen.

IPA (1999) *Handbook of the International Phonetic Association: A Guide to the Use of the International Phonetic Alphabet*. Cambrige University Press, Cambrige.

Jain A. K., Dubes R. C. (1988) *Algorithms for Clustering Data*. Englewood Cliffs, Prentice Hall, New Yersey.

Kessler B. (1995) Computational Dialectology in Irish Gaelic. In *Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics*, EACL, Dublin, pp. 60–67.

Ladefoged P., Maddieson I. (1996) *The Sounds of the World's Languages*. Blackwell, Oxford etc.

Nerbonne J., Heeringa W. (1997) Measuring Dialect Distance Phonetically. In Coleman, J. (ed.), *Workshop on Computational Phonology*, Special Interest Group of the Association for Computational Linguistics, Madrid, pp. 11–18.

Nerbonne J., Heeringa W. (1998) Computationele vergelijking en classificatie van dialecten. *Taal en Tongval, Tijdschrift voor Dialectologie*, 50(2), pp. 164–193.

Nerbonne J., Heeringa W. (2001) Computational Comparison and Classification of Dialects. *Dialectologia et Geolinguistica. Journal of the International Society for Dialectology and Geolinguistics*, 2001(9), pp. 69–83.

Nerbonne J., Heeringa W., Kleiweg P. (1999a) Comparison and Classifications of Dialects. In *Proceedings of the 9th Meeting of the European Chapter of the Association for Computational Linguistics*, Bergen, pp. 281–282.

Nerbonne J., Heeringa W., Kleiweg P. (1999b) Edit Distance and Dialect Proximity. In Sankoff, D. and Kruskal, J. (eds.), *Time Warps, String Edits, and Macro Molecules; The Theory and Practice of Sequence Comparison*, 2nd edition, CSLI, Stanford, pp. v–xv.

Nerbonne J., Heeringa W., van den Hout E., van der Kooi P., Otten S., van de Vis W. (1996) Phonetic Distance between Dutch Dialects. In Durieux, G., Daelemans, W., and Gillis, S. (eds.), *CLIN VI, Papers from the Sixth CLIN Meeting*, University of Antwerp, Center for Dutch Language and Speech (UIA), Antwerp, pp. 185–202.

Te Winkel J. (1901) *Geschiedenis der Nederlandsche taal*. Blom & Olivierse, Culemborg, naar de tweede Hoogduitsche uitgave met toestemming van den schrijver vertaald door Dr. F. C. Wieder. Met eene Kaart.

Vieregge W. H., Rietveld A. C. M., Jansen C. I. E. (1984) A Distinctive Feature Based System for the Evaluation of Segmental Transcription in Dutch. In van den Broecke, M. P. R. and Cohen, A. (eds.), *Proceedings of the 10th International Congress of Phonetic Sciences*, Foris Publications, Dordrecht and Cinnaminson, pp. 654–659.