# Measuring Dialect Pronunciation Differences using Levenshtein Distance

Wilbert Heeringa

The research has been carried out at the University of Groningen, Faculty of Arts, the Humanities Computing department.

The work in this thesis has been carried out under the auspices of the Behavioral and Cognitive Neurosciences (BCN) research school, Groningen.

Groningen Dissertations in Linguistics 46
ISSN 0928-0030

Rijksuniversiteit Groningen

# Measuring Dialect Pronunciation Differences using Levenshtein Distance

Proefschrift

ter verkrijging van het doctoraat in de
Letteren
aan de Rijksuniversiteit Groningen
op gezag van de
Rector Magnificus, dr. F. Zwarts,
in het openbaar te verdedigen op
donderdag 8 januari 2004
om 14.15 uur

door

Wilbert Jan Heeringa

geboren op 2 augustus 1970
te Groningen

| | | |
|---|---|---|
| Promotor | : | Prof. dr. ir. J. Nerbonne |
| Beoordelingscommissie | : | Prof. dr. H. Goebl<br>Prof. dr. H. Niebaum<br>Prof. dr. G. De Schutter |

# Acknowledgements

This thesis is attributed to exactly one author as can be seen on both the cover and the title pages. The author is the one who is responsible for the content. But it should be emphasized that many people contributed to the coming about of this thesis. I would like to mention them.

First of all, I thank my promotor, John Nerbonne. More than five years ago, he encouraged me to work on the dialectometry project, and during the project he gave invaluable support. Without his support this thesis would never have been published.

Probably just the pictures in this thesis will catch the eye of the reader. While I implemented programs for calculating distances between language varieties and for clustering them, Peter Kleiweg developed software for creating dendrograms, multidimensional scaling plots and different types of (color) maps. I am grateful to Peter for developing and making available this excellent software, and for his extensive help when creating the figures.

During a visit on a cloudy afternoon, one of my best friends, Martin de Vries suggested that one seek speech segment distances on the basis of an acoustic representation. Becoming the inventor of a new type of voice-producing prosthesis, this approach seemed obvious to him. I thank him for this valuable suggestion.

In cooperation with Roberto Bolognesi I worked on the comparison of Sardinian dialects. In this small project, the use of acoustic segment distances was developed and the use of the Levenshtein distance was improved. I thank Roberto for his help and his friendly cooperation.

In the field of phonetics and phonology I got the help of many persons. I thank David Weenink, Dicky Gilbers, Vincent van Heuven, Wouter Jansen, Angelika Braun, Tjeerd de Graaf and Christine Siedle for explanation and advices. I thank Paul Boersma and David Weenink for making available their excellent PRAAT program.

Norwegian dialects play an important role in this thesis. Jørn Almberg made Norwegian recordings and transcriptions of the fable 'The North Wind and the Sun'. I am grateful to him for this his permission to use this material and for his help during the whole investigation. I thank Charlotte Gooskens for the good cooperation in our Norwegian research, and for her permission to use the results of her perception experiment in Norway. Thanks are due to Charlotte Gooskens and Sabine Rosenhart for cutting the Norwegian word samples. I thank Arnold Dalen for his help in finding a reliable dialect map and for classifying each of the Norwegian varieties in the right dialect group in accordance with this traditional dialect map.

When exploring the Dutch dialect area, it is important to do this on the basis of a well-chosen network of sites. I thank Jo Daan, Arjen Versloot (Friesland), Sybren Dyk (Friesland), G. H. Kocks (Drenthe), Jacques van Keymeulen (French, West, East and Zealand Flanders) and Joep Kruijsen (Brabant and Limburg) for their useful advice when selecting the varieties. Once the dialects were chosen, the transcriptions had to be digitized. I owe a lot to Saakje van Dellen, Rogier Nieuweboer, Marcus Bergman and Johan Dijkhuis who did the greater part of this donkey work.

Now and then I asked Cor Hoppenbrouwers some questions about the interesting book 'De indeling van de Nederlandse streektalen' which he wrote in cooperation with his brother. I thank him for patiently answering all my questions. I thank Henk Kiers for his help with the Mantel test. I thank Frits Steenhuisen for some help in geography.

Many people help me prevent my English becoming too much like Dutch. I thank John Nerbonne, Jennifer Spenader, James Hammerton and Menno van Zaanen for making corrections. Remaining errors are for my own responsibility! Furthermore, Menno van Zaanen gave also some useful remarks as regards content. Thanks!

Special thanks are due to the members of the reading committee: Hans Goebl, Hermann Niebaum and G. de Schutter. During the hot summer they read the manuscript. Especially I thank Hermann Niebaum for his valuable comments, which I used profitably in this thesis.

I thank all who contributed to this thesis anyway. In the department of Alfa-Informatica, sometimes I felt like a dwarf among giants metaphorically speaking. I am grateful to all colleagues for their collegiality. Furthermore, I thank my two paranimfs for rendering assistance.

Last but not least I thank my parents for their empathy during the more than five years of this project.

In the center of the coat of arms of the university of Groningen, an open Bible is shown. On the left page the words *Verbum Dni* are found, and on the right page the word *lucerna* is printed. This text is a short reproduction for *Verbum Domini lucerna pedibus nostris*: 'the Word of the Lord is a lamp unto my feet' (Psalms 119:105). In the light of this lamp I confess that language variation is a gift of the Creator. It is my wish that this thesis will help scholars in further research of language variation.

# Contents

# Chapter 1

# Introduction

## 1.1 Motivation

Dialect speakers are aware that there exist borders in the dialect landscape. This is reflected in the dialect map of Daan and Blok (1969). In the Netherlandic part of this map dialect borders are found on the basis of the arrow method. Dialects which speakers judge to be similar are connected by arrows. Bare strips, where no arrows are placed, show dialect area borders.

In Chambers and Trudgill (1998, p. 5) the dialect landscape is described from the perspective of a traveler:

> "If we travel from village to village, in a particular direction, we notice linguistic differences which distinguish one village from another. Sometimes the differences will be larger, and sometimes smaller, but they will be *cumulative*. The further we get from our starting point, the larger the differences will become."

For the most part the villagers of two successive villages will understand each other's dialects very well, but the longer the chain, the greater the chance that dialects on the outer edges of the geographical area are not mutually intelligible.

> "At no point is there a complete break such that geographically adjacent dialects are not mutually intelligible, but the cumulative effect of the linguistic differences will be such that the greater the geographical separation, the greater the difficulty of comprehension."

We illustrate this by a set of 27 dialects, found on a straight line from the northeast to the southwest in the Dutch language area. The locations of the dialects are shown in Figure 1.1. For each variety pronunciations for the words *wijn* 'wine', *potten* 'pots' and *deur* 'door' are given in Figure 1.2. The transcriptions are taken from the *Reeks Nederlandse Dialectatlassen* (Blancquaert and

Peé, 1925–1982). Assume we travel from Scheemda to Bellegem. In Scheemda we notice that the three words are pronounced as [ʋin], [pɔtn̩] and [døːrə]. These pronunciations remain about the same until we arrive in Putten. In this location *deur* is pronounced as [døˑᵊr]. The final [ə] is lost. Going one location further, we arrive in Amersfoort. In Amersfoort *potten* is pronounced as [pɔtə]. The final [n̩] is replaced by a schwa. Traveling further we find that *wijn* is pronounced as [wa̞ᶦn] in the variety of Driebergen. The monophthong [i] is replaced by the diphthong [a̞ᶦ]. In the following locations of Vianen and Hardinxveld the strongly related diphthongs [a̞ˑⁱ] and [aⁱ] are used, but in Zevenbergen the diphthong [ɛⁱ] is used. In Oudenbosch and Roosendaal the strongly related diphthongs [æⁱ] and [ɛˑⁱ] are used. However in Ossendrecht we find the monophthong [ɛˑ]. In the subsequent locations the same or a strongly related sound is found. When we arrive in Moerbeke, we notice that *potten* is pronounced as [pɔtn]. The final [ə] is replaced by [n].

During our travel, other small changes are also observed, but in the description which we gave here we focused on the more systematic ones. We found that systematic changes did not appear in the three words simultaneously, but the changes are found at different places in the chain. It is not the case that the landscape is divided perfectly into areas of nearly homogeneous speech habits, instead the dialect landscape we study may be regarded as a continuum. Both Bloomfield (1933, p. 51) and Chambers and Trudgill (1998, p. 5) mentioned that differences accumulate if one travels in any one direction. This may be the perception of the traveler, but in reality this is not always the case. For example *potten* is pronounced as [pɔtn̩] in the Northeast, as [pɔtə] in the middle, but as [pɔtn] in the Southwest again.[1]

On the one hand, dialect speakers find borders, but on the other hand, the traveler finds a continuum with gradual transitions which are sometimes larger and sometimes smaller. The one does not necessarily exclude the other. Dialect borders bound the dialect continuum which was investigated by the traveler. Bloomfield (1933, p. 51) defines a *dialect area* as a "geographic area of gradual transitions." Chambers and Trudgill (1998, p. 7) call such an area a *geographic dialect continuum*.[2]

Considering the Netherlandic part of the map in Daan and Blok (1969), we find that this small area is divided into no less than 20 dialect areas or dialect continua (approximately 40,000 km$^2$). In Chambers and Trudgill (1998, p. 6) a map of Europe is given. In this map Europe is divided into five dialect continua: the Scandinavian dialect continuum, the West Germanic dialect continuum, the West Romance dialect continuum, the North Slavic dialect continuum and the

[1] The final [n̩] in the Northeast and the final [n] in the Southwest are probably different notations of different transcribers of the same phenomenon.

[2] An investigation to the relation between dialect areas and dialect continua on the one hand and geography on the other hand with the Chambers-Trudgill traveler as starting point can be found in Heeringa and Nerbonne (2001).

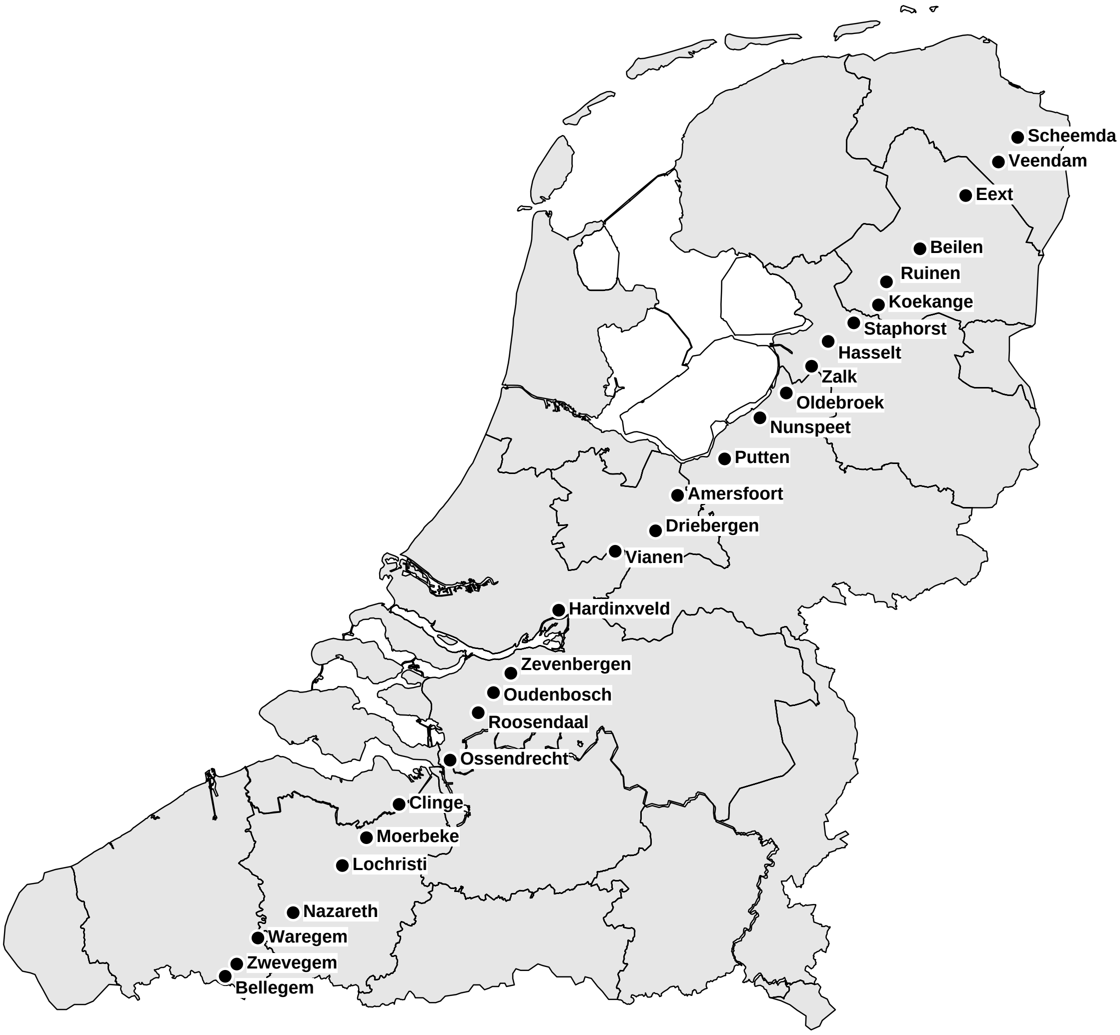


Figure 1.1: Locations of 27 Dutch dialects which may be visited by a traveler who walks from the northeast to the southwest.

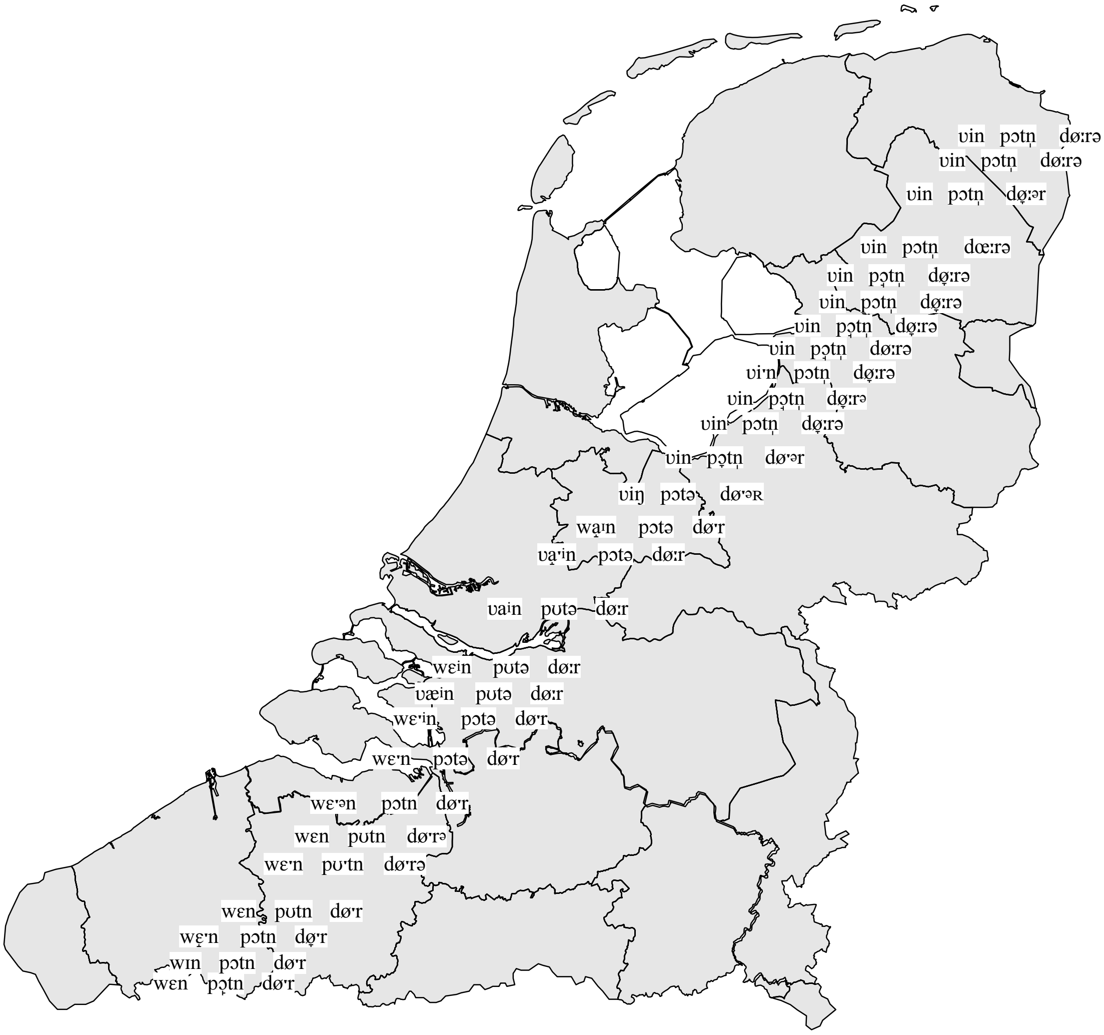


Figure 1.2: Variation of the words *deur* 'door', *potten* 'pots' and *wijn* 'wine' as perceived by a traveler who starts in the northeast and ends in the southwest. Extra-short sounds are noted in superscript. The transcriptions correspond with the locations in Figure 1.1.

South Slavic dialect continuum. Compared to this European map, the borders on the Dutch map mark weak differences, and the Netherlandic area is just a small piece of the larger West Germanic dialect continuum. The comparison of the two maps shows that the meaning of the terms *border* and *continuum* depends on the degree of detail in which the dialect landscape is investigated. Dialect speakers themselves will be sensitive to relatively small differences, while a 'foreign' traveler may regard the dialect landscape more globally.

In this thesis we present a method for finding dialect borders and exploring dialect continua for any given degree of detail. For this purpose we need a 'ruler' with which the linguistic distances between any pair of dialects can be measured in an objective way. The first to develop a method of measuring dialect distances was Jean Séguy, assisted and inspired by Henri Guiter. Séguy and his associates published six volumes of the *Atlas linguistique de la Gascogne*. Using the data in this atlas, Séguy and his research team counted "the number of items on which the neighbors *disagreed*" for each pair of contiguous sites. The number of disagreements between two neighbors was expressed as a percentage. This percentage represented the linguistic distance between two varieties (Chambers and Trudgill, 1998, p. 138).

At about the same time Hans Goebl worked on methods for measuring dialect distances which are strongly related to the methodology of Séguy. The basis of the work of Goebl was developed mainly independent of Séguy, and can be characterized by three innovations. First, Goebl searched for close connection to the international numerical classification. Second, methods from the field of geography and cartography were taken into account. Third, starting with a consistent setup of theoretical fundamental questions about classification, data compression and typology Goebl came to real philosophical questions. Especially the issue of data compression is a matter of major concern for interdisciplinary research. More about the work of Goebl can be found in Goebl (1982, 1984, 1993, 2002).

In 1995 Kessler used the *Levenshtein distance* for finding linguistic distances between dialects (Kessler, 1995). The Levenshtein distance is a sensitive measure with which distances between strings (in this case transcriptions of word pronunciations) are calculated. The algorithm finds the cost of the least expensive set of insertions, deletions or substitutions that would be needed to transform one string into the other (Kruskal, 1999). Kessler applied this measure successfully to Irish Gaelic. Due to its sensitivity we find the Levenshtein distance promising and use this measure as well. In our research we improved the method further. The goal of this thesis is to show that the Levenshtein distance is a useful tool for measuring dialect word pronunciation distances, and thus for measuring dialect distances. We will show different ways in which the Levenshtein distance can be refined, validate the method and apply it to different data sets.

## 1.2 Overview

In Chapter 2 we give an overview of the main methods for showing geographical distribution patterns. They can be divided into traditional methods, perceptual methods and computational methods. In the section about computational methods (Section 2.3) we discuss among others the *corpus frequency method* (developed by Hoppenbrouwers and Hoppenbrouwers (2001)), the *frequency per word method* and, of course, the Levenshtein distance. This range of methods reflects different steps of improvement. The corpus frequency method treats a list of words simply as a set which contains a large number of segments. The method does not distinguish between different words and does not consider different segment orders. The frequency per word method distinguishes different words but still does not consider the order of segments in a word. The Levenshtein distance distinguishes different words *and* takes the order of segments in a word into account. Although the Levenshtein distance is the focus of this thesis, the two other methods will be involved indirectly.

In our research dialects are compared on the basis of word pronunciations. A word pronunciation consists of the concatenation of speech segments. When attempting to quantify distances in pronunciation between dialects, we need to base our measurements on the relations among different speech segments. In Chapter 3 these relations are found on the basis of discrete representations. First we discuss a representation where speech segments are simply equal or not equal, excluding graduations. Second we discuss the use of feature descriptions of phoneticians and phonologists from which we derive finer segment distances. The different discrete representations are used for all of the three computational comparison methods we mentioned above. In Chapter 4 the relations among segments are determined on the basis of acoustic representations. Acoustic representations cannot be used for frequency-based methods, so we used them in the Levenshtein distance only.

In Chapter 5 Levenshtein distance is described. First we describe the application of Levenshtein distance to transcriptions of word pronunciations. When using transcriptions the segment distances are used as determined in the Chapters 3 and 4. Second we explain the application of Levenshtein distance to acoustic recordings of word pronunciations. When using recordings of words a transcription is only used for finding the number of segments per word. The segment distances as measured in the Chapters 3 and 4 are not used.

Once the distances between dialects are calculated, the varieties can be classified. Classification results show relations between elements in a way which is easy to understand. Different classification techniques are discussed in Chapter 6. First we discuss cluster analysis, the result of which perfectly agrees with the idea that the dialect landscape can be divided by borders. The result is a dendrogram, a hierarchically structured tree in which the varieties are the leaves. In this tree for *each* degree of detail the number of groups can be found. The groups can be drawn on a geographic map. Second we discuss multidimensional scaling. The result of this technique is a plot, where the geographic distance between kindred varieties is small, and between different dialects great. On the basis of multidimensional scaling results a map can also be made in which each dialect has its own unique color and in which color contrasts represent linguistic differences. This type of representation perfectly agrees with the idea of the traveler who has traversed the dialect continuum, perceiving sometimes larger and sometimes smaller differences.

Using different computational comparison methods on the basis of different segment representations, the question arises which methods are most suitable in general. In Chapter 7 different versions of frequency-based methods and the Levenshtein distance are validated by applying them to a small set of 15 Norwegian varieties and comparing their results with the judgments which are given by the dialect speakers themselves. Subsequently, the method which appears to be the best method is applied to a larger set of 55 Norwegian varieties in Chapter 8. On the basis of distances which are found with this method we apply cluster analysis and multidimensional scaling. The results are compared to the traditional map of Skjekkeland (1997). In Chapter 9 the same computational comparison method is applied to a set of 360 Dutch dialects. First the distances between the varieties are calculated, and on the basis of these distances cluster analysis is applied and multidimensional scaling is performed. The results are compared to the map of Daan and Blok (1969). Second the varieties are compared to Standard Dutch and a ranking of differences with respect to Standard Dutch is given. Finally conclusions are drawn and future prospects are given in Chapter 10.

# Chapter 2

# Overview of methods in dialectology

The awareness of the existence of different dialect areas dates at least since the Middle Ages, as appears from an example cited by Niebaum and Macha (1999, p. 76). About 1300 the Franconian Hugo von Trimsberg mentioned in his didactic poem "Der Renner" in chapter "Von manigerleie sprâch" (Von Trimberg, 1970, p. 220 ff.) a list of dialect groups. The speakers of the groups are characterized by slogans. However, the oldest known attempts to find dialect divisions in a more scholarly way dates from 1821. In France C. F. Dupin suggested drawing dialect maps in 1814, and in 1821 the first French dialect map was created by Coquebert de Montbret (Weijnen, 1966, p. 188). In the same period in Germany J. A. Schmeller published a dialect map as a résumé of his grammatical description of the "Mundarten Bayerns" (Niebaum and Macha, 1999, pp. 52–54).

In this chapter, we will give a brief overview of the main methods for showing geographical distribution patterns. We divided them in traditional methods (Section 2.1), perceptual methods (Section 2.2) and computational methods (Section 2.3). We do not pretend to give a complete overview, but just give some outlines to locate our research within the scholarly field. For more details we refer to Weijnen (1966), Goossens (1977), Inoue (1996a), Inoue (1996b), Chambers and Trudgill (1998), Niebaum and Macha (1999) and Hoppenbrouwers and Hoppenbrouwers (2001). At the end of this chapter we account for our decision to use the Levenshtein method (Section 2.4). This method is the central theme in this thesis.

## 2.1 Traditional methods

### 2.1.1 Tribes and intuition

The oldest dialect classifications were based on knowledge about dialectal contrasts and intuition, and tried to demonstrate a connection with early tribal history. The Dutch language area could be divided into Frisian, Saxon and Franconian, a division given by Winkler (1874). Transition areas are also identified. Following the proposals of Winkler, Jellinghaus (1892) created a map in which dialect areas are separated by lines. Similar maps were published by Te Winkel (1901), Van Ginneken (1913) and Lecoutere and Grootaers (1926), in which the different dialect areas were given different colors. The color distinctions give a visual representation of the borders between different dialect areas. Therefore, Goossens classifies the maps just mentioned under the 'plane method'. However, this is not helpful since this term points to the visualization of the classification, not to the classification method itself. We agree with Hoppenbrouwers and Hoppenbrouwers (2001) who order these maps under 'tribal divisions'.

### 2.1.2 The isogloss method

In the field of meteorology *isotherms* play an important role. An isotherm is a line on a map connecting places having the same temperature at a given time or on average over a given period (OUP, 1998). Using an idea similar to isotherms, the field of geolinguistics uses *isoglosses*. An isogloss is a line on a map dividing areas whose dialects differ in some specific respect (Matthews, 1997). The equivalents of 'chicken' in the Dutch language area are a good example of a lexical isogloss. In the west and midland areas of the Netherlands, the dominant pronunciation is [kɪpə] (or something related), but in the east along the border with Germany the word is [hundər] or something related. An example of a pronunciation isogloss can be found in the pronunciation of the final syllable in the Dutch word *dopen* 'to baptize', which is pronounced as [dopm̩] in the northeastern part of the Netherlands and the western part of Flemish-speaking Belgium, and as [dopə] in the intervening area and in Frisian (the northwest of the Netherlands). Using the isogloss method, isoglosses of different phenomena are drawn on a map. Coinciding isoglosses are interpreted as borders. The two main Dutch isogloss maps were made by Weijnen, where the first is published in Weijnen (1941) and the second in both Weijnen (1958) and Weijnen (1966).

The advantage of an isogloss map is that it shows verifiable facts. However Goossens (1977) mentioned that the isogloss method cannot be applied without making subjective choices. This fact is described in more detail by Kessler (1995) who mentioned three problems when trying to find dialect areas on the basis of isoglosses. First isoglosses do not always coincide. They can be parallel, forming vague bundles, or even cross each other, describing contradictory binary divisions.

In this connection we mention the famous *Rhenish fan*, as described by Bloomfield (1933, pp. 343–345).[1] Features separating Low German and High German form nearly coincident isoglosses for much of their length, but then they diverge at the Rhine valley (see also Chambers and Trudgill (1998)). In practice, well-known isoglosses which form bundles are selected, but this makes the method subjective. A second problem Kessler mentioned is that many isoglosses do not neatly bisect the language area. Often variants do not neatly line up on two sides of a line, but are intermixed to some degree. Furthermore, information may be lacking for some sites, or the question is not applicable. Kessler illustrates this by an example. "When comparing how variuos sites pronounce the first consonant of a particular word, it is meaningless to ask that question if the site does not use that word." The third problem which Kessler pointed out is the fact that in case of a dialect continuum with very gradual changes, it seems arbitrary to draw major dialect boundaries between two villages with very similar speech patterns. Most languages have dialect continua.

### 2.1.3 The structure geographic method

A language area can be divided in dialect areas on the basis of structure geographical data. Dialects with the same phoneme inventory form a dialect area. So each dialect area is characterized by its own phoneme inventory. Structure geographic classifications can also be made by lexical, syntactic or morphological data. Until now, the structure geographic method has only been used for smaller areas. Several examples of classifications on the basis of especially phoneme structures exist. Moulton (1960) classified dialects in northern Switserland on the basis of short vowel systems. In 1960 Wortmann investigated the development of the Middle Low German *ê* and *ô* sounds in the Westphalian area. On the basis of this research Foerste (1960) made a structural phonologic classification of the Westphalian dialects. A corresponding map is also given by Niebaum and Macha (1999, p. 83). Heeroma (1961) published a map in which the northeastern part of the Netherlands is divided on the basis of systems of the long vowels from the *aa* and *ie* series. Goossens (1965) applied the structure geographic method to material from the *Reeks Nederlandse Dialectatlassen* (RND) (Blancquaert and Peé, 1925–1982), a series of atlasses covering the Dutch dialect area (The Netherlands, north Belgium, northwest France and the German county Bentheim) (see Section 9.1). In 1965 only the RND parts covering the northwestern and the southern part were finished. Goossens investigated whether it is possible to find the phoneme system of a dialect on the basis of the corresponding transcription in the RND. For a west Flemish dialect, a west Brabant dialect, an east Brabant dialect, a west Limburg dialect and an east Limburg dialect he made a matrix where the rows represent different short vowel segments as found in the RND

[1] See Niebaum and Macha (1999, pp. 100–101) for a clearer visualization of the *Rhenish fan*.

transcription, and the columns short vowel phonemes as given in literature about that dialect. In the matrix for each segment-phoneme pair the number of times that the segment in the transcription is noted as the phoneme in the literature is given. Goossens concluded that the RND transcriptions form mostly suitable material for the use of the structure geographic method. Furthermore, Goossens divided the central dialects in the southern part of the Dutch language area on the basis of different vowel inventories. Only the /i/, /iː/, /ɪ/, /ɛ/, /æ/ and /ɑ/ were considered (see p. 30). He found a division in south Brabant, northwest Brabant, east Flemish and Zeeland dialect groups.

Goossens (1977, p. 169) pointed out that differences in phoneme inventories do not form sufficient information for finding dialect areas. Different dialects may have the same phoneme inventory. Kocks (1970) was also faced with this problem when he classified dialects in and around the southeastern part of the Dutch province of Drenthe on the basis of phoneme inventories. His solution was to use the frequencies of phonemes, found on the basis of translations of 163 words which he retrieved for several places. Actually he applied the phone frequency method, which we discuss in Section 2.3.2.

## 2.2 Perceptual methods

### 2.2.1 The arrow method

In 1939, the Department of Dialects of the Royal Netherlands Academy of Sciences and Letters in Amsterdam, which has about 1500 correspondents in all parts of the county, held a survey in which the following questions were asked:

1. In which nearby location(s) do people speak the same or nearly the same dialects as yours?
2. In which nearby location(s) is it absolutely certain that a dialect different from yours is spoken? Could you mention some deviations?

In 1946 Weijnen published a map which was constructed on the basis of the first question in this survey. On the map, places in which, according to the speakers, (nearly) the same dialects are spoken are connected by arrows. In that way, white strips arise where there are no arrows; these are the dialect borders (Weijnen, 1966). This approach is called the *arrow method* and aims to find dialect areas and borders on the basis of the language awareness of the dialect speakers.

Later, on the basis of the same survey, an arrow map was published for the Netherlands by Rensink (1955). For this map as well only the first question was used. Rensink stressed that the map should be regarded as a temporary result.

A definitive map, based on the same survey, the same question and the same area was published by Daan and Blok (1969). To cover the complete Dutch language area, the Flemish part of Belgium was also included. However, because the Belgian dialectologists did not have such a large group of correspondents at their disposal, a different procedure was applied. Language geographers who often belonged to dialect-speaking groups themselves were consulted. According to Daan and Blok this gave sufficient certainty that in this region the experience of the dialect speaker was properly expressed too. It was convenient that the South-Netherlandic dialects are usually regarded as more homogeneous than those of the North (p. 43).

In the map of Daan each dialect area has its own unique color. The colors are more or less intuitively chosen, but fit with the tribal division into Frisian (blue), Saxon (green) and Franconian (white, yellow, orange, red). In the (nearly) white area the dialects which are closest to Standard Dutch are found. The choice of the colors corresponds to a gradually increasing divergence from Standard Dutch.

Sometimes the user of the arrow method had to correct the results. According to Goossens (1977) this means that the designer did not trust his or her own method. Indeed the designer made corrections (see p. 31 of Daan and Blok (1969)). However, these are made:

1. in case of a very low response of correspondents for an area, e.g. Drenthe;
2. in case of contradictory responses, i.e. speakers at location A judge the dialect at location B as the same, but not vice versa.

In these cases dialectologists were consulted, tape recordings were examined or literature was consulted. If none of this was possible, the designer personally went to the area to find the right border. The fact that corrections led to consulting expert opinion rather than further subjective judgments suggests that the latter were regarded as general indications.

A disadvantage of the arrow method is that the method cannot be used for comparing dialect areas which are clearly related but do not border on each other. Such situations exist as a result of migration or emigration.

### 2.2.2 Perception experiments

Distances between varieties can be obtained on the basis of a perception experiment. Gooskens (1997) investigated perceptual distances between Standard Dutch and some Dutch varieties (some dialects and standard Flemish), focusing on the verbal level and the prosodic level. Subjects listen to a series of fragments and rate the similarity with respect to Standard Dutch with a number between 1 and 10, where 1=language variety in question and 10=Standard Dutch.

Just as perceptual distances of varieties with respect to a standard language can be obtained, mutual perceptual distances between varieties can be measured.

This is showed by Gooskens (2002) on the basis of 15 Norwegian dialects. In each of the 15 places listeners listen to fragments of each of the 15 varieties. While listening to the dialects the listeners were asked to judge each of the 15 dialects on a scale from 1 (similar to native dialect) to 10 (not similar to native dialect). In this way a perceptual distance matrix is obtained, on the basis of which cluster analysis (see Section 6.1) and multidimensional scaling (see Section 6.2) was performed. The experiment is described in more detail in Section 7.4.1.

## 2.3 Computational methods

### 2.3.1 Counting differences or similarities

Jean Séguy was director of the *Atlas linguistique de la Gascogne.* He and his associates published six atlas volumes. In these volumes maps are published in which single answers were plotted (Chambers and Trudgill, 1998, p. 137). However, Séguy looked for a way to analyse the maps in a more objective way than was possible with traditional analytic methods. For each pair of contiguous sites Séguy and his research team counted "the number of items on which the neighbors *disagreed.*" The number of disagreements between two neighbors was expressed as a percentage, "and the percentage was treated as an index score indicating the *linguistic distance* between any two places" (Chambers and Trudgill, 1998, p. 138).

The items fell into five types: 170 lexical variables, 67 pronunciation, 75 phonetic/phonological, 45 morphological, and 68 syntactic. Séguy weighted all types equally by calculating percentages for each type rather than for each item. The final linguistic distance was calculated as the mean of the five percentages. Séguy and his team calculated the linguistic distances for each item, for each item type and for the composites. They were plotted on maps, which can be found in the last ten pages of the sixth volume of the atlas which was published in 1973.

To make dialect areas more or less visible, Séguy and his associates divided the percentages in four classes: under 13 %, 14-17 %, 18-23 %, and over 23 %. On a interpretive map these classes are represented by respectively unmarked, dotted, light and heavy line-types. "The patterns of lines divided Gascony into regions of greater dialect diversity and regions of relative homogeneity" (Chambers and Trudgill, 1998, p. 140).

Strongly related to the methodology of Séguy is the work of Goebl, although the basis of Goebl's work was developed mainly independent of Séguy. Our description of the work of Goebl is based on Goebl (1982) and Goebl (1993). As data source in his work he used *l'Atlas Linguistique de l'Italie et de la Suisse Méridionale* (AIS) which was compiled by Karl Jaberg and Jakob Jud in about the first quarter of the 20th century. From this atlas he selected 251 varieties and 696 working maps. This means that for each dialect and for each of the 696 items a nominal value is given. 569 working maps represent lexical variation and 127 working maps represent morpho-syntactic variation. Comparable to the way in which Séguy calculated distances, Goebl calculated similarities. The similarity between two variaties based on 696 item pairs as a percentage is calculated as:

$$\frac{\#\text{equal nominal values}}{\#\text{equal nominal values} + \#\text{different nominal values}} \times 100$$

In order to provide different visualizations, distances are also calculated. They are found as the complement of the similarities: $100 - \textit{similarity percentage}$.

On the maps the basic grid consists of 251 polygons, found using the Thiessen geometry, a technique based on the idea of drawing tiles around points so that tiles are as evenly apportioned as possible. In Goebl (1993) three types of maps are shown, namely *choropleth maps*, *interpoint maps* and *beam maps*. A *choropleth map* is a map which is divided into spatial units, and each unit is shaded or colored to the value of a variable for that area. In the work of Goebl choropleth maps are often used for visualizing similarity with respect to a reference variety. *Interpoint maps* visualize distances between neighboring dialects. The darker the 'wall' between two adjacent polygons, the greater the distance between the corresponding varieties. This way of visualizing is called the *honeycomb method* (Inoue, 1996b). The counterpart of this visualization technique is the *beam map*, which is the same method as used by Séguy. Close dialects are connected by darker beams, and more remote ones by lighter beams.

Just as in the work of Séguy and his associates, Goebl's distances are divided into classes. On the maps, each class has its own shade or color. For the division into classes three procedures are mentioned: *MINMWMAX*, *MEDMW* and *MED*. In the MINMWMAX procedure the range from the minimum (MIN) to the mean (German: 'Mittelwert' = MW) is divided in $n$ equally sized classes and the range from the mean to the maximum (MAX) is also divided in $n$ equally sized classes. This gives $n + n$ intervals, based on the percentages of overlap. Using the MEDMW procedure the range from the minimum to the mean is divided in $n$ classes so that each class contains the same number of different percentages. Next the range from the mean to the maximum is also divided in $n$ classes so that each class has the same number of different values. We get $n + n$ intervals again. In the MED procedure the range from minimum to maximum is divided in $n$ classes so that each class contains the same number of different values.

In Goebl's work cluster analysis is also performed on the basis of the distances using complete linkage (see Section 6.1). The 24 most significant groups are drawn on a map, where adjacent polygons of different groups are separated by a dark 'wall', while neighboring polygons of the same group are separated by a light line.

Since 1993 Goebl's group has considerably expanded their empiric foundations with the inclusion of a number of new linguistic atlasses. First the *Atlas linguistique de la France* (ALF) has been entirely dialectometrized. This atlas was compiled by J. Gilliéron and E. Edmont in the period 1902–1920. In Figure 2.1 an interpoint map is shown on the basis of the ALF data. It shows distances between 640 varieties. In Figure 2.2 a choropleth map is given which shows the distances of 640 varieties with respect to Standard French using the ALF data as well.[2] More about the ALF can be found in Goebl (2002). In 1985 Goebl started a project with the goal of compiling a linguistic atlas of Dolomitic Ladinian and neighbouring Dialects. A collaborator of Goebl, Roland Bauer has undertaken the dialectometrization of the first part of the *Atlante linguistico del ladino dolomitico e dei dialetti limitrofi* (ALD-I). This atlas was published by Hans Goebl, Roland Bauer and Edgar Haimerl in 1998.

More about the work of Goebl and his associates can be found at `http://ald.sbg.ac.at/dm/`.

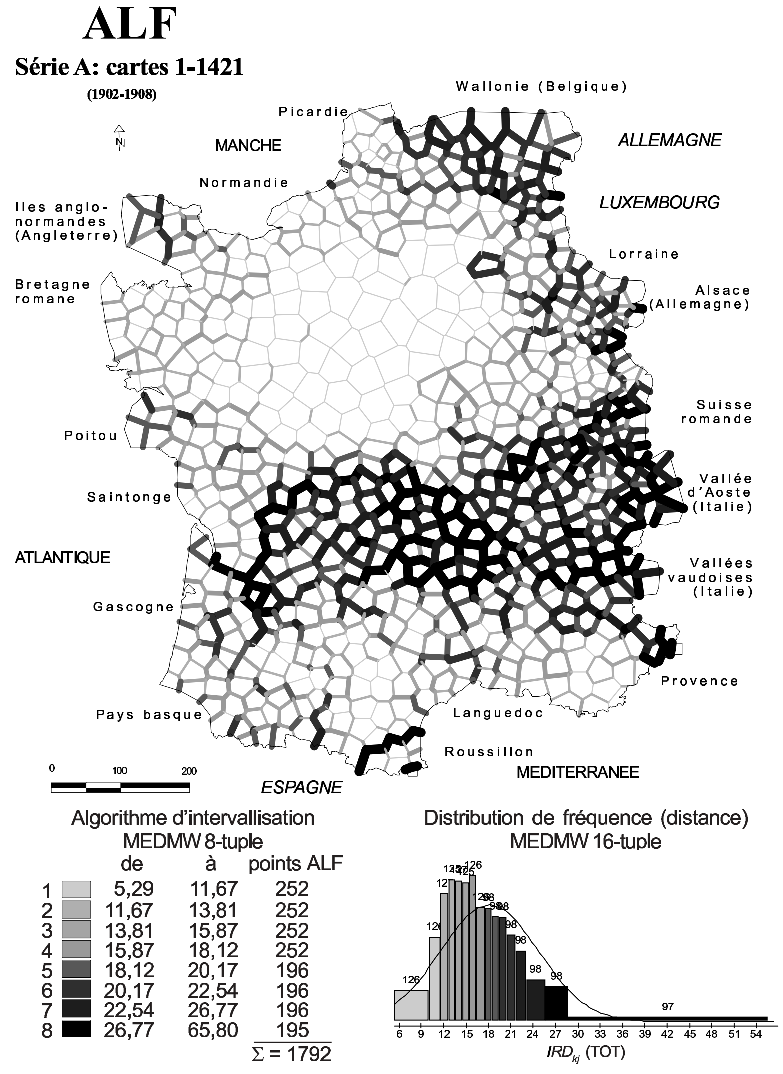
### 2.3.2 Corpus frequency method

In 1988 the Hoppenbrouwers brothers (H & H) introduced the feature frequency method (Hoppenbrouwers and Hoppenbrouwers, 1988). Our description is based on their most mature publication (Hoppenbrouwers and Hoppenbrouwers, 2001). This section is based on an extended analysis of their work (Heeringa, 2002).

The letter frequency method and the phone frequency method are predecessors of the feature frequency method. Using the letter frequency method for each language variety the unigram frequencies of letters are found on the basis of a corpus. Such a corpus is a sample of letters. Since not all samples have the same size, the frequencies should be expressed as percentages. The distance between two languages is equal to the sum of the differences between the corresponding letter frequencies. H & H verify that this approach correctly shows that the distance between Afrikaans and Dutch is smaller than the distance between Afrikaans and the Samoan language. H & H correctly pointed out that different spellings do not always represent different pronunciations (e.g. Dutch *academie* versus Frisian *akademy*), and equal spellings not always represent equal pronunciations (e.g. English *we* versus Dutch *we*), "observations" that show the limits of this approach.

A more phonetically oriented approach is the phone frequency method, in which phonetic texts are used. H & H write (on p. 1) that they started ten years ago with an experiment using texts from *The Principles of the International Phonetic Association* (1949). In this pamphlet for 51 languages a translation of the fable 'The North Wind and the Sun' is given in phonetic (IPA) script. For each text the frequencies of phones are determined. However, this approach also deserves comment. Assume we have three languages with the following phone percentages:

[2] Since color printing is expensive, the Figures 2.1 and 2.2 show black-and-white versions of the color maps which can be found in Goebl (2002) on pp. 40 and 41 respectively.



Algorithme d'intervallisation
MEDMW 8-tuple

| | | de | à | points ALF |
|---|---|---|---|---|
| 1 | ▨ | 5,29 | 11,67 | 252 |
| 2 | ▨ | 11,67 | 13,81 | 252 |
| 3 | ▨ | 13,81 | 15,87 | 252 |
| 4 | ▨ | 15,87 | 18,12 | 252 |
| 5 | ■ | 18,12 | 20,17 | 196 |
| 6 | ■ | 20,17 | 22,54 | 196 |
| 7 | ■ | 22,54 | 26,77 | 196 |
| 8 | ■ | 26,77 | 65,80 | 195 |
| | | | | Σ = 1792 |

Figure 2.1: Example of an interpoint map based on 1687 working maps of the ALF and created by Goebl's research team. Darker and thicker lines separate different varieties, lighter and thinner lines separate more related ones. The 1792 distances are divided in eight classes according to the MEDMW algorithm. Each class has its own thickness and darkness.

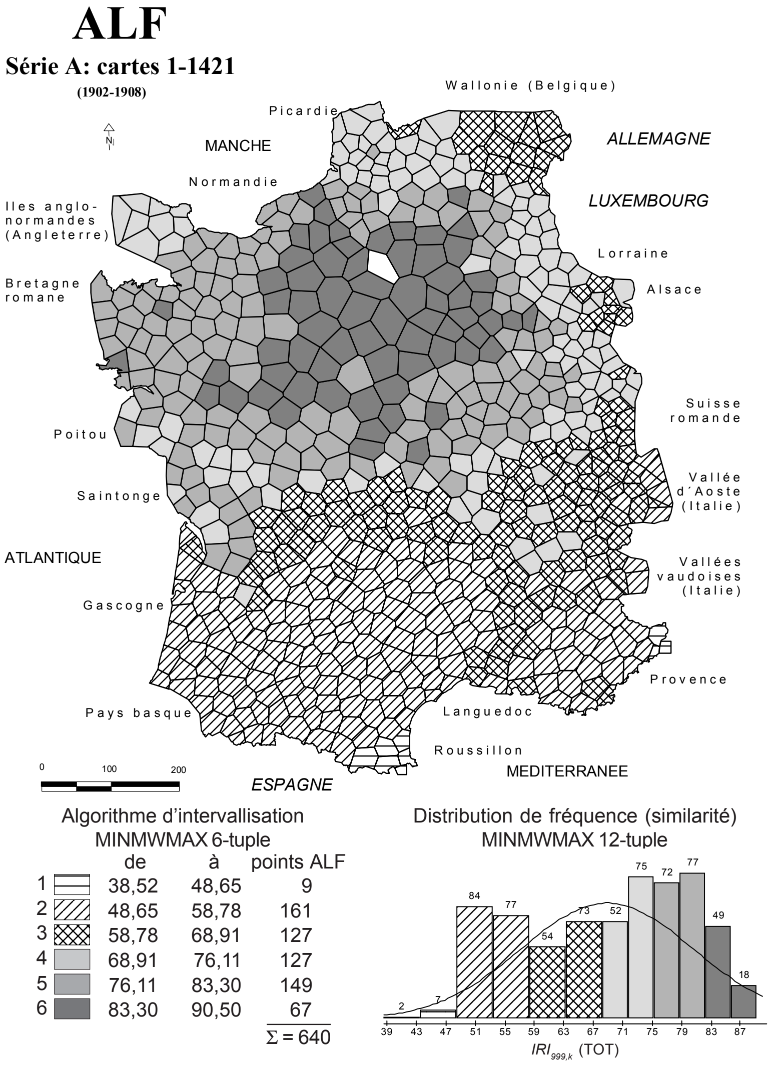


Figure 2.2: Example of an choropleth map based on 1687 working maps of the ALF and created by Goebl's research team. The map shows the distances of 640 varieties with respect to Standard French. The distances are divided in six classes according to the MINMWMAX algorithm. Each class has its own shade.

| | [e] | [ɪ] | [u] |
|---|---|---|---|
| language 1 | 100 % | 0 % | 0 % |
| language 2 | 0 % | 100 % | 0 % |
| language 3 | 0 % | 0 % | 100 % |

Using the phone frequency method there is no basis to conclude that language 1 and language 2 are more related than language 1 and language 3 or language 2 and language 3. H & H wrote that they soon had the insight that a more refined approach was desirable. Therefore, they developed the feature frequency method (FFM).

Speech sounds can be described by a range of distinctive features. For example vowels may be pronounced in front, in the middle or in the back of the oral cavity (described by the features *front* and *back*), or they can be pronounced with the tongue low, central or high (described by the feature *low*), or they can be pronounced with spread or rounded lips (described by the feature *round*). If we have a transcription we can count the number of sounds pronounced in front of the oral cavity, the number of low sounds, or the number of sounds pronounced with rounded lips. In other words: we find the feature frequencies. On the basis of a transcription, the feature frequency method finds the frequencies for the series of features which are fixed in advance. The result is a histogram. The frequencies are expressed as percentages. The distance or similarity between two histograms may be calculated in different ways (see Section 3.6). H & H calculated the similarity by using the Pearson's correlation coefficient.

Finding the frequencies of the features, all speech sounds which can appear in the transcriptions must be defined in terms of features. Therefore, the right features have to be selected. H & H selected *The Sound Pattern of English* (SPE) (Chomsky and Halle, 1968) as starting point, an articulation-based system. H & H applied their method to material from the RND. Therefore, they modified and extended the SPE system so that with the use of this system the RND material is done justice as much as possible. A more detailed description of this feature system can be found in Section 3.1.2.

Once a similarity matrix is obtained, each variety is defined as a vector of similarity values with respect to all other varieties and to itself. Between each pair of two vectors the Euclidean distance can be calculated (see Section 3.6). In that way a distance matrix is obtained. On the basis of this distance matrix cluster analysis was applied, where H & H used average linkage (between groups) (see 6.1).

In the RND for each variety the same 139 sentences have been translated and transcribed in phonetic script. H & H selected 156 varieties. They added Standard Dutch, the dialect of the Amsterdam quarter of the Jordaan, and two adjusted RND transcriptions of Zwolle and Scheveningen. Comparing the H & H results with traditional results, Frisian and Saxon emerge clearly as groups, but the Franconian dialects are split into a Limburg group and a group of remaining dialects. In Figure 2.3 the locations of the 156 varieties are given. In Figure 2.4 for each location which of the ten main groups the variety belongs to can be seen. H & H distinguish between core dialects and edge dialects. On the map core dialects are given in upper case and edge dialects in lower case.

Heeringa (2002) reviews Hoppenbrouwers and Hoppenbrouwers (2001). The review shows that it is possible to build a clone of H & H's computer program 'Polyphon' on the basis of the description given by H & H. With this clone very similar results were obtained.

### 2.3.3 Frequency per word method

A disadvantage of the corpus frequency method is that is does not attach any significance to words. Therefore, the frequency per word method was developed which considers words as separate entities. The frequency per word method was examined in Nerbonne and Heeringa (1998), and later in Nerbonne and Heeringa (2001). With this method two words are compared exactly in the same way H & H compared two corpora. As we saw in Section 2.3.2 the phonetic transcriptions may be compared to each other by comparing histograms of phone frequencies or feature frequencies, where the frequencies are expressed as percentages.

Just as in H & H's work, the frequency per word method was applied to material of the RND. However, rather than using the complete texts (and complete sentences), for each text, a selection of the transcriptions of the same words was made. For each variety a word list was made. When $n$ words are selected, the comparison of two varieties results in $n$ word distances. The dialect distance is found by dividing the sum of the word distances by the number of examined word pairs. In this way we get a distance matrix on the basis of which cluster analysis or multidimensional scaling can be applied (see Chapter 6).

This method was never developed extensively because it is overshadowed by the methodologically superior Levenshtein distance, which we present in Section 2.3.4. However, in validation work it offers the possibility of showing that a word-based approach performs significantly better than a corpus-based approach (see Chapter 7). More details about word based dialect comparison as applied in our research can be found in Chapter 5.

### 2.3.4 Levenshtein distance

A disadvantage of the frequency per word method is that this method is not sensitive to the order of phonetic segments in a word. The better alternative for finding word distances is to use the Levenshtein distance, which considers for each word its sequential structure. In 1995 Kessler introduced the use of the Levenshtein distance as a tool for measuring dialect distances (Kessler, 1995). He applied it successfully to Irish Gaelic.

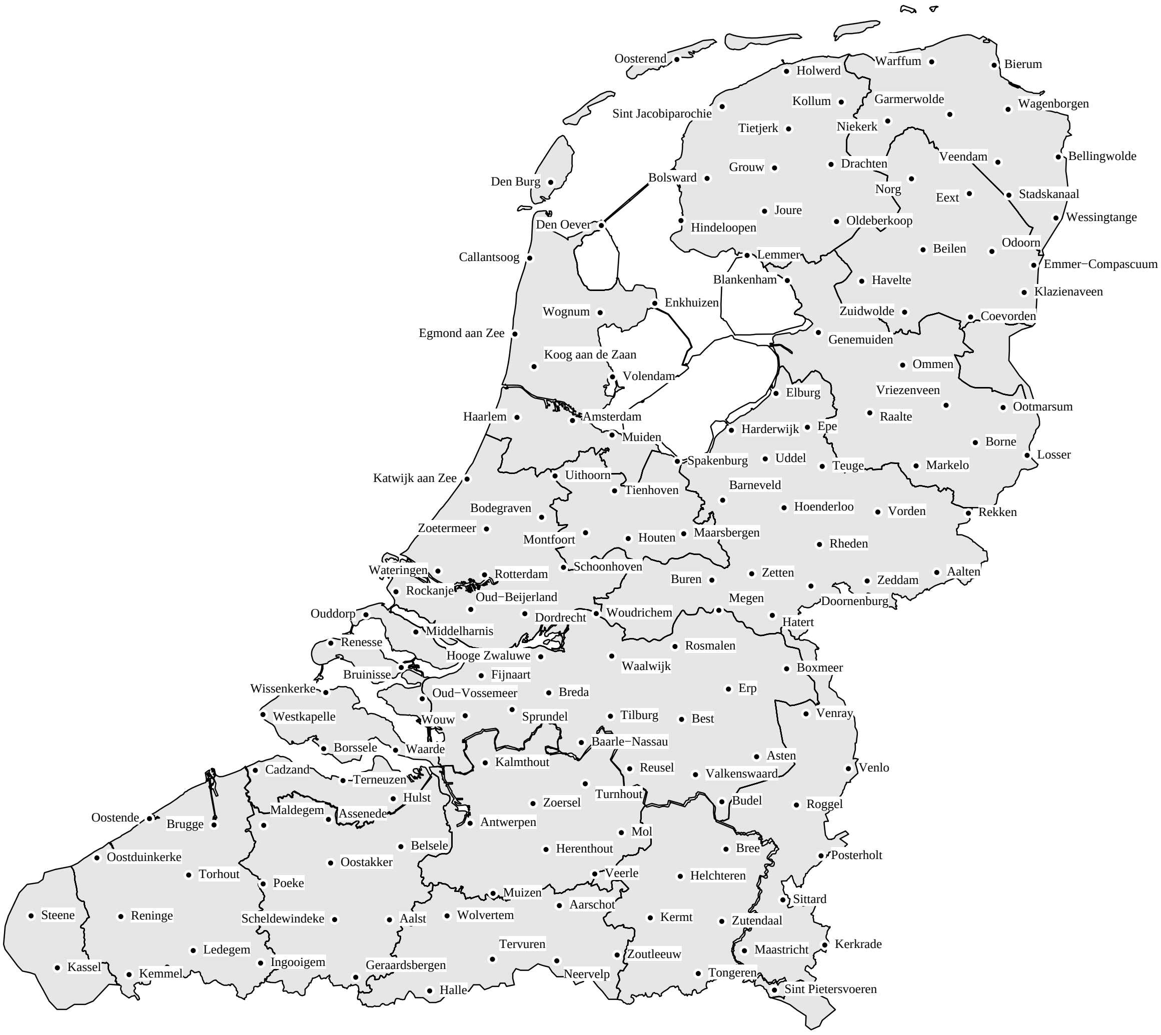


Figure 2.3: The locations of the 156 varieties which were selected from the RND by the Hoppenbrouwers brothers.

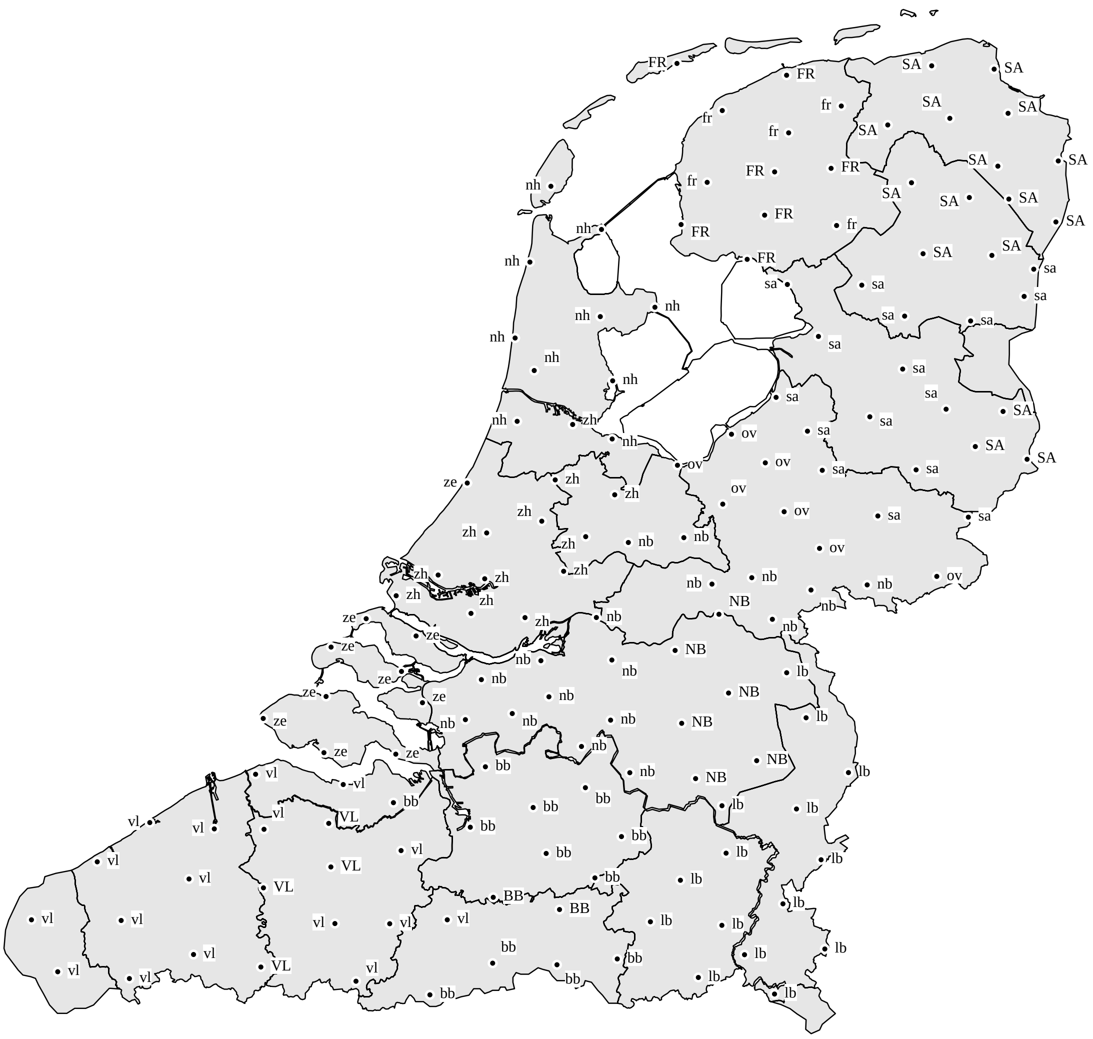


Figure 2.4: The main division of the Dutch dialects according to the feature frequency method of the Hoppenbrouwers brothers distinguishes ten areas: fr=Frisian, sa=Saxon, ov=Overijssel, nh=Noord-Holland, zh=Zuid-Holland, ze=Zeeland, nb=Noord-Brabant, lb=Limburg, bb=Belgian Brabant, vl=Flemish. Core dialects are given in upper case, and peripheral dialects in lower case. The corresponding names of the locations can be found in Figure 2.3.

The Levenshtein distance is a numerical value of the cost of the least expensive set of insertions, deletions or substitutions that would be needed to transform one string into another (Kruskal, 1999). The simplest technique is *phone string comparison*. In this approach all operations have the same cost, e.g., 1. In Kessler's approach when two phones are basically equal but have different diacritics, they are regarded as different phones. So [a] versus [aː] costs 1 unit just as [a] versus [p] costs 1 unit.

In the above technique it is not possible to take into account the affinity between phones that are not equal, but are still related. Methods based on phones will not regard the pair [b,p] as more related than [a,p]. This problem can be solved by replacing each phone by a bundle of features, just as in the feature frequency method. A feature bundle is a range of feature values. For each of the corresponding features a value is given which indicates to what extent that property is instantiated (see Section 2.3.2). Since diacritics influence feature values, they likewise figure in the mapping from transcriptions to feature bundles, and thus automatically figure in calculations of phonetic distance. The resulting metric is called *feature string comparison*.

Using the *phone string comparison* Kessler calculated Levenshtein distances not only when words are phonetic variants of each other, but also when they lexically differ. He called this the *all-word* approach. However, when he used the *feature string comparison*, not only the *all-word* approach was used, but also an approach was used in which the Levenshtein distance is only applied when words are phonetic variants of each other. Kessler called this approach the *same-word* approach.

Kessler applied the Levenshtein distance to data from the *Linguistic Atlas and Survey of Irish Dialects*. This atlas was compiled by Heinrich Wagner and published in 1958. In the atlas the dialect pronunciations are presented in a very narrow phonetic transcription based on the International Phonetic Alphabet. Kessler used 95 varieties and selected 51 concepts for each. Using the Levenshtein distance he calculated the distances between the dialects. On the basis of these distances cluster analyses were performed. The distance between two clusters was calculated as the average distance between all pairs of elements that are in the different clusters. The resulting dialect areas were continuous, aligned with traditional provincial boundaries and agreed with commonly accepted taxonomies. Kessler notes that dialect groupings at narrower levels were unstable, but explained this by the relatively small number of concepts on which the distance metrics were based (51). In this context Kessler refers to Séguy (1973) who cites empirical research suggesting that general dialectometry requires about a hundred concepts. Séguy (1973 and elsewhere) developed dialectometry based on measures of lexical overlap.

The Levenshtein algorithm is the focus of this thesis. An extensive explanation is given in Section 5.1.

### 2.3.5 Gravity center method

A useful method for showing the geographical distribution patterns of dialects is the *gravity center method.* An extensive explanation is given by Inoue (1996b). We will explain it by an example. Assume we know for a number of dialects in the Dutch language area the distance of each dialect with respect to standard Dutch. These distances are then the weights of the dialects. Furthermore, each dialect has geographical coordinates $(x, y)$. Now the gravity center is calculated so that when the survey area rests on a pin at the point of the gravity center, the area will be balanced and remain horizontal. The coordinates for the gravity center are calculated as follows:

$$gravity\ center(x) = \frac{w_1 \times x_1 + ... + w_n \times x_n}{w_1 + ... + w_n}$$

$$gravity\ center(y) = \frac{w_1 \times y_1 + ... + w_n \times y_n}{w_1 + ... + w_n}$$

where $w_1 \ldots w_n$ are the weights and $x_1 \ldots x_n$ and $y_1 \ldots y_n$ refer to the positions of the locations in two dimensions. Now the gravity center may be seen as the center of the Dutch language area. When distances with respect to standard Dutch are given for different words separately, for each word a gravity center can be calculated. Using this, a map of accumulated centers of gravity can be drawn.

## 2.4 Our choice of method

Following Kessler (1995) we used the Levenshtein distance for finding distances between dialects and for finding dialect classifications which are based on the dialect distances. Compared to other methods mentioned in this chapter, the use of the Levenshtein distances has obvious advantages.

The Levenshtein distance is completely objective, and its results are verifiable, an advantage it shares with other computational methods, in contrast to dialect maps based on tribes and intuition (see Section 2.1.1). However, a condition for using Levenshtein is that the data used consists of representative samples of the varieties.

Using the isogloss method, isoglosses cannot simply be added. They are selected so that satisfactory boundaries emerge (see Section 2.1.2), which make this method subjective. However, the Levenshtein distance and other computational methods are able to add differences. This allows one to relate entire varieties, aggregating the atomic differences. None of the differences need to be excluded.

Until now, with the structure geographic method, different dialect areas are characterized by different phoneme inventories and/or different phoneme changes. However, even if frequencies of phonemes are considered, the method is rather insensitive. Words of different dialects may be different, although the phoneme inventories are the same.

The arrow method, as an attempt to process subjective impressions in a objective way (Goossens, 1977), has the shortcoming that only relations between adjacent varieties can be found. E.g., it is not possible to compare varieties of Afrikaans with Dutch varieties. However, when using the Levenshtein distance or other computational methods, such comparisons can easily be made.

Both the arrow method and the use of controlled perception experiments base the classification of dialects on the perception of dialect speakers (see Section 2.2.2). An advantage of perception experiments compared to the arrow method is that perception experiments can compare varieties which do not border on each other. In general the listeners in an experiment judge the distance with respect to their own dialect or standard language (see e.g. Gooskens (1997), Gooskens (2002)). For a listener in a perception experiment it may be much harder to judge the distance between two unknown varieties. However, with the Levenshtein distance and other computational methods the distance for any pair of varieties can be found.

In the methods of Séguy and Goebl the number of (dis)agreements is counted (see Section 2.3.1). With this computational method distances between varieties are found in an objective way. Distances are the aggregate of atomic differences. However, the method is rather rough. Two items are equal or not equal, either lexically, phonologically, morphologically or syntactically. Using Levenshtein, gradual distances between words are found. Lexical, phonological and morphological differences need not be explicitly distinguished, but can be processed with the same algorithm. However, since the algorithm compares word pronunciations, syntactic differences are not processed.

Using the corpus frequency method two varieties are compared by comparing the frequencies of positively marked features of segments in a corpus of the first variety with the frequencies of positive marked features in a corpus of the second variety (see Section 2.3.2). In this method words are not processed as linguistic units. This problem is solved when using the frequency per word method (see Section 2.3.3). However, in both frequency-based approaches the order of segments is ignored. E.g. *it is* may be pronounced as [ɪts] ‘it’s’ in English, and the Dutch equivalent *het is* may be pronounced as [tɪs] ‘t’is’ in Dutch. Using the corpus frequency method or frequency per word method no difference between these two pronunciations is found. However, when using the Levenshtein distance, the order of segments is taken into account.

We conclude that the Levenshtein distance is superior to traditional methods because of its objectivity and sensitivity. Furthermore, the Levenshtein distance does not have the limitation of perceptually-based methods. Compared to previous computational methods, with the Levenshtein distance the data is used most exhaustively. This makes the Levenshtein distance most sensitive. Therefore, in this thesis we focus on the application of the Levenshtein distance in dialectology.

# Chapter 3

# Measuring segment distances discretely

A language variety allows the expression of sentences, which consist of words, which in turn consist of speech segments. When attempting to quantify distances in pronunciation between dialects, we should first make clear how the different speech segments are related to each other. In other words: the distances between the different speech segments should be determined. The relations between speech segments and the way in which distances are found are studied in this chapter and also in Chapter 4. In this chapter we focus on discrete representations of segments. These representations can be used for the corpus frequency method (see Section 2.3.2), the frequency per word method (see Section 2.3.3), and the Levenshtein distance (see Section 5.1). In our research, language varieties are mainly compared with the Levenshtein distance. Using the Levenshtein distance also allows acoustic representations of the segments to be used. The way in which distances are obtained on the basis of acoustic representations is described in Chapter 4.

In Section 3.1 we describe the different ways in which sounds can be represented. We will look at both vowels and consonants. In Section 3.2 we discuss how diphthongs are represented in the different systems. Section 3.3 discusses the way in which affricates are processed. Section 3.4 will explain how suprasegmentals and diacritics are processed. In the feature system of Hoppenbrouwers & Hoppenbrouwers rules are applied to remove redundancy in feature specifications. This is explained in Section 3.5. In Section 3.6 it is described how distances between sounds are calculated on the basis of the different representations. For each feature representation, a vowel distance matrix and a consonant distance matrix can be calculated. Once segment distances are obtained, they can be used unchanged in the Levenshtein algorithm. Another possibility is to use the logarithms of the distances. This is discussed in Section 3.7. In Section 3.8 the different feature representations are compared to each other by correlating vowel and consonant distances based on the one system with vowel and consonant distances based on the other system. Finally we draw some conclusions in Section 3.9.

When discussing the representation of segments and the processing of suprasegmentals and diacritics, we consider two data sources. One consists of a number of transcriptions of Norwegian dialects, compiled by Jørn Almberg. Each text is a translation of the fable 'The North Wind and the Sun', in Norwegian: 'Nordavinden og sola' (NOS) (see Section 7.2). The other is the *Reeks Nederlandse Dialectatlassen* (RND), a series of Dutch atlasses edited by Blancquaert and Peé (1925–1982) (see Section 9.1). For application to other dialect comparison work based on the modern IPA system, the remarks for the NOS data source should be kept in mind where there are differences between the NOS and the RND.

## 3.1 Representation of segments

In transcriptions words are transcribed as a series of speech segments: phones. In the simplest case the phones are not further defined. Two phones are equal or different. This simple representation is described in Section 3.1.1.

Using the phone representation it is not possible to take into account the affinity between different, but kindred segments. Methods based on phones will not regard the pair [ɪ,e] as more kindred than [ɪ,ɒ]. This problem can be solved by replacing each phonetic symbol by a bundle of features. Each feature can be regarded as a phonetic property which can be used for classifying sounds. A feature bundle is a range of feature values. Each value indicates to what extent the corresponding property is instantiated.

We present the results of experiments on three feature systems. The first feature system is described by Hoppenbrouwers and Hoppenbrouwers (2001) (H & H). This is an articulation-based system, based on Chomsky and Halle (1968). The system is interesting because the developers themselves used this system for (Dutch) dialect comparison. In Section 3.1.2 we give a more detailed description. The second feature system is developed by Vieregge et al. (1984) and Cucchiarini (1993) (V & C). This system was developed for a comparison task similar to dialectological comparison, that of checking the quality of phonetic transcriptions. This involves comparison to consensus transcriptions. This system is interesting since it is partly perception-based. We describe the system in Section 3.1.3. The last feature system is developed by Almeida and Braun (1986) (A & B), intended for checking the quality of phonetic transcriptions as well. We also included this system in our research because the well-known IPA system is directly used for finding sound distances. The system is described more detailed in Section 3.1.4.

Although each of the three feature systems seems to be a good candidate for use in dialect comparison, none of three feature systems is originally developed for the approach in which we used them. The system of Hoppenbrouwers & Hoppenbrouwers was originally meant to be used in their feature frequency method (see Section 2.3.2). Instead of comparing sounds, histograms are compared. A histogram represents for each feature the number of sounds which are positively marked for that feature in a dialect transcription. Both the system of Vieregge & Cucchiarini and the system of Almeida & Braun were developed for checking phonetic transcriptions, not for dialect comparison. Because our goal is to find dialect distances that approach human perception, the perception-based system of Vieregge & Cucchiarini may give the most promising sound distances. In Section 7.4.3 dialect distances based on the different segment representations are validated.

Note that the sounds used in the RND form a subset of the sounds of the IPA system. The RND vowels are given in Appendix A Figure A.1 and the IPA vowels are given in Figure A.2. The RND consonants are given in Figure A.3 and the IPA consonants are given in Figure A.4.

### 3.1.1 Phones

In transcriptions, words consist of a sequence of phones. In the IPA system each phone is noted with a basic symbol (vowels, pulmonic and non-pulmonic consonants and other symbols) and optionally supplemented with one or more suprasegmentals and/or one or more diacritics.[1] The combination of a basic symbol supplemented with some additional marks is regarded as a phone.

### 3.1.2 Features Hoppenbrouwers and Hoppenbrouwers

For the feature system of H & H the *Sound Pattern of English* (Chomsky and Halle, 1968) was the starting point. According to H & H this articulation-based system is considered as a standard work, generally followed in studies and modern handbooks about modern phonology (Hoppenbrouwers and Hoppenbrouwers, 2001, p. 33 and 34). The SPE system was modified and extended so that with the use of this system the RND material is done as much justice as possible. This resulted in a system with 21 features. Since we also want to dispose of a system which is suitable for (partially) processing the complete IPA system in general and the NOS data in particular, we should be able to process some suprasegmentals and diacritics which do not appear in the RND data. Therefore, it was necessary to extend the H & H system with six extra features, obtaining a total of 27 features. Both the original and the added features are given in Table 3.1. All features are initially binary, where 0 means ‘absent’ and 1 ‘present’. In the

[1] Note that in the IPA system *half-long* (ˑ) and *long* (ː) are ordered under suprasegmentals. Although this may be debatable, we will use the same ordering as starting point throughout this thesis.

| Vowel features | | Consonant features |
|---|---|---|
| vowel | | consonant |
| front | | anterior |
| back | | coronal |
| round | | posterior |
| low | | laryngeal |
| polar | | sonorant |
| long | | voiced |
| peripheral | | high |
| diphthong | | continuant |
| nasal | | lateral |
| | | syllabic |
| *breathy* | | *apical* |
| *creaky* | | |
| *toneme 1* | | |
| *toneme 2* | | |
| *circumflex* | | |

Table 3.1: The features of the Hoppenbrouwers and Hoppenbrouwers system. The names of the original features were given in Dutch. In this table they are given in regular font style in English. These original features are used for the RND only. Features added for the NOS data are given in italics.

left column the vowel features are given and in the right column the consonant features can be found. The feature *nasal* is shared by both vowels and consonants.

In this section we describe the definitions of the vowels first and we give explanations about the definitions of the consonants next.

#### 3.1.2.1 Vowels

In this section we focus on the vowels and show the relation between the IPA vowel quadrilateral and the table of H & H.

Ladefoged (1975, p. 245) uses the feature *tense* to distinguish vowels which are on the periphery of the vowel area [+tense], and the corresponding lax vowels which are slightely lower and more central [−tense]. Following Ladefoged H & H use the feature *peripheral*, which distinguishes between centralized and non-centralized vowels.

Short vowels are always specified as [−peripheral]. However, a number of vowels get [+peripheral] when they are half-long or long. From the feature table given by Hoppenbrouwers and Hoppenbrouwers (2001, pp. 37–41) it can be concluded that the [e] and [ɪ], the [ø] and [ʏ], and the [o] and [ʊ] are equal to each

| | | front | | central | | back | |
|---|---|---|---|---|---|---|---|
| polar | high | i | y | | | ɯ | u |
| | high | e | ø | | | ɤ | o |
| | low | | | | | | |
| polar | low | æ | | a | ɶ | ɑ | ɒ |

Table 3.2: IPA sounds which are (or should be) defined as peripheral sounds in the feature system of H & H. Elements left in a cell are spread, and elements right are rounded.

other if they are short. However, when they are half-long or long, the [e], [ø] and [o] are specified as [−peripheral], and the [ɪ], [ʏ] and [ʊ] as [+peripheral].

Table 3.2 presents the vowels which are or should be defined as [+peripheral] according to the feature table of H & H when they get half-long or long. [ɯ], [ɤ] and [ɶ] are not given by H & H since they are not used in the RND. The [ɒ] is not given by H & H as well, although this vowel is used in part 1 of the RND. Table 3.3 shows the vowels which are always specified as [−peripheral], regardless their length. [ɨ], [ɘ], [ɜ] and [ɐ] are not given by H & H since they are not used in the RND. The [ʏ] is in the RND and in the feature table of H & H noted as [ʌ].[2] The [ɵ] and [ɜ] are noted respectively as [ᵿ] and [ə] in the feature table of H & H. Although the [ʉ], [ɵ] and [ɜ] are given by H & H, we did not find these vowels in the RND transcriptions we processed.

In the RND the schwa is noted as [ə], just as in the IPA system. The [ə] is defined in IPA as a half-round central vowel exactly between close-mid and open-mid. In the feature table of H & H the schwa is defined as a sound for which all features are absent, i.e., all features are set to zero. Only the features *vowel*, *sonorant*, *voiced*, *continuant* and *syllabic* are positively marked. The result is that the schwa is defined as a high central unrounded sound. So the schwa is defined as the IPA [ɘ]. Therefore, in the system of H & H the [ə] and the [ɘ] are not distinguished.

In the SPE system, intended for English segments, exactly three degrees of height can be defined using the features *high* and *low*. Defining Dutch vowels, it is necessary to be able to distinguish four degrees of height (Hoppenbrouwers and Hoppenbrouwers, 2001, p. 35). In the system of H & H this is realized with the features *low* and *polar*:

[2] In the RND the [ʌ] is introduced as a symbol representing the vowel in the Dutch word *bus* 'bus'. However, this vowel sounds approximately as the [ʏ] of the modern IPA system. For the English pronunciation of *bus* the use of the [ʌ] (as given in the modern IPA system) is correct.

| | | front | | central | | back | |
|---|---|---|---|---|---|---|---|
| polar | high | | | ɨ | ʉ | | |
| | high | ɪ | ʏ | ɘ | ɵ | | ʊ |
| | low | ɛ | œ | ɜ | ɞ | ʌ | ɔ |
| polar | low | | | ɐ | | | |

Table 3.3: IPA sounds which are (or should be) defined as non-peripheral sounds in the feature system of H & H. Elements left in a cell are spread, and elements right are rounded.

| | | |
|---|---|---|
| close | - | + |
| close-mid | - | - |
| open-mid | + | - |
| open | + | + |

The result is that the difference between close and open-mid is greater than the difference between close and open. Likewise the difference between open and close-mid is greater than between open and close. The smaller difference between the extremes (open and close) may be intended to reflect that a vowel shift is cyclic: e.g. an [ɔ] changes in a [o], an [o] changes in an [u], and next an [u] changes in an [ɒ]. We suspect that because of the 'polar' feature this feature system does not reflect the distance between segments with as much fidelity as some competitors.

Besides the features as given by Hoppenbrouwers and Hoppenbrouwers, some extra features are added. To be able to process some diacritics in the NOS data, we also needed to add the features *breathy*, *creaky*, *toneme 1*, *toneme 2* and *circumflex*.

For the use of the Levenshtein distance we will also need a definition of 'silence' (see Section 5.1). In the feature system of H & H all features can simply be defined as absent, i.e. set to 0. For vowels, this is equal to a [ə] or [ɘ] with [−vowel].

#### 3.1.2.2 Consonants

In this section we focus on the consonants and show the relation between the IPA consonant table and the table of H & H. We will consider all pulmonic IPA consonants. For consonants treated by H & H as well as consonants not given, we show the relation with the IPA consonant table. The relation with the manner of articulation (IPA columns) and the relevant H & H features is given in Table 3.4. The relation with the place of articulation (IPA rows) and the relevant H & H features is given in Table 3.5. In the H & H system all approximants (not lateral) are defined as [+high]. For the definition of the [w] the vowel feature *round* is

| IPA | ant | cor | post | lar | high |
|---|---|---|---|---|---|
| bilabial | + | - | - | - | - |
| labiodental | + | - | - | - | - |
| dental | + | + | - | - | - |
| alveolar | + | + | - | - | - |
| postalveolar | - | + | - | - | + |
| retroflex | - | + | - | - | + |
| palatal | - | + | - | - | + |
| velar | - | - | + | - | + |
| uvular | - | - | + | - | + |
| pharyngeal | - | - | - | - | - |
| glottal | - | - | - | + | - |

Table 3.4: Relation between IPA columns (manner of articulation) and the relevant H & H features.

specified as [+round].[3] The distinction between voiced and voiceless sounds is defined in the same way as in the IPA table of pulmonic consonants. H & H only defined the consonants which appear in the RND.

In the IPA system the [h] represents a voiceless glottal fricative, and the [ɦ] represents its voiced counterpart. It is striking that H & H specifies the Dutch /h/ as [+voiced], just as Booij (1995) does. However, in our opinion the Dutch /h/ is voiceless. Therefore, we specify this segment as [−voiced]. This agrees with Rietveld and Van Heuven (1997, p. 395) who transcribe the /h/ in *hond* 'dog' as [h], and not as [ɦ].

Using 21 features, all RND sounds get a unique definition. However, not all sounds of the complete set of pulmonic IPA sounds are uniquely defined. So H & H write that with close to thirty features all sounds which appear in natural languages can be defined (Hoppenbrouwers and Hoppenbrouwers, 2001, p. 9).

Besides the basic features given by Hoppenbrouwers and Hoppenbrouwers, some features were added to process some diacritics. For both the RND data and the NOS data we added the feature *syllabic*. For the NOS data only we added the feature *apical*.

For the use of the Levenshtein distance we will also need a definition of 'silence' (see Section 5.1). As mentioned in Section 3.1.2.1 for 'silence' all features can simply be defined as absent, i.e. set to 0. For consonants, this is equal to the [ʔ] with [−consonant −laryngeal].

[3] In our research the [w] is regarded as a voiced bilabial approximant which can be located in the IPA table of pulmonic consonants. However, in the IPA system (revised to 1993, updated in 1996) the [w] is ordered under 'Other Symbols' and mentioned as a voiced labial-velar approximant.

| IPA | nas | cons | son | cont | lat |
|---|---|---|---|---|---|
| plosive | - | + | - | - | - |
| nasal | + | + | + | - | - |
| trill | - | + | + | + | - |
| tap or flap | - | + | + | + | - |
| fricative | - | + | - | + | - |
| lat. fric. | - | + | - | + | + |
| approximant | - | - | + | + | - |
| lat. appr. | - | + | + | + | + |

Table 3.5: Relation between IPA rows (place of articulation) and the relevant H & H features.

### 3.1.3 Features Vieregge and Cucchiarini

The reliability of transcriptions can be measured by determining the degree of similarity between transcriptions carried out either by the same transcriber at different times, which corresponds to the use of “reliability” in its strict sense (Bürkle, 1986), or by different transcribers, an option also left open by Vieregge et al. (1984). The validity of transcriptions is measured by comparing individual transcriptions with expert transcriptions (Vieregge et al., 1984).

In 1984 Vieregge et al. 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. Tables for vowels and consonants are given by Vieregge (1987). An advantage of Vieregge’s system is that it is partly based on real measurements. The vowel system is based on experimental data which was found in the literature. The consonant system is based on a perception experiment, in which subjects were asked to give the distance between two consonants on a scale from 1 (minimal dissimilarity) to 10 (maximal dissimilarity). Next a feature system was developed, such that sound distances on the basis of features approach the perceptual distances maximally. The complete system was originally developed for Dutch.

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 had not been included earlier, as well as other sounds that probably could crop up in the transcriptions that 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 filled differently. 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". The use of a Dutch Vieregge system which is extended and applied to e.g. Czech will probably reflect the perception of Dutch people listening to Czech, rather than the perception of the Czech speakers themselves.

Distance measures which are developed to assess transcriptions can also be 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.

For our purpose of quantifying dialect distances we use a feature system which is a combination of the vowel feature system of Vieregge et al. and the consonant feature system of Cucchiarini. We first describe the vowel system and then we give a description of the consonant system.

#### 3.1.3.1 Vowels

For the construction of the vowel feature system Vieregge et al. consulted various data in the literature. The existing literature also provided enough experimental data on the basis of which Dutch vowel distances could be found. As examples the authors refer to Nooteboom (1971), Nooteboom (1972) and Rietveld (1979). The vowel feature system of Vieregge et al. consists of four features: *advancement*, *high*, *long* and *rounded*. The possible values for the features are listed in Table 3.6. The features *advancement*, *high* and *rounded* correspond with the dimensions of the IPA vowel quadrilateral as can be seen in Table 3.7. From the values of the feature *advancement* it appears that this feature has extra weight. The authors refer to Rietveld (1979) who show that 'the proprioceptive articulatory dissimilarities can be predicted quite satisfactorily by using a traditional vowel scheme and giving extra weight to differences on the front/back dimension'. Although this statement refers to a subset of Dutch vowels, namely [i, e, ɛ, y, ø, u, o, ɔ, ɑ], Vieregge et al. assume that this finding can be applied to all Dutch vowels.

The tables of Vieregge (1987) show that Dutch [ɪ], [ʏ], [ʊ], [ə], [ɛ] and [ɑ] can only be short. The [i], [y], [u], [ɛ], [œ] and [ɔ] can only be short or long. The [e], [ø], [o] and [a] can only be half-long or long. The [ɪ] is defined as a short [e], and the [ʊ] is defined as a short [o].

For our research we extended the vowel system so that it contains all vowels of the IPA vowel quadrilateral. The result can be seen in Table 3.8. All IPA vowels are defined by analogy with the IPA vowel quadrilateral. The result is that the [ɪ] and the [e], and the [ʊ] and the [o] no longer have the same values for the features *advancement* and *high*. Now the [ʏ] is defined as a rounded [ɪ] and the [œ] as the rounded [ɛ]. The [a] is defined as a front vowel instead of a central

| Feature | Value | Meaning |
|---|---|---|
| *vowel* | 0 | no |
| | 1 | yes |
| advancement | 2 | front |
| | 4 | central |
| | 6 | back |
| high | 1.0 | open |
| | 1.5 | near-open |
| | 2.0 | open-mid |
| | 2.5 | central |
| | 3.0 | close-mid |
| | 3.5 | near-close |
| | 4.0 | close |
| long | 1 | short |
| | 2 | half-long |
| | 3 | long |
| rounded | 0 | no |
| | 1 | yes |
| *nasal* | 0.0 | not nasal |
| | 0.5 | half-nasal |
| | 1.0 | nasal |
| *breathy* | 0 | no |
| | 1 | yes |
| *creaky* | 0 | no |
| | 1 | yes |
| *toneme 1* | 0 | no |
| | 1 | yes |
| *toneme 2* | 0 | no |
| | 1 | yes |
| *circumflex* | 0 | no |
| | 1 | yes |

Table 3.6: The vowel features of Vieregge et al. and their possible values. We extended the system with some extra features, in this table given in italics. Only the first seven features in this table are used for the RND data, the last five features are added for the NOS data.

| | front | | central | | back | |
|---|---|---|---|---|---|---|
| close | i | y | | ʏ | | u |
| close-mid | e/ɪ | ø | ə | | | ʊ/o |
| open-mid | ɛ | | | œ | | ɔ |
| open | | | a | | ɑ | |

Table 3.7: The Dutch vowels as defined by the features of Vieregge et al.. Elements left in a cell are spread, and elements right are rounded.

vowel, and the [ə] is defined as half-rounded instead of not-rounded. In the IPA vowel quadrilateral we interpreted the [æ] and [ɐ] as not rounded, the [ə] as half rounded and the [ʊ] as rounded.

In the original system, all the possible lengths were not available for all the vowels. In our adapted system for vowels, all lengths are allowed. The correct use of length marks is assumed to be the responsibility of the transcriber. In the original system, for short sounds the feature *long* is set to 3, for half-long sounds to 2 and for long sounds to 1. We have reversed this order: for short sounds the feature *long* gets the value 1, for half-long sounds the value 2 and for long sounds the value 3.

Besides the features given by Vieregge et al., some extra features are added. To be able to process some diacritics, for both the RND data and the NOS data we added the feature *nasal*. For the NOS data we also needed to add the features *breathy*, *creaky*, *toneme 1*, *toneme 2* and *circumflex*.

A feature *vowel* was also added. Usually for vowels this feature is set to 1 and for consonants this feature is set to 0. However, for the [j] and [w] the feature is set to 1 as well. In our system the [i], [j], [u] and [w] are defined as both vowels and consonants. The [j] and the [w] share all the vowel features of respectively the [i] and the [u], and the [i] and the [u] share all the consonant features of the [j] and the [w] (see Section 3.1.3.2). When counting frequencies, both the vowel features and the consonant features are counted for these sounds, however, they are weighted by half. When finding the distance between two segments which are defined as both vowel and consonant, first the distance on the basis of the vowel features is calculated, and next the distance on the basis of the consonant features is found. The final distance between the two segments is equal to the mean of vowel distance and the consonant distance.

The feature *vowel* also plays a role in the definition of silence; a definition of silence in terms of vowel features will be used in the Levenshtein algorithm (see Section 5.1). We defined it to be equal to the schwa, except that the feature *vowel* is set to 0.

| | front | | central | | back | |
|---|---|---|---|---|---|---|
| close | i | y | ɨ | ʉ | ɯ | u |
| near-close | ɪ | ʏ | | | | ʊ |
| close-mid | e | ø | ɘ | ɵ | ɤ | o |
| central | | | ə | | | |
| open-mid | ɛ | œ | ɜ | ɞ | ʌ | ɔ |
| near-open | æ | | ɐ | | | |
| open | a | ɶ | | | ɑ | ɒ |

Table 3.8: The IPA vowels defined using the features of Vieregge et al. by analogy with the IPA vowel quadrilateral. Elements left in a cell are spread, and elements right are rounded.

#### 3.1.3.2 Consonants

For getting perceptual distances between 18 Dutch consonants Vieregge et al. performed a perception experiment in which 25 first year speech therapy students were presented with pairs of consonants in medial word position. They were asked to rate each pair on articulatory dissimilarity on a scale from 1 (minimal dissimilarity) to 10 (maximal dissimilarity). The stimulus material consisted of $(18^2 - 18)/2 = 153$ word pairs which differed as little as possible. The stimuli were offered in random order on paper.

Next, a feature system was developed to model perceptual distances as accurately as possible. Features for both place and manner of articulation can be found, comparable to the IPA system, as well as a feature for distinguishing between voiced and voiceless consonants.

The system was originally developed for Dutch, where only for a subset of the Dutch consonants the perceptual distances were measured, viz., the [p], [b], [t], [d], [k], [f], [v], [s], [z], [x], [m], [n], [ŋ], [l], [ʀ], [w], [j] and [h]. Along the lines of Vieregge et al., Cucchiarini (1993) extended the system so that it can be used for Dutch, Limburg and Czech. She replaced the feature *flap* by the feature *trill.* Along the lines of Vieregge et al. she added a number of consonants. For consonants we used the system of Cucchiarini as a basis. Along the lines of Cucchiarini's system we introduced extensions so that it contains all pulmonic consonants of the IPA system. The possible values for the features are listed in Table 3.9. The relation between the manner of articulation (IPA columns) and the relevant features of Cucchiarini is given in Table 3.10. However, although palatal consonants are defined as distributed, the [j] is defined as non-distributed. The relation between the place of articulation (IPA rows) and the relevant Cucchiarini features is given in Table 3.11. The feature *voice* is defined exactly as in the IPA consonant table.

| Feature | Value | Meaning |
|---|---|---|
| *consonant* | 0 | no |
| | 1 | yes |
| place | 1.0 | bilabial/labiodental |
| | 1.5 | dental |
| | 2.0 | alveolar/postalveolar |
| | 2.5 | retroflex |
| | 3.0 | palatal |
| | 4.0 | velar/uvular |
| | 4.5 | pharyngeal |
| | 5.0 | glottal |
| voice | 0 | voiceless |
| | 1 | voiced |
| nasal | 0.0 | not nasal |
| | 0.5 | half-nasal |
| | 1.0 | nasal |
| stop | 0 | no |
| | 1 | yes |
| glide | 0 | no |
| | 1 | yes |
| lateral | 0 | no |
| | 1 | yes |
| fricative | 0 | no |
| | 1 | yes |
| trill | 0 | no |
| | 1 | yes |
| high | 0 | no |
| | 1 | yes |
| distributed | 0 | no |
| | 1 | yes |
| *syllabic* | 0 | no |
| | 1 | yes |
| *apical* | 0 | no |
| | 1 | yes |

Table 3.9: The consonant features of Cucchiarini and their possible values. We extended the system with some extra features, in this table given in italics. Only the first twelve features in this table are used for the RND data, the last feature is added for the NOS data.

| IPA | place | high | distributed |
|---|---|---|---|
| bilabial | 1.0 | 0 | 1 |
| labiodental | 1.0 | 0 | 0 |
| dental | 1.5 | 0 | 0 |
| alveolar | 2.0 | 0 | 0 |
| postalveolar | 2.0 | 1 | 1 |
| retroflex | 2.5 | 0 | 0 |
| palatal | 3.0 | 1 | 1 |
| velar | 4.0 | 1 | 0 |
| uvular | 4.0 | 0 | 0 |
| pharyngeal | 4.5 | 0 | 0 |
| glottal | 5.0 | 0 | 0 |

Table 3.10: Relation between IPA columns (manner of articulation) and the relevant features of Cucchiarini.

Besides the basic features given by Cucchiarini, some features are added to process some diacritics. For both the RND data and the NOS data we added the feature *syllabic*. For the NOS data only we added the feature *apical*.

A feature *consonant* is added. Usually for the vowels this feature is set to 0, and for the consonants this feature is set to 1. However, for the [i] and the [u] the feature is set to 1 as well. In our system the [i], [j], [u] and [w] are defined as both vowels and consonants. The [i] and the [u] share all the consonant features of respectively the [j] and the [w], and the [j] and the [w] share all the vowel features of [i] and [u] (see Section 3.1.3.1 for more details).

The feature *consonant* also plays a role in the definition of silence; a definition of silence in terms of consonant features will be used in the Levenshtein algorithm (see Section 5.1). We defined it to be equal to the glottal stop, except that the feature *consonant* is set to 0.

### 3.1.4 Features Almeida and Braun

At the same time as the Vieregge system was developed an alternative system with the same goal was developed which was based on the IPA tables. The system was first developed in the phonetics department of the research institute for German language “Deutscher Sprachatlas” (Marburg, Germany) in 1980 and was further developed and formalized later (Almeida and Braun, 1986). 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, followed by an inference on the part of the transcriber of the articulatory gestures of the speaker, and finally in a phonetic

| IPA | nasal | stop | glide | lateral | fricative | trill |
|---|---|---|---|---|---|---|
| plosive | 0 | 1 | 0 | 0 | 0 | 0 |
| nasal | 1 | 0 | 0 | 0 | 0 | 0 |
| trill | 0 | 0 | 0 | 0 | 0 | 1 |
| tap or flap | 0 | 0 | 0 | 0 | 0 | 0 |
| fricative | 0 | 0 | 0 | 0 | 1 | 0 |
| lat. fric. | 0 | 0 | 0 | 1 | 1 | 0 |
| approximant | 0 | 0 | 1 | 0 | 0 | 0 |
| lat. appr. | 0 | 0 | 0 | 1 | 0 | 0 |

Table 3.11: Relation between IPA rows (place of articulation) and the relevant features of Cucchiarini.

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) which essentially consists in an abbreviation for a combination of articulatory features.[4] The Almeida & Braun system is an articulatory system in which sound distances are derived from the IPA vowel quadrilateral and the IPA consonant table. 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 our research we introduced adjustments to the Almeida & Braun system. The description given in this section is based on Heeringa and Braun (2003). We describe first the definitions of the vowels and next we explain how the definitions for the consonants were determined.

#### 3.1.4.1 Vowels

The basis for finding vowel distances is the IPA vowel quadrilateral as given in Appendix A Figure A.2. The quadrilateral reflects three features: *advancement*, *height* and *roundedness*. The possible values for the features are listed in Table 3.12. 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 (rounded: no vs. yes) as one step. So when simply subtracting the corresponding feature values from each other and taking the absolute value, we get a distance of one for each of these three pairs.

In the IPA vowel quadrilateral we interpreted the [æ] and [ɐ] as not rounded, the [ə] as half rounded and the [ʊ] as rounded.

[4] The system can be found in the *Handbook of the International Phonetic Association* (IPA, 1999) as well as via: `http://www2.arts.gla.ac.uk/IPA/ipachart.html`.

| Feature | Value | Meaning |
|---|---|---|
| *vowel* | 0 | no |
| | 1 | yes |
| advancement | 1 | front |
| | 2 | central |
| | 3 | back |
| height | 1 | close |
| | 2 | near-close |
| | 3 | close-mid |
| | 4 | central |
| | 5 | open-mid |
| | 6 | near-open |
| | 7 | open |
| roundedness | 0 | no |
| | 1 | yes |
| *long* | 0.0 | short |
| | 0.5 | half-long |
| | 1.0 | long |
| *nasal* | 0.0 | not nasal |
| | 0.5 | half-nasal |
| | 1.0 | nasal |
| *breathy* | 0 | no |
| | 1 | yes |
| *creaky* | 0 | no |
| | 1 | yes |
| *toneme 1* | 0 | no |
| | 1 | yes |
| *toneme 2* | 0 | no |
| | 1 | yes |
| *circumflex* | 0 | no |
| | 1 | yes |

Table 3.12: The vowel features of Almeida and Braun and their possible values. We extended the system with some extra features, in this table given in italics. Only the first seven features in this table are used for the RND data, the last five features are added for the NOS data.

Besides the basic features derived from the vowel quadrilateral, some extra features are added. To be able to process some suprasegmentals and diacritics, we added the features *long* and *nasal* for both the RND data and the NOS data. Only for the NOS data we added the features *breathy*, *creaky*, *toneme 1*, *toneme 2* and *circumflex*.

A feature *vowel* was also added. Usually for the vowels this feature is set to 1 and for the consonants this feature is set to 0. However, for the [j] and [w] the feature is set to 1 as well. In our system the [i], [j], [u] and [w] are defined as both vowels and consonants. The way in which these sounds are defined and processed is similar as in the V & C system (see Section 3.1.3.1 for more details).

The feature *vowel* also plays a role in the definition of silence. A definition of silence in terms of vowel features will be used in the Levenshtein algorithm (see Section 5.1). We defined it to be equal to the schwa, except that the feature *vowel* is set to 0.

#### 3.1.4.2 Consonants

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 Appendix A Figure A.4. In this table it can be seen that in our system the voiced labial-velar approximant [w] is regarded as, and will be treated as, 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 3.13. 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 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 diacritics. For both the RND data and the NOS data we added the feature *syllabic*. For the NOS data only we added the feature *apical*.

A feature *consonant* is added. Usually for the vowels this feature is set to 0, and for the consonants this feature is set to 1. However, for the [i] and the [u] the feature is set to 1 as well. In our system the [i], [j], [u] and [w] are defined

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

Table 3.13: The consonant features of Almeida and Braun and their possible values. We extended the system with some extra features, in this table given in italics. Only the first five features in this table are used for the RND data, the last feature is added for the NOS data.

as both vowels and consonants. The way in which these sounds are defined and processed is similar as in the V & C system (see Section 3.1.3.2).

A definition of silence in terms of consonant features will be used in the Levenshtein algorithm (see Section 5.1). We defined it to be equal to the glottal stop, except that the feature *consonant* is set to 0.

## 3.2 Diphthongs

A diphthong is a vowel with a changing color. In the feature table of H & H diphthongs are combinations of two vowels, where the first segment is short or half-long, and the second short. When the first element is long, two succeeding vowels are not regarded as a diphthong.

When processing diphthongs, we need an accurate representation of them. Vieregge et al. write that they follow Moulton (1962) and consider a diphthong as vowel+vowel sequence, the second vowel being non-syllabic allophonically. On the other hand, H & H regard regular diphthongs as segmental units (Hoppenbrouwers and Hoppenbrouwers, 2001, p. 35). We make no a priori decision here but experimented with both two-phone and one-phone representations. Validation work will make clear whether the different representations result in different results, and which representation gives the better results (see Chapter 7).

In this section we explain how diphthongs can be defined as segmental units using the different segment representations. An accurate representation should reflect the color at the onset, the transition, and the color of the offset. The second element as noted in a transcription may not always be the real offset. It may also be a target position which is not really pronounced (Rietveld and Van Heuven, 1997, p. 74). However, with the discrete representations we used it is not possible to represent diphthongs in such a refined way. E.g. the way in which the transition takes place cannot be represented. In our research we used a simplied approach in which the definition of a diphthong is based only on the color of the onset and the color of the target position.

When using the phone representation, all sounds are equally different. Therefore, when a diphthong is regarded as one segmental unit, e.g. the [a] and [au] will be regarded as equally different as the [a] and the [i], except that the [au] (like all diphthongs) is treated as a long sound. So no special specifications for diphthongs need to be made. This is a very rough approach where the color of the onset and the target position plays no role. However, when using feature representations, the definition of diphthongs is based on the definitions of the onset color and the target color. Below we explain how diphthongs are specified in the feature system of H & H and how we defined them for the systems of V & C and A & B.

H & H make a distinction between *closing diphthongs* and *centering diphthongs*. A closing diphthong is a long vowel with a movement toward a non-central position in the vowel space. On the contrary a centering diphthong is a vowel with a movement toward a central position in the vowel space, the schwa. In Section 3.2.1 we describe the specification of closing diphthongs, and in Section 3.2.2 we explain the specification of centering diphthongs.

### 3.2.1 Closing diphthongs

As mentioned above a closing diphthong is a long vowel with a movement toward a non-central position in the vowel space. The term *closing* indicates that there is a movement in the direction of a closer vowel. This movement can be vertical (e.g. [ɔu]) or diagonal (e.g. [ɔi]). In a diphthong, the color of a sound changes from a start position to an end position. Therefore, when specifying a diphthong the feature values are derived from the feature values of the vowel corresponding with the start position and the vowel corresponding with the end position. If the second element of a diphthong is noted as a [j] or a [w], we used the feature values of the [i] and [u] respectively. In the system of H & H the feature values of closing diphthongs are defined as follows:

| | | |
|---|---|---|
| front | : | mean of both segments |
| back | : | mean of both segments |
| round | : | mean of both segments |
| low | : | value of first element |
| polar | : | value of first element |
| long | : | always [+long] |
| peripheral | : | always [+peripheral] |
| diphthong | : | always [+diphthong] |

A movement from front to back (or vice versa) is specified by using the mean of the start position and the end position. H & H symbolize the mean value using '$*$'. It is striking that a similar procedure is not followed with respect to the height. The features *low* and *polar* simply get the value of the first segment. Does this mean that the first segment is most dominant? H & H write that this vertical closing movement is specified by specifying the feature *diphthong* as positive (Hoppenbrouwers and Hoppenbrouwers, 2001, p. 46). Because there is always a movement to either the [i] or [u] in closing diphthongs, they are specified as [+peripheral] which makes closing diphthongs a bit more related to the (half-)long [i] and [u], which are also specified as [+peripheral] (see Table 3.2). Closing diphthongs are specified as peripheral sounds regardless whether the (half-)long version of the first element is specified as peripheral or not.

For the system of V & C we define the feature values of closing diphthongs as follows:

| | |
|---|---|
| advancement | mean of both segments |
| high | mean of both segments |
| long | always *long*=3 |
| rounded | mean of both segments |

In contrast to H & H the value of *high* is determined in the same way as the value for *advancement.* In our opinion the value for *high* should also be based on both segments. A disadvantage of this approach compared to that of H & H is that the order of segments is not represented. For example the [ɑu] and the [uɑ] are defined in exactly the same way. For both the RND and the NOS data this is no problem since in the selection of diphthongs in both data sources none of the diphthongs has a symmetric counterpart. However, when a selection of diphthongs contains symmetric cases, a solution may be to weight the advancement, height and rounding of the first element 75%, and the same features of the second element 25%.

For the system of A & B we define the feature values of closing diphthongs as follows:

| | |
|---|---|
| advancement | mean of both segments |
| height | mean of both segments |
| roundedness | mean of both segments |
| long | always *long*=1 |

Just as in the V & C system for the feature *high* in the A & B system the value for the feature *height* is based on both segments, which is different from the approach of H & H. Again symmetric diphthongs cannot be distinguished from each other when using our definition. As noted above, however, for both the RND and the NOS data no symmetric cases were selected.

Using the RND data we adopted the selection of closing diphthongs as made by H & H. In Table 3.14 the closing diphthongs which were included in the feature table of H & H are listed, extended with six diphthongs which are lacking in the H & H table. The fact that these six diphthongs were missing has to do with the fact that the monophthong [ɒ] is also missing from the table.

When a closing diphthong is noted as a combination of vowel+vowel, for some diphthongs the first element is short, for others the first element is half long. The first element is never long. If a closing diphthong is noted as a combination of vowel+consonant, then the first element is always short. The first element is never half-long or long. According to this H & H write that only sequential diphthongs of the type [aːj], consisting of a long vowel followed by a [i] or [j] as in Dutch *fraai* and *mooi* are biphonemically interpreted (Hoppenbrouwers and Hoppenbrouwers, 2001, p. 35). For all closing diphthongs the second element should be short.

| Example | | Diphthong | Defined as |
|---|---|---|---|
| | | [ʏy] | [œy] |
| | | [ʊi] | |
| Dutch: | hooi | [ʊˑi] | |
| English: | bay | [ei] | |
| | | [eˑi] | |
| | | [øi] | [øy] |
| | | [øy] | |
| | | [oj] | [ʊj] |
| | | [ou] | |
| | | [oˑu] | |
| | | [ɛi] | |
| Dutch: | wijn | [ɛˑi] | |
| | | [ɛj] | |
| | | [œi] | [œy] |
| | | [œy] | |
| Dutch: | huis | [œˑy] | |
| English: | boy | [ɔi] | |
| | | [ɔˑi] | |
| | | [ɔj] | |
| | | [ɔu] | |
| Dutch: | koud | [ɔˑu] | |
| | | [ɔw] | |
| | | [æi] | |
| | | [æˑi] | |
| | | [æj] | |
| English: | line | [ai] | [æi] |
| | | [ay] | [ɶy] |
| | | [aˑy] | [ɶy] |
| | | [ɑi] | |
| German: | drei | [ɑˑi] | |
| | | [ɑj] | |
| | | [ɑu] | |
| Dutch: | saus | [ɑˑu] | |
| | | [ɑw] | |
| | | [ɒi] | |
| | | [ɒˑi] | |
| | | [ɒj] | |
| | | [ɒu] | |
| | | [ɒˑu] | |
| | | [ɒw] | |

Table 3.14: Dutch closing diphthongs as found in the feature table of H & H. The last six diphthongs were not originally included in the table.

The selection of closing diphthongs for the NOS data is much smaller. In the Norwegian text ‘Nordavinden og sola’ the only probable diphthong is the *ei* as in *dei* ‘them’, *ein* ‘a’, *seg* ‘him, her’, *dei* ‘they’, *einige* ‘agreed’, *skein* ‘shone’ and *meir* ’more’. In the cases where the ‘potential diphthongs’ in these words were perceived, one of the following six transcriptions could be used: [ei], [eɪ], [ɛi], [ɛɪ], [æi] and [æɪ]. Both the first and the second element are short. All suprasegmentals and diacritics which may be applied to monophthongs may also be applied to these diphthongs.

In the H & H table no closing diphthongs are defined with suprasegmentals and diacritics (except for length as just explained). In our research all suprasegmentals and diacritics which we allow to be applied to monophthongs (see Section 3.4) may also be applied to closing diphthongs. When a suprasegmental or diacritic is noted after the first or second segment of a diphthong, it is applied to the diphthong as a whole.

In the RND the second element of a closing diphthong is often noted as extra-short (i.e. in superscript or with a smaller character), probably with the goal to express that the first element is more significant than the second element. In the table of H & H the second element of the diphthong is always noted in normal script. However, when it was noted as extra-short, we treated it as a short sound. In the NOS data a second element of a closing diphthong noted as extra-short was never observed. However, if this should occur, it will be treated as short as well, just as for the RND data.

### 3.2.2 Centering diphthongs

A centering diphthong is a vowel with a movement toward a central position in the vowel space, the schwa. So when specifying a centering diphthong, the feature values should be derived from the the feature values of the first vowel and the ending schwa. In the feature system of H & H the feature values of centering diphthongs are defined as follows:

| | | |
|---|---|---|
| front | : | mean of both segments |
| back | : | mean of both segments |
| round | : | mean of both segments |
| low | : | mean of both segments |
| polar | : | mean of both segments |
| long | : | always [+long] |
| peripheral | : | mean of both segments |
| diphthong | : | always [−diphthong] |

A movement from front to back (or vice versa) is defined by taking the mean of the start position and the end position. Different from the definition for closing

| Example | | Diphthong |
|---|---|---|
| Afrikaans: | tee | [iɘ] |
| | | [yɘ] |
| Afrikaans: | voet | [uɘ] |
| Frisian: | each | [ɪɘ] |
| | | [ʏɘ] |
| | | [ʊɘ] |
| | | [eɘ] |
| Afrikaans: | deur | [øɘ] |
| Frisian: | roas | [oɘ] |
| | | [ɛɘ] |
| | | [œɘ] |
| | | [ɔɘ] |
| | | [æɘ] |
| | | [aɘ] |
| | | [ɑɘ] |
| | | [ɒɘ] |

Table 3.15: Dutch centering diphthongs as found in the feature table of H & H. The last diphthong was not originally included in the table.

diphthongs, a movement from low to high (or vice versa) is also defined by taking the mean of the start position and the end position. Centering diphthongs also differ from closing diphthongs in the way the feature *peripheral* is defined. For centering diphthongs the mean of both elements is used. For finding this value as first element the (half-)long sound is used, which may be specified as either [−peripheral] or [+peripheral] (see Section 3.1.2.1).

The feature values for the centering diphthongs in the systems of V & C and A & B are found in the same way as for the closing diphthongs (see Section 3.2.1) which is analogous to the way in which H & H find the feature values for centering diphthongs.

Using the RND data we adopted the selection of centering diphthongs as made by H & H, just as we did for the closing diphthongs. In Table 3.15 the centering diphthongs which were included in the feature table of H & H are listed, extended with one diphthong which is missing in the H & H table, due to the monophthong [ɒ]'s being missing.

In the combination vowel+vowel the first element is always short. The first element is never half-long or long. The second element is always short as well.

For the NOS data no centering diphthongs are selected. In the Norwegian text 'Nordavinden og sola' the only probable diphthong is the [iə] as in [$^2$eːniə] *enige* 'agreed'. This pronunciation (or something similar) was found in the dialects of Bø, Borre, Larvik, Stavanger and Trysil. Here the [iə] may be an shortened form of [igə] since we found in the dialect of Fyresdal the pronunciation [²eːnigə]. Therefore, we decided not to include the sequence [iə] as centering diphthong.

In the H & H table for all centering diphthongs a nasalized version is also defined. In our research all suprasegmentals and diacritics which we allow to be applied to monophthongs (see Section 3.4) may also be applied to centering diphthongs, just as for closing diphthongs. When a suprasegmental or diacritic is noted after the first or second segment of a diphthong, it is applied to the diphthong as a whole.

In the RND the second element of a centering diphthong is often noted as extra-short, just as for closing diphthongs (see Section 3.2.1). However, when it was noted as extra-short, we treated it as a short sound. For the NOS no centering diphthongs were selected. However, if centering diphthongs would be processed and the second element is noted as extra-short, the second element will be treated as short as well, just as for the RND data.

## 3.3 Affricates

In the RND data no affricates are used. However, in the NOS data they do appear. When processing them, both elements are processed as extra-short independent elements in sequence. E.g. [t͡s] is treated as [t̆s̆]. The idea behind this is that two extra-short elements are one segment of normal length together. This can be illustrated with an example. The Standard German word for ‘plough’ is *pflügen*. In Low-German dialects the word *plögen* (or something similar) is often used. However, in a small southerly part of this northern area the words *flichen* and *flien* are used, while in the Prussian area in the northwest the word *flieje* (besides *pleje*) is used (König and Paul, 1991, p. 198). Here we see that the [p], [pf] and [f] correspond to each other. Regarding a [pf] as a sequence of an extra-short [p] and an extra-short [f] allows us to match both elements with one segment of normal length, either a [p] or [f].

In some cases the first element of an affricate is stressed more than the second, or the second element is stressed more than the first. E.g. in the Sardinian dialect of Atzara the equivalent for ‘daughters’ is pronounced as [ˈfid͡ʒaza]. In the affricate [d͡ʒ] the first element should be processed as a short sound, and the second as an extra-short sound. In the Sardinian dialect of Abbasanta the equivalent for ‘policeman’ is pronounced as [polit͡sɔtˑɔ]. In the affricate [t͡s] the first element should be processed as an extra-short sound, and the second as a short sound. In our research these cases should not be noted as affricates, but as sequences of a short sound followed by an extra-short sound and an extra-short sound followed by a short sound respectively.

## 3.4 Suprasegmentals and diacritics

Using different sound representations it is possible to process suprasegmentals and diacritics. We did not implement the processing of all suprasegmentals and diacritics, so in this section only a subset is examined. The selection of suprasegmentals and diacritics was determined by the fact that they appear in the transcriptions we used on the one hand, and by the possibilities of the sound representations on the other hand.

When processing suprasegmentals and diacritics it is important to find and use the right weights which represent as precisely as possible the effect as perceived by listeners. In the next sections, all weights proposed are not based on real measurements. They are intuitively assigned. Besides, for different language groups different weights should be used. Our starting point is mainly the Dutch language area.

### 3.4.1 Stress and tonemes

In the RND, one symbol is available for marking stress [ˈ] which we interpret as corresponding with primary stress. The RND does not consistently mark which syllable is stressed for every word. It may be that stress is only noted when a syllable is stressed that differs from the one which the transcriber expected to be stressed.

In most Dutch dialects tonemes play no role. Only in the Limburg dialects can tonemes be found. In the RND, the dialects of the Limburg area were recorded by four transcribers. The transcribers did not note tonemes in equal detail (see part 8 of the RND). The fact that stress and tonemes are not consequently noted in the RND material may be the reason why these are not processed in the H & H system. For the RND, we also did not process these suprasegmentals.

In the IPA system we find symbols for primary stress [ˈ] and secondary stress [ˌ]. However, most Norwegian dialects are pitch accent varieties. All syllables with primary stress generate tone, or ‘accent’ (or ‘tonal accent’). Also: tonal accent can only be generated from primary stressed syllables. Only a few Norwegian dialects lack the tonal/accentual opposition. They generate the same tone, or accent, for all words, which is the same as primary stress in the IPA system. These dialects are found in an area around Bergen, in the Brønnøy area north of Trondheim and in many dialects of the two northernmost counties, Troms and Finnmark. In the varieties with tonal/accentual opposition three tonemes may occur: toneme 1 and toneme 2 (Kristoffersen, 2000) and circumflex (Almberg, 2001). Since no symbols are available in the IPA system for these tonemes, extra symbols are introduced and used in the NOS transcriptions. Syllables with toneme 1 are preceeded by a [ˈ], syllables with toneme 2 by a [$^2$], and syllables with circumflex by a [˜].[5] All transcriptions of the NOS data were made by one transcriber, who noted stress and tonemes consistently. Therefore, we process stress and tonemes for the NOS data when using feature representations. For the phone representation we found no way to deal with them.

In the transcriptions stress and toneme marks are noted before a syllable. To be able to process these marks, we shift them to the first vowel in the syllable.[6] So stress and tonemes are processed like properties (features) of a vowel. We suppose that stress and tonemes are mainly realized by the way in which vowels in syllables are pronounced. When diphthongs are processed as one sound and a stress or toneme mark is noted before the second element of a diphthong, it is shifted before and applied to the first vowel to the right. This may happen when the first element is the last segment of the one syllable, and the second element is the first segment of the next syllable.

To be able to process stress and tonemes as properties of vowels, we extended the feature systems of H & H, V & C and A & B with three features: *toneme 1*, *toneme 2* and *circumflex*. With these features the different stresses and tonemes are represented as follows:

| | toneme 1 | toneme 2 | circumflex |
|---|---|---|---|
| primary stress | 0.250 | 0.250 | 0.250 |
| secondary stress | 0.125 | 0.125 | 0.125 |
| stress and toneme 1 | 0.500 | 0.250 | 0.250 |
| stress and toneme 2 | 0.250 | 0.500 | 0.250 |
| circumflex | 0.250 | 0.250 | 0.500 |

On the basis of these representations, the distances between the stresses and tonemes can be calculated as the sum of the differences per feature (see Section 3.6.2 for more explanation and other alternatives). This results in the following distances:

| | nothing | secondary | primary | toneme 1 | toneme 2 | circumflex |
|---|---|---|---|---|---|---|
| nothing | | 0.375 | 0.750 | 1.000 | 1.000 | 1.000 |
| secondary | | | 0.375 | 0.625 | 0.625 | 0.625 |
| primary | | | | 0.250 | 0.250 | 0.250 |
| toneme 1 | | | | | 0.500 | 0.500 |
| toneme 2 | | | | | | 0.500 |
| circumflex | | | | | | |

The scheme reflects the view that primary stress weighs more heavily than secondary stress. The three tonemes are regarded as equally different from one

[5] The reader might expect that the circumflex should be noted by a [ˆ], but this symbol is reserved for a tone with a falling contour in the IPA system.

[6] Here the [w] and the [j] are considered as consonants.

another. In our scheme this distance is equal to 0.500. Because all tonemes imply primary stress, the distance between the toneme of an accent variety and the primary stress of a non-accent variety should not be too large. Therefore, in the scheme the relative small distance of 0.250 is found. With respect to 'nothing' the tonemes weigh a little bit more than primary stress only: 1.000 versus 0.750.

## 3.4.2 Quantity

In this section we discuss the processing of quantity marks: extra-short, half-long, long and syllabic. In the RND extra-short sounds are noted in superscript or with a smaller character. In the IPA system extra-short sounds are noted with a ˘ on top of the sound symbol. In both the RND and the IPA system half-long sounds are followed by a ˑ, long sounds are followed by a ː.

### 3.4.2.1 Extra-short

H & H process extra-short sounds in two ways. First, for some few scattered cases in Groningen dialects and eastern and western Flemish dialects the odd cases in the transcriptions were taken into account while the even cases were ignored. Second the feature table was extended with specifications for the extra-short versions of the [r], [ɣ], [m], [n], [ŋ] and [h]. For these sounds both the odd and even cases are processed, using the specifications of the extra-short versions as given in the feature table. In the specifications the values of non-redundant positively marked features are halved.

It was not clear to us why only a restricted set of sounds may be processed as extra-short. In our research, all sounds may be processed as extra-short. Furthermore, the way in which H & H process extra-short sounds works for feature-based comparison methods, but not for phone-based methods. Therefore, we used another way of processing them which works for phone-based representations as well, and gives the same effect as the approach of H & H when using the feature frequency method. In our approach the half weighting of an extra-short sound with respect to other sounds is realized by changing the transcription beforehand. We retain the extra-short sounds as they are and double all other sounds. E.g. the Dutch word *arm* 'arm' is sometimes pronounced as [ɑrə̆m]. This word is processed as [ɑɑrrəmm]. The Dutch word *timmerman* 'carpenter' is sometimes pronounced as [tɪmər̆mɑn]. This word is changed in [ttɪɪmməərmmɑɑnn].

### 3.4.2.2 Half-long, long (1)

In the feature table of H & H all *monophthongs* are defined for three lengths: short, half-long and long. In the table half-long and long vowels are not really distinguished. In fact half-long vowels are processed as long vowels. Both half-long and long vowels are defined as [+long] (and [+peripheral] for peripheral vowels). For the RND data this is a sound approach to eliminate the influence of the different uses of length marks per transcriber or per atlas part. So for the RND data we follow H & H when deriving feature representations for these sounds.

However, for the NOS data we do distinguish half-long sounds from long sounds. Using the system of H & H half-long NOS vowels are symbolized as [∗long] (and [∗peripheral] for peripheral vowels). We recall that the '∗' is used to signify an intermediate value. In the system of V & C the feature *long* is set to 2, in the system of A & B to 0.5. Half-long RND vowels and long RND and NOS vowels are specified as [+long] (and [+peripheral] for peripheral vowels) in the system of H & H. In the system of V & C the feature *long* is set to 3, in the system of A & B to 1. Note that in the system of V & C length is weighted more heavily than in the two other feature systems.

For the [ɘ], [ʉ], [ɵ] and [ɜ] only versions without length marks are defined in the feature table of H & H. Furthermore, in Section 3.1.3 we saw that in the original V & C system not all vowels can have all lengths. In our adapted systems, for both the RND and the NOS for all monophthongs all lengths are allowed and processed in the way we explained above.

In the feature system of H & H "length" is also processed for a restricted number of *consonants*. While in the RND the nasals ([n], [m], [ŋ]) are noted as long, H & H define them as syllabic sounds by specifying them as [+vowel] and [+syllabic].[7] Long nasals are always treated as syllabic, regardless of the context in which they appear, following H & H. When using the systems of V & C and A & B, we treat such nasals as syllabic sounds as well. In the two systems the feature *syllabic* is set to 1. For half-long nasals H & H ignore the length mark, possible because they judge a half-long nasal as too similar to a short nasal. Here again we follow H & H. For other consonants, H & H did not process length marks. For Dutch varieties it is not common to lengthen other consonants. In the RND some half-long non-nasal consonants can be found when they simultaneously form the last segment of the one word and the first segment of the next word. Since these are rare cases on the one hand, and as we limit our study to single words on the other hand, we also ignore length symbols of non-nasal consonants.

Although it is justified for the RND to interpret a long nasal as a syllabic sound, in a more general approach long nasals and syllabic nasals should be distinguished. In the NOS data we found that *the northwind* (nordavinden) was pronounced as [$^{2}$nuːɾɑˌʋinːn̩] in the dialect of Oslo. This makes clear that long nasals and syllabic nasal are not the same. Therefore, for the NOS data we did not process long nasals as syllabic, but expect that the transcriber has put a syllabic mark [ ̩ ] under a sound if it should be interpreted as syllabic. The problem how to process (half-)long nasal and non-nasal consonants remains, for

[7] In the feature table of H & H syllabic consonants are specified as [+vowel] *and* [+consonant]. Syllabicity makes consonants more vowel like.

example geminates. For both vowels and consonants length refers to duration. Even so, we have the feeling that the feature which represents vowel length should not be used for consonant length. Therefore, for the NOS data consonant length is not processed, unless an author transcribes for example a [tː] as [tt]. We did not find this type of notations in the NOS data.

As we saw, when using a feature-based representation, length can easily be processed by changing one or more feature values. However, this does not work when using the phone-based representation. In some languages length is redundant to some extent. E.g., in Standard Dutch the [e], [a] and [o] (written as <ee>, <aa> and <oo> respectively in closed syllables) are usually long, while the [ɛ], [ɑ] and [ɔ] (written as <e>, <a> and <o> respectively in closed syllables) are short. Therefore, for the phone-based representation we experimented with an approach in which half-long and long are not processed.

#### 3.4.2.3 Half-long, long (2)

Although half-long and long may sometimes be redundant to some extent, this will never consistently be the case. When ignoring both length marks, there is no difference between e.g., the Dutch word *ver* [fɛr] 'far' and *fair* [fɛːr] 'fair'. In this section we present a second approach for processing *half-long* and *long* which we examined in addition to the approach which we described in the previous section. The benefit of this approach is that *half-long* and *long* can be processed not only when using a feature representation, but also when using the phone representation as well.

In Zwaardemaker and Eykman (1928) it was found that the duration of short vowels in Dutch is 40-50% of that of long vowels (p. 298). In a study of durational properties of vowels in Dutch, Nooteboom (1972) found that the duration of long vowels is about two times the duration of short vowels (p. 115). However, this ratio may be affected by stress, the position within the word, position within the sentence, speech rate, etc. For Norwegian Fintoft (1961) found that the duration of short vowels is 53% of that of long vowels (p. 24). This ratio is based on a number of nonsense words built up on the structural principles of real Norwegian words. The words were read by Norwegian speakers. More ratios of short to long vowels for different languages can be found in Elert (1964).

Although the duration ratios may vary per language and under different conditions, we take as a starting hypothesis that the duration of long vowels is twice the duration of short vowels. The duration of half-long vowels is intermediate between the duration of short and long vowels. Analogous to vowels, for short, half-long and long consonants the same ratios are taken as starting point. This implies that a long consonant is processed as if it were just as long as a long vowel. In a more refined system vowel-consonant ratios should also be taken into account. In the present system the different durations are implemented by changing the transcription. Above we explained that extra-short sounds are retained

```
if extra-short
  then retain sound
  else if normal
         then double sound
         else if half-long
                then treble sound
                else if long
                       then quadruple sound
                       else{nothing}
```

Figure 3.1: Procedure followed when more than one length mark is noted for the same phone.

as they are, and short sounds are doubled. In this outline we treble half-long sounds, and quadruple long sounds. In this approach, the suprasegmentals *half-long* and *long* are processed for all vowels and all consonants. E.g. the Dutch word *ook* 'ook' is pronounced as [oːk]. This is changed into [ooookk]. In the Sardinian dialect of Abbasanta the word for 'water' is pronounced as [abˑa]. This becomes [aabbbaa]. In the same dialect the equivalent for 'then' is pronounced as [asːɔɾa]. This is changed in [aassssɔɔɾɾaa].

We applied this approach to both the phone-based and feature-based representations. We are aware of the fact that length is heavily weighted in this procedure, but judge that length plays a rather strong role in perception. E.g. it is for a listener striking when a speaker lengthens vowels at positions where the listener himself would not.

For those cases where a transcriber unfortunately noted more than one length mark for one phone, we follow the procedure as given in Figure 3.1.

#### 3.4.2.4 Syllabic

In the RND consonants may also be vocalized. We process vocalized sounds as syllabic sounds. Vocalized (RND) or syllabic sounds (NOS) are marked with the diacritic *syllabic*. 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 we have to decide how to deal with it since it occurs in both the RND and the NOS transcriptions. We consider two approaches for processing syllabic sounds which corresponds with the two approaches which are regarded when processing *half-long* and *long*.

In the first approach *syllabic* is not processed at all when using the phone representation. This is potentially interesting since *syllabic* may be redundant. E.g. a nasal after a stop at the end of a syllable can hardly be pronounced as a non-syllabic consonant. However, when using feature systems, this diacritic can easily be processed by changing feature values. In the system of H & H a syllabic sound is specified as [+vowel] and [+syllabic]. In the system of both V & C and A & B the feature *syllabic* is set to 1.

In the H & H feature table only for the [m], [n], [ŋ], [r] and [l] syllabic versions are specified. For the RND data we retained the same restriction in the use of the diacritic *syllabic*. For the NOS data we do not check whether the transcriber noted the syllabic diacritic [ ̩ ] under a nasal, trill or lateral approximant but assume a correct use of this diacritic by the transcriber. For the NOS data the same way of processing is followed as for the RND.

In the second approach *syllabic* is processed by changing the transcription. In this way *syllabic* is processed for both the phone and the feature representation. Syllabic sounds are processed as long sounds, i.e. they are quadrupled. Here for both the RND and the NOS data the correct use of this diacritic is not checked, but is assumed to be the responsibility of the transcriber.

## 3.4.3 Place of articulation

### 3.4.3.1 Advanced, retracted

In the RND vowels and consonants can be followed by a [ ˧ ] or a [ ˫ ], which means respectively 'more to the back' and 'more to the front'. The same diacritical marks are found in the IPA system, but there they represent respectively the diacritics *advanced tongue root* and *retracted tongue root*. Following H & H, we did not process these diacritics for the RND data, since their use is probably too transcriber-dependent. For the NOS data these diacritics were ignored as well. It was not clear how these diacritics should be processed. However, in the NOS data the diacritical marks [ ̟ ] and [ ̠ ] also appear, representing respectively the diacritics *advanced* and *retracted*. We processed them for vowels only. We only found a satisfying way to process them in the systems of V & C and A & B.

Using V & C the feature *advancement* is decreased by 1 for an advanced vowel and increased by 1 for a retracted vowel. For the A & B system the feature *advancement* is decreased by 0.5 for an advanced vowel and increased by 0.5 for a retracted vowel. The different weighting of V & C and A & B are due to the different weighting of the feature *advancement*. For both systems the result is that e.g. the [i̠] and the [ɨ̟] are equal, both located exactly in the middle between the [i] and the [ɨ]. Using V & C *advanced* is not processed for vowels with *advancement*=2 (front) while *retracted* is not processed for vowels with *advancement*=6 (back). In the A & B system *advanced* is not processed for vowels with *advancement*=1 (front) while *retracted* is not processed for vowels with *advancement*=3 (back).

#### 3.4.3.2 Raised, lowered

In the RND vowels and consonants can be followed by a [ ̝] or a [ ̞], which means respectively 'more closed' and 'more open'. In the IPA system the same diacritical marks are found, representing the diacritics *raised* and *lowered*. Following H & H, we did not process these diacritics for the RND data, since their use is probably too transcriber-dependent. However, for the NOS data they are processed. We processed them for vowels only. We only found a satisfying way to process them in the systems of V & C and A & B.

Using the feature system of V & C the feature *high* is increased by 0.25 for a raised vowel and decreased by 0.25 for a lowered vowel. For the A & B system the feature *height* is decreased by 0.5 for a raised vowel and increased by 0.5 for a lowered vowel. The different weighting for both systems is due to the different weighting of the features *high* and *height*. For both systems the result is that e.g. the [ɛ̞] and the [æ̝] are equal, both located exactly in the middle between the [ɛ] and the [æ]. Using the V & C system *raised* is not processed for vowels with *high*=4 (close) while *lowered* is not processed for vowels with *high*=1 (open). For the system of A & B *raised* is not processed for vowels with *high*=4 (close) while *lowered* is not processed for vowels with *high*=1 (open).

#### 3.4.3.3 Labialized, palatalized, velarized, pharyngealized

In the feature table of H & H the sequence [tj] and the [tʲ] are specified as a [c]. The [c] is only used in part 16 of the RND, therefore, it is obvious to interpret and to process the [tj] and the [tʲ] as substitutes for the [c]. In our research in the RND transcriptions we replaced all [tj]'s and [tʲ]'s by the [c]. In the RND data the diacritic *palatalized* is noted by putting a dot on top of or below the consonant, or by putting a [ ̥ ] below the consonant. Except for the [t], as just explained, we did not process this diacritic, following H & H. The use of this diacritic may be too transcriber dependent. In the NOS data the diacritic *palatalized* was not found.

However, for the NOS data the diacritics *labialized* (ʷ), *palatalized* (ʲ), *velarized* (ˠ) and *pharyngealized* (ˤ) are taken into account when using the feature systems of V & C and A & B. They are processed by changing the feature *place*. The new place is based on the bit representation of the original place of articulation and the bit representation of respectively bilabial, palatal, velar and pharyngeal. The original place of articulation is weighted for 75% and the secondary place of articulation for 25%. Since e.g. a velarized [t] is still recognized as a [t] and not as a [k], the original place of articulation should be weighted more heavily then the secondary place of articulation. The weightings are applied to each bit of each pair of bits separately. This assures that the bit representation of the place of a velarized [t] is distinguished from one of the existing places between alveolar and velar (see Section 3.6.2 for a more extended explanation).

## 3.4.4 Manner of articulation

### 3.4.4.1 Apical

We found no way to process the diacritic *apical* using phones (e.g. s̺). However, when using a feature system it is possible to process this diacritic. In the RND data this diacritic is not used. However, when processing data based on the modern IPA system such as the NOS data, the occurrence of this diacritic will be taken into account. In the NOS data this diacritic is not found, but in data of e.g. Roman languages this mark may occur. In all three feature systems we added an extra feature *apical.* In the H & H system apical sounds are specified as [∗apical], in the V & C and A & B systems the feature *apical* is set to 0.5.

### 3.4.4.2 Nasalized

In the RND data sounds can be nasalized (e.g. [ã]), but also half-nasalized (e.g.[ã̸]). Half-nasalized sounds may be conceived of being produced with the velum in an intermediate position between fully raised and fully lowered. In the feature table of H & H a nasalized and a half-nasalized version is defined for all monophthongs (possibly for different lengths) and centering diphthongs. In our research both diacritics may also be applied on closing diphthongs and consonants. The use of the diacritic *half-nasalized* is specific for the RND data, in the NOS data only the diacritic *nasalized* is used.

H & H pay special attention to instances of the combination (half-)nasalized vowel + (extra-short) nasal consonant (Hoppenbrouwers and Hoppenbrouwers, 2001, p. 46). All possible combinations are listed in Table 3.16. If we understand the explanation of H & H correctly, then the combinations 1, 3 and 4 are impossible because of undervaluing nasality, and the combinations 6, 9 and 10 are not possible due to overrating nasality. Only the combinations 2, 5, 7 and 8 are possible. H & H corrected the combinations which they judged to be impossible. We are not convinced that all of the combinations are impossible. Therefore, we made no changes in the transcription and process nasality as given by the transcriber.

In the feature table of H & H half-nasalized sounds are specified as [∗nasal], and nasalized sounds as [+nasal]. For both V & C and A & B the feature *nasal* is set to 0.5 for half-nasalized vowels and to 1 for nasalized vowels. Note that for both systems the feature *nasal* is a vowel feature. So with the use of this feature it is not expressed that a nasal consonant is more related to a nasalized vowel than to a non-nasalized vowel, which is a disadvantage of the fact that vowel features and consonant features are separated.

| | | | |
|---|---|---|---|
| 1. | Not nasalized vowel | + | extra short nasal |
| 2. | Not nasalized vowel | + | nasal |
| 3. | Half nasalized vowel | + | nothing |
| 4. | Half nasalized vowel | + | non-nasal |
| 5. | Half nasalized vowel | + | extra-short nasal |
| 6. | Half nasalized vowel | + | nasal |
| 7. | Nasalized vowel | + | nothing |
| 8. | Nasalized vowel | + | non-nasal |
| 9. | Nasalized vowel | + | extra-short nasal |
| 10. | Nasalized vowel | + | nasal |

Table 3.16: All possible combinations of a (half-)nasalized vowel and an (extra-short) nasal consonant.

#### 3.4.4.3 No audible release

The diacritic *no audible release* (e.g. [d˺]) appears in the NOS data. However, this diacritic was not processed since it was not clear how it could be processed in the feature system.

### 3.4.5 Voice

#### 3.4.5.1 Aspiration

H & H did not process the diacritic *aspirated* in their feature system. Possibly the use of this diacritic in the RND is too transcriber-dependent. We follow H & H and ignore this diacritic when processing the RND data. However, for the NOS data source (where all data is transcribed by one transcriber) it is processed. An [h] is inserted after the phone which was noted to be aspirated. This [h] is noted as extra-short, so the weighting is halved. When using feature systems, another way to process *aspirated* would have been to use an extra feature. In that case the weight of aspiration would be much lower. However, our approach reflects the fact that an aspirated sound is perceived as a sound followed by a small [h]. The rather strong weighting accords with the fact that people who aspirate sounds are quickly associated with certain regions.

#### 3.4.5.2 Voiceless, voiced

H & H mention that phonological interpretation systematically seems to play a role in the RND (Hoppenbrouwers and Hoppenbrouwers, 2001, p. 31). In case of assimilation of voice this results in notations like [nit v̥eːl] *niet veel* ‘not much’.

In that case H & H replace the voiceless [v] by a [f]. We chose a more cautious approach. When using phones, the diacritics are ignored. However, when using the feature system of H & H, both a voiceless version of a normally voiced consonant (e.g. [v̥]) and a voiced version of a normally voiceless consonant (e.g. [f̬]) are specified as [∗voiced], which means that the consonant is half-voiced and half-voiceless. When using the feature systems of V & C or A & B the feature *voice* is set to 0.5. The procedure for phones and features as described here applies for both, the RND and the NOS data.

#### 3.4.5.3 Breathy, creaky

For the NOS data only the diacritics *breathy voiced* (e.g. [ə̤]) and *creaky voiced* (e.g. [æ̰]) are processed. They are only processed when using a feature representation, since we found no possibility of processing them when using the phone representation. Since these diacritics should not weigh too heavily, we assign only a value of 0.25 to the features *breathy* and *creaky* respectively. These weightings are chosen intuitively. Both diacritics are only processed when noted below a vowel.

## 3.5 Redundancy

We only find rules that are applied to remove redundancy from feature bundles in the feature system of H & H. This can be explained from the fact that the system was originally developed to be used for the feature frequency method in which features are counted. H & H write that a text which contains many coronals and anteriors, will also contain many consonants, and that in a language with a relatively great number of vowels there will be relatively less space for consonants (Hoppenbrouwers and Hoppenbrouwers, 2001, p. 43). Therefore, a feature specification may contain redundant information. H & H decided to ignore feature values which may be predicted on the basis of other feature values using rules.

In the feature table the presence of a feature is indicated by a '+' and the absence by a '−'. Also a '∗' can be used which signifies intermediate values, needed to express that a feature value changes during the realization of the segment (see Section 3.2), or when a property is only weakly present (e.g. half nasalized, see Section 3.4.4.2). However, when presence or absence can be predicted on the basis of one or more of the other features, respectively a '0' and '1' are noted. When H & H apply their feature frequency method, redundant positive marked features (1's) are processed as absent.

In our implementation only 0's, ∗'s and 1's are used. The ∗'s and 1's have the same meaning as respectively the ∗'s and +'s of H & H, and a 0 means that either the feature is absent or redundant or both (the −'s, 1's and 0's of H & H).

H & H give five rules, where rule 1 and rule 2 are subdivided in four subrules. Because of the restriction to and the different meaning of 0's, $*$'s and 1's in our implementation, we first have to perform the rules 2a to 2d, and the rules 1a to 1d afterwards. In this way, we get the same effect as when following the operating procedure of H & H exactly. Table 3.17 contains the rules as given by H & H, extended with some more rules, which we will justify below.

The rules 1a to 1e indicate that the features [+front] ... [+polar] predict the feature [+vowel]. For the sake of closing diphthongs we added the rules 1f to 1j, and for the purpose of the centering diphthongs we added the rules 1k to 1o.

The rules 1e, 1j and 1o are added for the sake of sounds marked as [−front −back −round −low +polar]. Sounds with these specifications do not appear in the feature table of H & H because it contains only sounds which appear in the RND. However, in view of the NOS, we extend the system so that all IPA vowels can be processed.

Vowels are always marked as [+sonorant +voiced +continuant +syllabic]. This is indicated in the rules 2a to 2d. We added the condition [−consonant]. Syllabic consonant are also specified as [+vowel], however, since they are also [+consonant], they are excluded by this extra condition, in accordance with what is suggested by the feature table of H & H. For the benefit of the centering diphthongs we added the rules 2e to 2h.

H & H mention that for diphthongs the feature [+long] is redundant. From the feature table it can be concluded that only closing diphthongs are intended to fall under this rule. Centering diphthongs are – perhaps surprisingly – specified as [−diphthong]. The prediction of [+long] for closing diphthongs is reflected in rule 3.

For consonants pronounced in the back the feature [+high] is superfluous: This is reflected in rule 4a. From the feature table follows that also rule 4b applies.

H & H write that for the sonorant laryngeal, the guttural [ɦ] the feature [+continuant] can be predicted. They suggest as rule: [+laryngeal +sonorant] → +continuant. However, in the feature table the [ɦ] is specified as [−sonorant +voiced]. Upon investigation, it appears that as a second condition it was not [+sonorant], but [+voiced] that was meant. In our overview the rule is corrected and given as rule 5.

From the feature table it appears that for vocalized (or syllabic) consonants the feature [+vowel] is redundant. Therefore, we added rule 6.

H & H do not give rules which predict negative marked features since there is no real difference between absent features and redundant features when processing the feature specifications.

| | | | | |
|---|---|---|---|---|
| 1a | +front | | → | +vowel |
| 1b | +back | | → | +vowel |
| 1c | +round | | → | +vowel |
| 1d | +low | | → | +vowel |
| 1e | +polar | | → | +vowel |
| 1f | *front | +diphthong | → | +vowel |
| 1g | *back | +diphthong | → | +vowel |
| 1h | *round | +diphthong | → | +vowel |
| 1i | *low | +diphthong | → | +vowel |
| 1j | *polar | +diphthong | → | +vowel |
| 1k | *front | −diphthong | → | *vowel |
| 1l | *back | −diphthong | → | *vowel |
| 1m | *round | −diphthong | → | *vowel |
| 1n | *low | −diphthong | → | *vowel |
| 1o | *polar | −diphthong | → | *vowel |
| 2a | +vowel | −consonant | → | +sonorant |
| 2b | +vowel | −consonant | → | +voiced |
| 2c | +vowel | −consonant | → | +continuant |
| 2d | +vowel | −consonant | → | +syllabic |
| 2e | *vowel | −consonant | → | +sonorant |
| 2f | *vowel | −consonant | → | +voiced |
| 2g | *vowel | −consonant | → | +continuant |
| 2h | *vowel | −consonant | → | +syllabic |
| 3 | +diphthong | | → | +long |
| 4a | +posterior | | → | +high |
| 4b | *posterior | | → | +high |
| 5 | +laryngeal | +voiced | → | +continuant |
| 6 | +syllabic | +consonant | → | +vowel |

Table 3.17: Redundancy rules which predict positive feature specifications as used in the feature system of H & H.

## 3.6 Comparison of segments

### 3.6.1 Phones

Calculating sound distances on the basis of the phone representation is trivial. There exist only two distances: 0 (phones are equal) and 1 (phones are different). So the distance between e.g. a [p] and an [a] is 1 unit, the distance between a [p] and a [b] is 1 unit, and also the distance between an [au] and an [a] is 1 unit. In the approach of Kessler (1995) two phones which are basically equal but have different diacritics, are regarded as different phones. So [a] versus [aː] costs 1 unit just as [a] versus [p] costs 1 unit. Our approach only deals with basic symbols. Suprasegmentals and diacritics can only be taken into account by changing the transcription beforehand (see Section 3.4). Using the resulting transcription only the basic symbols are processed. This approach is motivated by the idea that we should take care not to overvalue the influence of suprasegmentals and/or diacritics and retain the relation between a sound with and without one or more additional marks. So in our research an [a] and an [aː] are considered to be equal.

### 3.6.2 Features

One of the properties of the feature system of H & H is that all features are binary. H & H developed their system for their feature frequency method. In the systems of V & C and A & B also multivalued features are used. A disadvantage of multivalued features is that they may neutralize each other. In this section we describe how a multivalued feature can be changed into a vector of binary features. The neutralizing effect is illustrated on the basis of the three cases where the effect was found. For each of the cases we show how this effect is eliminated when using binary vectors. Finally we describe how the distance between two histograms or between two feature bundles is actually calculated, using the binary vector representation.

#### 3.6.2.1 Vector representation per feature value

Comparing the feature representations in the system of H & H with the systems of V & C and A & B, we see that H & H only use values 0 and 1, represented by + and −. For diphthongs and extra-short sounds also the value 0.5 is used, represented by a ∗. In the two other systems a wider range of values is used. In the three feature systems vowel advancement for instance is defined as follows:

| | H & H | | V & C | A & B |
|---|---|---|---|---|
| | front | back | advancement | advancement |
| front | 1 | 0 | 2 | 1 |
| central | 0 | 0 | 4 | 2 |
| back | 0 | 1 | 6 | 3 |

Where V & C and A & B use one multivalued feature, H & H use two binary features. In the V & C system the values of the A & B system are weighted two times. The weighting in the H & H system is the same as in the A & B system. In fact H & H give a vector representation for the A & B values. In general the following applies: one feature having $n$ integers with stepsize 1 can always be converted to a vector of $n - 1$ binary values. Other possibilities of representing the A & B advancement feature by a binary vector are:

| | value 1 | value 2 |
|---|---|---|
| front | 0 | 0 |
| central | 1 | 0 |
| back | 1 | 1 |

and:

| | value 1 | value 2 | value 3 |
|---|---|---|---|
| front | 1 | 0 | 0 |
| central | 1 | 1 | 0 |
| back | 1 | 1 | 1 |

The first representation is most efficient. However, we prefer the second representation. The idea behind the second is: 1 is represented by one 1, 2 by two 1's, 3 by three 1's, etc. If necessary, the value 0 can also be used and represented as a vector containing only 0's. In that case the following applies: a multivalued feature which may have as its highest value the value $n$, and which only contains integers, can always be converted to a vector containing $n$ binary features.

As mentioned above, besides the 0 and 1 H & H use also the value 0.5. When processing suprasegmentals and diacritics (see Section 3.4) it appeared that we needed some more fractions. In our research we use the following fractions: 0.125, 0.250, 0.375, 0.500, 0.625, 0.750 and 0.875.

#### 3.6.2.2 Summing feature values

When using the frequency-based corpus frequency method or the frequency per word method per feature the values as specified for the segments in the corpus or in the word are added. When using multivalued features, low and high values may neutralize each other. Assume a hypothetical example where one dialect has one vowel which is fronted and one vowel which is back. Another dialect has two vowels which are central. Using the A & B system, the sum of the values for the feature advancement in the first dialect is $1 + 3 = 4$, and in the second dialect $2 + 2 = 4$. This suggests erroneously that the two dialects are equal when only considering the feature advancement. When using binary vector representations this will not be the case. We use the vector representation as suggested in the previous section. In this representation the feature advancement is split in three binary features. Now for each dialect the values of the features of the one vector are added to the values of the corresponding features of the other vector:

| | v1 | v2 | v3 |
|---|---|---|---|
| front | 1 | 0 | 0 |
| back | 1 | 1 | 1 |
| *sum* | 2 | 1 | 1 |

| | v1 | v2 | v3 |
|---|---|---|---|
| central | 1 | 1 | 0 |
| central | 1 | 1 | 0 |
| *sum* | 2 | 2 | 0 |

Next we calculate the absolute differences between the corresponding sum values of the one dialect and those of the other dialect:

| | v1 | v2 | v3 |
|---|---|---|---|
| dialect 1 | 2 | 1 | 1 |
| dialect 2 | 2 | 2 | 0 |
| *abs. diff.* | 0 | 1 | 1 |

The distance between the two dialects is found by taking the sum of the absolute differences: $0 + 1 + 1 = 2$. This outcome shows correctly that the two dialects are different.

### 3.6.2.3 Representation of diphthongs

As mentioned in Section 3.2, the definition of diphthongs is based on the features corresponding with the start position and with the end position. To be more concrete, the average feature values of the values of the start position and those of the end position are taken. Assume the four degrees of height are defined as 1 (close), 2 (close-mid), 3 (open-mid) and 4 (open). Now we want to find the correct height value for the [ɛi]. The [ɛ] is open-mid (3), the [i] is close (1). The average would be $(1 + 3)/2 = 2$, i.e. close-mid. However, the [e] is close-mid is well. So we get a neutralizing effect when taking the average feature values, resulting in a specification which suggests that the [ɛi] has the same height as the [e]. Since the [e] is a stable sound, and the [ɛi] is a sound with a changing height, this outcome is undesirable. We found the solution by representing the multivalued feature height as a binary vector. Analogous to the example in Section 3.6.2.1 the different degrees of height are represented as follows:

| | v1 | v2 | v3 | v4 |
|---|---|---|---|---|
| close | 1 | 0 | 0 | 0 |
| close-mid | 1 | 1 | 0 | 0 |
| open-mid | 1 | 1 | 1 | 0 |
| open | 1 | 1 | 1 | 1 |

Using these representations, the values of the corresponding features of the two vectors are averaged:

| | v1 | v2 | v3 | v4 | |
|---|---|---|---|---|---|
| [i] | 1 | 0 | 0 | 0 | ×50% |
| [ɛ] | 1 | 1 | 1 | 0 | ×50% |

gives:

| | v1 | v2 | v3 | v4 |
|---|---|---|---|---|
| | 0.5 | 0 | 0 | 0 |
| | 0.5 | 0.5 | 0.5 | 0 |
| [ɛi] | 1 | 0.5 | 0.5 | 0 |

Next we calculate the absolute differences between the corresponding vector values of the [ɛi] and the [e]:

| | v1 | v2 | v3 | v4 |
|---|---|---|---|---|
| [e] | 1 | 1 | 0 | 0 |
| [ɛi] | 1 | 0.5 | 0.5 | 0 |
| *abs. diff.* | 0 | 0.5 | 0.5 | 0 |

The sum of the differences is $0 + 0.5 + 0.5 + 0 = 1$, which shows that the stable [e] and the changing [ɛi] are different.

#### 3.6.2.4 Representation of the place of articulation

Using vector representations for multivalued features is also useful when finding the representation of the place of articulation of a sound with a secondary place of articulation. In our research the primary place of articulation is weighted for 75% and the secondary place of articulation for 25% (see Section 3.4). Assume we have to process a velarized t ([tˠ]). The place of articulation of the [t] is alveolar. Assume alveolar is represented by 1, postalveolar by 2, retroflex by 3, palatal by 4 and velar by 5. Using a single value, the new place of articulation would be $75\% \times 1 + 25\% \times 5 = 2$, which is postalveolar. However, a velarized t is not postalveolar at all. Here we are faced again with the difficulty that values of multivalued features neutralize each other. With the use of vector representations this problem will be solved. Analogous to the example in Section 3.6.2.1 the different places of articulation are represented as follows:

| | v1 | v2 | v3 | v4 | v5 |
|---|---|---|---|---|---|
| alveolar | 1 | 0 | 0 | 0 | 0 |
| postalveolar | 1 | 1 | 0 | 0 | 0 |
| retroflex | 1 | 1 | 1 | 0 | 0 |
| palatal | 1 | 1 | 1 | 1 | 0 |
| velar | 1 | 1 | 1 | 1 | 1 |

Using this representations, we should weight the primary place of articulation 75% (alveolar) and the secondary place of articulation (velar) 25%. When adding the weighted vector values of the primary place of articulation to the corresponding weighted vector values of the secondary place of articulation, we get the place of articulation of the velarized t:

| | v1 | v2 | v3 | v4 | v5 | |
|---|---|---|---|---|---|---|
| alveolar | 1 | 0 | 0 | 0 | 0 | ×75% |
| velar | 1 | 1 | 1 | 1 | 1 | ×25% |
| | | | | | | |

gives:

| | v1 | v2 | v3 | v4 | v5 |
|---|---|---|---|---|---|
| | 0.75 | 0 | 0 | 0 | 0 |
| | 0.25 | 0.25 | 0.25 | 0.25 | 0.25 |
| *sum* | 1 | 0.25 | 0.25 | 0.25 | 0.25 |

Next we calculate the absolute differences between the corresponding vector values of the velarized t and postalveolar:

| | v1 | v2 | v3 | v4 | v5 |
|---|---|---|---|---|---|
| velarized t | 1 | 0.25 | 0.25 | 0.25 | 0.25 |
| postalveolar | 1 | 1 | 0 | 0 | 0 |
| *abs. diff.* | 0 | 0.75 | 0.25 | 0.25 | 0.25 |

The sum of the differences is $0 + 0.75 + 0.25 + 0.25 + 0.25 = 1.5$, which shows that the place of articulation of a velarized t is distinguished from postalveolar.

#### 3.6.2.5 Comparison of segments

The comparison of feature histograms and feature bundles is basically the same. Therefore, the same metric for the comparison of feature histograms is used for the comparison of feature bundles as well. There are several metrics for finding the distance between two feature histograms or feature bundles (Jain and Dubes, 1988; Hoppenbrouwers and Hoppenbrouwers, 1988). We restricted ourselves to the most common ones: Manhattan (or taxicab, or city block) distance, Euclidean distance and Pearson's correlation coefficient.

Assume we compare a histogram or bundle $X$ with a histogram or bundle $Y$ where $n$ is the number of features. The Manhattan distance (Jain and Dubes, 1988) is simply the sum of all feature value differences for each of the n features:

$$\delta(X,Y) = \sum_{i=1}^{n} |X_i - Y_i| \tag{3.1}$$

The Euclidean distance is the square root of the sum of squared differences in feature values:

$$\delta(X,Y) = \sqrt{\sum_{i=1}^{n} (X_i - Y_i)^2} \tag{3.2}$$

The Pearson correlation coefficient (Hogg and Ledolter, 1992) is calculated as follows:[8]

$$r(X,Y) = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \overline{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \overline{Y})^2}} \tag{3.3}$$

When comparing two ranges, if all values are equal in one or both histograms, the correlation coefficient between both ranges is not defined since a division by zero occurs. This never happens in corpus-based histograms, but it is possible that all values in a word-based histogram or in a single feature bundle are the same. Normally the correlation coefficient ranges from −1 (inverse ranges) to +1 (parallel ranges). Therefore, when both feature bundles are constant, we set the correlation coefficient to 1. When only one range is constant, we set the correlation coefficient to 0.

In fact, Pearson's correlation coefficient is a similarity measure. As such Hoppenbrouwers and Hoppenbrouwers (2001) used this metric for the comparison of feature histograms. The minimum value is −1 (minimal similarity) and the maximum value is +1 (maximal similarity). In view of the use of cluster analysis (see 6.1) we only want to use distance metrics. We used the Pearson's correlations coefficient by calculating $1 - r$. In that case the minimum value is 0 (maximal similarity) and the maximum value is 2 (minimal similarity).

We are aware of the fact that the use of the Pearson's correlation coefficient is more correct for the comparison of histograms than for the comparison of feature bundles. Histograms which are parallel to each other show that in the corresponding varieties the different features are positively marked in the same proportions. However, when comparing feature bundles this approach may give the wrong results when ranges are parallel and the one range is consistently higher than the other range. The Pearson's correlation coefficient will give the impression that they are equal. We include measurements of correlation coefficients among feature bundles for the sake of completeness, even though we do not expect it to function well.

[8] Before using this formula, $\overline{X}$ and $\overline{Y}$ should be calculated. This means that for calculating the correlation coefficient at least two passes are needed. In our research we used another formula, which allowed more efficient processing and avoids some of the rounding errors that are made with the earlier formula (Hogg and Ledolter, 1992):

$$r(X,Y) = \frac{\sum_{i=1}^{n} X_i Y_i - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n}}{\sqrt{(\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n})(\sum_{i=1}^{n} Y_i^2 - \frac{(\sum_{i=1}^{n} Y_i)^2}{n})}}$$

## 3.7 Linear and logarithmic distances

Using the feature bundles the distance between two segments can be calculated. The simplest way in which this can be done is taking the sum of the absolute differences of each pair of corresponding feature values. Other metrics are described in Section 3.6.2. Characteristic for the original A & B system is that when a distance exceeds a certain ceiling, that distance is set to the value of a ceiling. However, the question arises as to what value the ceiling should be set to. In Heeringa and Braun (2003) an adjusted version of the Almeida & Braun system is proposed where the logarithm of the feature bundle distances is taken instead of using a ceiling. The effect of taking the logarithm is that small distances are weighted relatively more heavily than large distances. This may be in accordance with our perception, where small differences in pronunciation play a relatively strong role in comparison with larger differences. We experimented with both linear and logarithm feature bundle distances for all three feature systems: the H & H, the V & C and the A & B system.

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 3.2 the effect of taking the logarithm of the IPA vowel distances as found with the A & B system is shown. For each of the 28 IPA vowels the distance with respect to silence is calculated. Next the distances are sorted from short to long. 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. By taking the logarithm, greater distances are decreased to a greater degree than short distances.

In Figure 3.3 the effect of taking the logarithm of the IPA consonant distances as found with the A & B system is shown. For each of the 59 IPA consonants the distance with respect to silence is calculated. Next the distances are sorted from short to long. In both linear and logarithmic distance, [ʔ] is most like silence and [w] is most unlike silence. The graph shows sorted distances in the same way as was shown for the vowels. The points corresponding with distances are connected by lines. Of course the same effect as for the vowels is seen here: greater distances are decreased to a greater degree than shorter distances when taking the logarithm.

We only apply the logarithm to segment distances as used in the Levenshtein distance (see Section 5.1), not to histogram distances as calculated in the corpus frequency method or the frequency per word method. In the corpus feature frequency method the distance between two histograms corresponds to a dialect distance, and in the feature frequency per word method the distance between two histograms corresponds to a word distance (see Section 2.3.2 and Section 2.3.3).

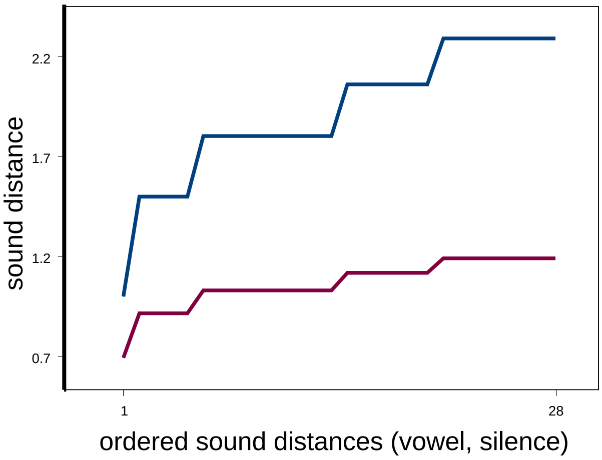


Figure 3.2: Linear (upper) and logarithmic (lower) A & B distances of 28 IPA vowels with respect to silence. Distances are calculated as the sum of the differences between corresponding features. The graph shows the distances sorted from low (left) to high (right). Greater distances are reduced more than smaller ones by using the logarithm.

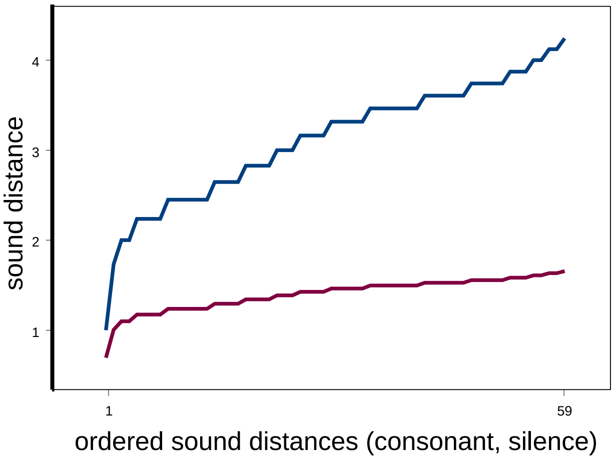


Figure 3.3: Linear (upper) and logarithmic (lower) A & B distances of 59 IPA consonants with respect to silence. Distances are calculated as the sum of the differences between corresponding features. The graph shows the distances sorted from low (left) to high (right). Greater distances are reduced more than smaller ones by using the logarithm.

The result of taking logarithmic distances is that great dialect distances (in the corpus feature frequency method) or great word distances (in the feature frequency per word method) will stay relatively small. For dialect distances and word distances we judge this to be undesirable.

Special attention should be paid when the [i], [j], [u] and [w] are compared to each other when using the feature systems of V & C and A & B. In Section 3.1.3.1 we described that these sounds are defined as both vowel and consonant. When a doubly defined segment is compared to another doubly defined segment, the distance is equal to the average of the vowel distance and the consonant distance. When calculating the logarithmic distance, we calculate this as:

$$\frac{ln(vowel\ distance + 1) + ln(consonant\ distance + 1)}{2}$$

## 3.8 Correlation between systems

In this section we compare the different feature systems of H & H (see Section 3.1.2), V & C (see Section 3.1.3) and A & B (see Section 3.1.4). For each of the systems all vowels and consonants are specified. We ask the question to what degree the various systems for calculating segment distances correlate in the distances they assign, and whether there remain interesting differences as well. The distances between vowels and consonants are calculated using the three metrics which are proposed in Section 3.6.2: the Manhattan distance (M.), the Euclidean distance (E.) and the Pearson correlation coefficient (P.). For the H & H system redundancy is removed from the feature bundles (see Section 3.5). For all systems the linear distances are used (see Section 3.7).

In Section 3.8.1 we give an overview of all matrices among which the correlation coefficients are calculated. Section 3.8.2 explains how to determine if a correlation coefficient is significant. In this section we also explain how to determine whether two correlation coefficients are significantly different. In Section 3.8.3 we examine the influence of the different feature systems while in Section 3.8.4 the influence of different feature bundle metrics is investigated.

In Section 7.4.3 the influence of different feature systems and different feature bundle metrics will be further examined.

### 3.8.1 Matrices

When correlating results of one particular segment representation using one particular feature bundle metric with results of another segment representation using a feature bundle metric, this is done on the basis of distances which are arranged in a matrix. The matrices may have four different sizes.

In the RND data 18 vowels are used. Although the vowel [œ] does not appear in RND transcriptions, it is included because it is used for the definition of diphthongs (see Table 3.14). Calculating the distances between the vowels results in $(18 \times 17)/2 = 153$ distances. In the RND 27 consonants are used. When calculating the distances between the consonants we get $(27 \times 26)/2 = 351$ distances.

In the NOS data the modern IPA system is used. In the IPA system 28 vowels are given. If we calculate the distances between the vowels, we get a distance matrix of $(28 \times 27)/2 = 378$ distances. In the IPA system 58 pulmonic consonants are given. We added the [w] which is ordered under 'Other Symbols' in the IPA system, so we get 59 consonants. When calculating the distances between the consonants, a distance matrix of $(59 \times 58)/2 = 1711$ distances is obtained.

### 3.8.2 Significance

For finding the significance of a correlation coefficient we used the Mantel test. In classical tests the assumption is made that the objects which are correlated are independent. However, values in distance matrices are usually correlated in some way, and not independent (Bonnet and Van de Peer, 2002). A widely used method to account for distance correlations is the Mantel test (Mantel, 1967). Mantel developed an asymptotic test, in which the null hypothesis is that distances in the one matrix are independent of the corresponding distances in the other matrix. The significance of the statistic can also be evaluated by randomly reallocating the order of elements in one of the matrices (Bonnet and Van de Peer, 2002).

The program we used is also based on a series of random permutations. Assume we have two matrices $D_1$ and $D_2$. We would like to know whether $r(D_1, D_2)$ is significant. To determine the significance a number of iterations is performed. In each iteration the order of the elements of $D_1$ is changed by swapping each element with another element where the other element is randomly chosen. In fact it does not matter whether $D_1$, $D_2$ or both are randomly permuted. Next the following condition is tested:

$$r(P_1, D_2) > r(D_1, D_2)$$

If this condition is true, a counter is incremented by 1. After the iterations are finished, the counter is divided by the number of the iterations. The outcome gives the chance that randomly permuted matrices correlate more strongly than the two unchanged matrices. The number of iterations determine the overall precision of the test. Since we use $\alpha = 0.05$, the number of repetitions should be equal to about 1000 (Manly, 1997).

Besides finding the significance of a correlation coefficient we would like to know whether one correlation coefficient is significantly higher than another. Assume we have four matrices $D_1$, $D_2$, $D_3$ and $D_4$. We want to know whether $r(D_1, D_2)$ is significantly higher than $r(D_3, D_4)$. For this purpose, a number of iterations is performed again. In each iteration $D_1$ and $D_3$ are randomly permuted. We call the permuted matrices respectively $P_1$ and $P_3$. Just as in the procedure described above, a random permutation is generated by swapping each element with another element which is randomly chosen. Next the following condition is tested:

$$r(P_1, D_2) - r(P_3, D_4) \geq r(D_1, D_2) - r(D_3, D_4)$$

If the condition is true, a counter is incremented by one. After all iterations are finished, the counter is divided by the number of the iterations. This gives the chance of getting a difference which is equal to or greater than the given difference when using randomly permuted matrices. Just as in the previous procedure, we perform 1000 iterations and a significance level of $\alpha = 0.05$.

Examples of related applications of the Mantel test are found in Barbujani et al. (1994), Weng and Sokal (1995) and Manni (2001). Barbujani et al. (1994) investigated the relation between genetics and linguistics in the Caucasus. Genetic, geographic and linguistic distances were correlated. Weng and Sokal (1995) carried out a lexicostatistics study. In this study a series of tests was undertaken to relate lexicostatistical dissimilarities among 48 Indo-European languages to distances representing various causal hypotheses. The putative causal distance matrices include geographic distances, distances representing the origin of agriculture, and distances representing hypotheses concerning the origin and spread of Indo-European languages in Europe. Manni (2001) compared genetic and linguistic distances for the Italian province of Ferrara and for the Netherlands.

### 3.8.3 Feature representations

In Tables 3.18 and 3.19 the different feature representations can be compared. In the tables correlation coefficients are given, based on pairs of matrices where the distances in each matrix are based on different feature systems and on the same metric for the comparison of feature bundles. The columns divide results in Manhattan, Euclidean and Pearson metrics. Results are given for vowels and consonants for each metric. All correlations are significant for $\alpha = 0.05$.

For both the RND and the IPA it appears that all correlations between the V & C system and the A & B system are stronger – although not significantly stronger – than the corresponding correlations between any other pair of systems. Looking at the vowel features, height is defined in a similar way in the V & C system and the A & B system, while in the H & H system some relationship between most low and most high sounds is defined. For the consonants it holds that in both the V & C system and the A & B system the place of articulation is explicitly defined. This is not the case in the H & H system. In turn most correlations between the H & H system and the V & C system are stronger – but not significantly stronger – than the corresponding correlations between

| | | | M. | | E. | | P. | |
|---|---|---|---|---|---|---|---|---|
| | | | vow. | cons. | vow. | cons. | vow. | cons. |
| H & H | vs. | V & C | 0.77 | 0.68 | 0.77 | 0.71 | 0.59 | 0.68 |
| H & H | vs. | A & B | 0.70 | 0.58 | 0.72 | 0.60 | 0.51 | 0.65 |
| V & C | vs. | A & B | 0.79 | 0.80 | 0.80 | 0.81 | 0.67 | 0.77 |

Table 3.18: Correlation coefficients between RND segment distances obtained on the basis of different feature systems and as calculated using Manhattan (M.), Euclidean (E.) and Pearson (P.) procedures. Results are given per metric for vowels (vow.) and consonants (cons.).

| | | | M. | | E. | | P. | |
|---|---|---|---|---|---|---|---|---|
| | | | vow. | cons. | vow. | cons. | vow. | cons. |
| H & H | vs. | V & C | 0.72 | 0.61 | 0.71 | 0.62 | 0.42 | 0.63 |
| H & H | vs. | A & B | 0.64 | 0.57 | 0.65 | 0.61 | 0.42 | 0.64 |
| V & C | vs. | A & B | 0.80 | 0.76 | 0.79 | 0.76 | 0.71 | 0.74 |

Table 3.19: Correlation coefficients between IPA segment distances obtained on the basis of different feature systems and as calculated using Manhattan (M.), Euclidean (E.) and Pearson (P.) procedures. Results are given per metric for vowels (vow.) and consonants (cons.).

the H & H system and the A & B system. In the A & B system, manner of articulation is defined as a scale. This is not the case in the two other systems.

The correlation coefficients vary from at least 0.42 to at most 0.81. These rather low correlation coefficients show that it remains interesting to take all three feature representations into account in further research, even though the correlation coefficients are significant.

### 3.8.4 Feature bundle metrics

In Tables 3.20 and 3.21 the different metrics can be compared. In the tables correlation coefficients are given, based on pairs of matrices where the distances in each matrix are based on different metrics for feature bundle comparison and on the same feature system. The columns divide results in the H & H, V & C and A & B system. All correlations are significant for $\alpha = 0.05$.

For both the RND and the IPA it appears that all correlations between the Manhattan metric and the Euclidean metric are stronger than the corresponding correlations between any other pair of metrics. Examining the IPA results, the correlations between the Manhattan metric and the Euclidean metric are also significantly stronger when using the feature systems of H & H (vowels and con-

| | H & H | | V & C | | A & B | |
|---|---|---|---|---|---|---|
| | vow. | cons. | vow. | cons. | vow. | cons. |
| M. vs. E. | 0.97 | 0.99 | 0.98 | 0.99 | 0.99 | 0.98 |
| M. vs. P. | 0.89 | 0.90 | 0.90 | 0.94 | 0.94 | 0.95 |
| E. vs. P. | 0.87 | 0.90 | 0.90 | 0.94 | 0.94 | 0.93 |

Table 3.20: Correlation coefficients between RND segment distances obtained on the basis of different metrics. Results are given per feature system for vowels (vow.) and consonants (cons.).

| | H & H | | V & C | | A & B | |
|---|---|---|---|---|---|---|
| | vow. | cons. | vow. | cons. | vow. | cons. |
| M. vs. E. | 0.97 | 0.97 | 0.98 | 0.98 | 0.99 | 0.99 |
| M. vs. P. | 0.80 | 0.90 | 0.90 | 0.94 | 0.95 | 0.94 |
| E. vs. P. | 0.80 | 0.89 | 0.89 | 0.93 | 0.95 | 0.93 |

Table 3.21: Correlation coefficients between IPA segment distances obtained on the basis of different metrics. Results are given per feature system for vowels (vow.) and consonants (cons.).

sonants) and A & B (consonants). We observe that some correlations between the Manhattan metric and the Pearson correlation are higher than the corresponding correlations between the Euclidean metric and the Pearson correlation. However, these correlations are not significantly higher.

When looking at the correlation coefficients between the Manhattan distances and the Euclidean distances, we see that these vary from at least 0.97 to at most 0.99. These are all extremely high values, indicating that it is not necessary to regard both metrics in further research. The correlation coefficient between the Manhattan metric and the Pearson metric vary from 0.80 to 0.95, and between the Euclidean metric and the Pearson metric from 0.80 and 0.95. This indicates that the Pearson correlation coefficient metric is rather different from both other metrics.

## 3.9 Conclusions

In this chapter we proposed different representations of speech segments in order to find the relations among them. The roughest representation is the *phone* representation. In this representation speech segments are simply equal or not equal. There are no gradations. The more refined representation is the *feature* representation. Using this representation, finer segment distances are obtained. We investigated three feature systems, and three metrics for finding distances between feature histograms (used in frequency-based methods) and feature bundles (used in Levenshtein distance).

We found that the V & C system and the A & B system correlate for most metrics more strongly – although not significantly so – than the corresponding correlations between any other pair of systems. For both systems the vowel feature which defines the height is defined in a similar way and the place of articulation is explicitely defined for consonants. The rather low correlations between all three systems show that it remains interesting to take all three feature representations into account in further research, although the correlation coefficients are significant.

It appeared that the correlations between the Manhattan metric and the Euclidean metric are for all feature systems stronger than the corresponding correlations between any other pair of metrics. Some of them were also significantly stronger. The strong correlation indicates that it is actually not necessary to consider both metrics in later work. The Pearson correlation coefficient appeared to be rather different from the two other metrics.

# Chapter 4

# Measuring segment distances acoustically

In Chapter 3 we described three feature systems, namely the system of Hoppenbrouwers & Hoppenbrouwers (H & H), Vieregge & Cucchiarini (V & C) and Almeida & Braun (A & B). The systems can be used for the corpus frequency method (see Section 2.3.2), the frequency per word method (see Section 2.3.3), and the Levenshtein distance (see Section 5.1). The Levenshtein distance is the focus of this thesis. When using the Levenshtein distance the distances may be calculated on the basis of these feature systems.

While the use of features as linguists have developed them yields satisfactory results, one may nonetheless question the physical basis of the feature assignments, in fact seeking a more objective foundation. The advantage of the system of H & H is that vowels and consonants can be compared with each other. For vowels, the more consonant-specific features get default values, and for consonants the more vowel-specific features get default values. The disadvantage of Hoppenbrouwers' system is that feature values are not based on physical measurements. The SPE feature system, which H & H use, was not developed to reflect physical, perceptual or articulatory differences directly, but rather to facilitate the coding of phonological rules. It is of course, to be expected that this coding reflects the physical properties of speech in some way. The advantage of the system of V & C is that it is partly based on real measurements, found by experiments. The system of the A & B is interesting because of its use of the well-known IPA system. However, just as for the H & H system, the IPA system is not based on real measurements.

Another inadequacy of the three feature systems concerns the definition of 'silence', which is needed in the Levenshtein algorithm (see Section 5.1). The way in which 'silent vowels' and 'silent consonants' are defined was described for each system in Chapter 3. A definition of 'silence' in terms of features will always be somewhat artificial.

When acquiring language, children learn to pronounce sounds by listening to the pronunciation of their parents or other people. The acoustic signal seems be to sufficient to find the articulation which is needed to realize the sound. Acoustically, speech is just a series of changes in air pressure, quickly following each other. With a spectrograph or a computer a spectrogram can be made, representing an analysis of the speech sample. A spectrogram is a "graph with frequency on the vertical axis and time on the horizontal axis, with the darkness of the graph at any point representing the intensity of the sound" (Trask, 1996, p. 328). In a spectrogram the formant structure of a sound can be identified. A formant is "a concentration of acoustic energy within a particular frequency band, especially in speech" (Trask, 1996, p. 148). Especially for vowels, formants can be easily recognized in a spectrogram as thick dark bars.

In this chapter we present the use of spectrograms and formant tracks for finding sound distances. We will show that the disadvantages of feature systems as mentioned above do not apply to acoustic representations when plausible segment classifications may be obtained on the basis of acoustic representations. Both spectrograms and formant tracks are based on physical measurements. When using a spectrogram or formant definition instead of a feature definition, the distance between a vowel and a consonant can be measured in the same way as the distance between a vowel and a vowel, or between a consonant and a consonant. When using a spectrogram 'silence' is defined in a natural way: for all frequencies for all times the intensities are equal to 0. Something similar applies for the formant definition of 'silence': there are no vibrations, so the frequencies are set to 0.

We explored the use of the different acoustic representations for finding segment distances which are intended for use in the Levenshtein algorithm. In this section we show how distances between sounds can be found using spectrograms and formant tracks. In Section 4.1 it is shown that spectrograms can be regarded as pictures of sounds. In Section 4.2 we discuss the samples of the sounds we used. Section 4.3 discusses several spectrogram models, as well as the formant track model. Classifications on the basis of the different representations are also given in this section. The way we deal with diphthongs and affricates is explained in respectively Section 4.4 and Section 4.5, while Section 4.6 describes how suprasegmentals and diacritics are processed. The comparison of sounds is explained in Section 4.7. In Section 4.8 the different acoustic representations are compared with each other and with respect to discrete representations. Segment distances which are obtained from the different systems are correlated with each other. Finally, in Section 4.9 we draw some conclusions.

## 4.1 Visible speech

In Potter et al.'s (1947) *Visible Speech*, spectrograms are shown for all common English sounds (see pp. 54–56: *The ABC'S of Visible Speech*). Examining the spectrograms the formant structure of vowels and sonorants, the high frequency noise of certain fricatives and the periods of voicing and voicelessness can be clearly identified (Trask, 1996, p. 328). Looking at the spectrograms we can already see which sounds are similar and which are not. We expect that visible (dis)similarity between spectrograms reflects perceptual (dis)similarity between sounds to some extent. In Figure 4.1 the spectrograms of some sounds are shown as pronounced by John Wells on the audio tape *The Sounds of the International Phonetic Alphabet* (Wells and House, 1995). The x-axis gives the time and the y-axis the frequency. For each frequency at each time the intensity is visualized by the darkness. The spectrograms are made with the computer program PRAAT.[1]

## 4.2 Samples

For finding spectrogram distances or formant track distances between all IPA sounds, for each sound we need samples from one or more speakers. We found these samples on the tape *The Sounds of the International Phonetic Alphabet* on which all IPA sounds are pronounced by John Wells and Jill House. On the tape the vowels are pronounced in isolation. The consonants are sometimes preceded, and always followed by an [a]. We cut the part preceding the [a], or the part between the [a]'s. We are aware of the fact that information on the F2 transition is lost. Rietveld and Van Heuven (1997) explain that the F2 transition in the transition zone from consonant to vowel gives information about the place of articulation. In our research we focus only on the sound itself. We also realize that the pronunciation of sounds depends on their context. For both vowels and consonants Stevens (1998) gives a discussion of some influences of context on speech sound production (pp. 557–581). Since we use samples of vowels pronounced in isolation, and samples of consonants selected from a limited context, our approach is a simplification of reality. However, Stevens (1998, p. 557) also observes that

> "by limiting the context, it was possible to specify rather precisely the articulatory aspects of the utterances and to develop models for estimating the acoustic patterns from the articulation".

The two speakers on the tape give us two sets of IPA samples. However, some sounds were missing or not properly pronounced. Therefore, the [ʔ], [ʀ] and [ʋ]

[1] The program PRAAT is a free public-domain program developed by Paul Boersma and David Weenink at the Institute of Phonetic Sciences of the University of Amsterdam and available via `http://www.fon.hum.uva.nl/praat/`.

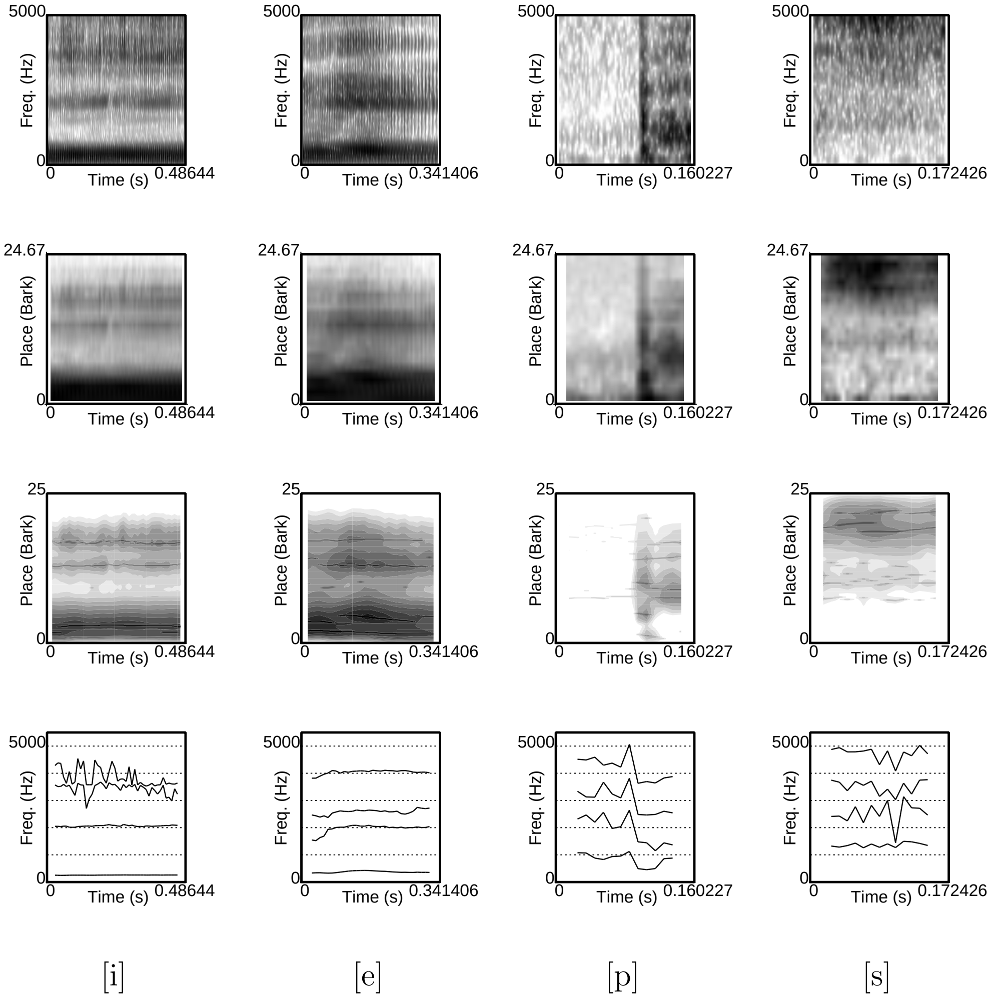
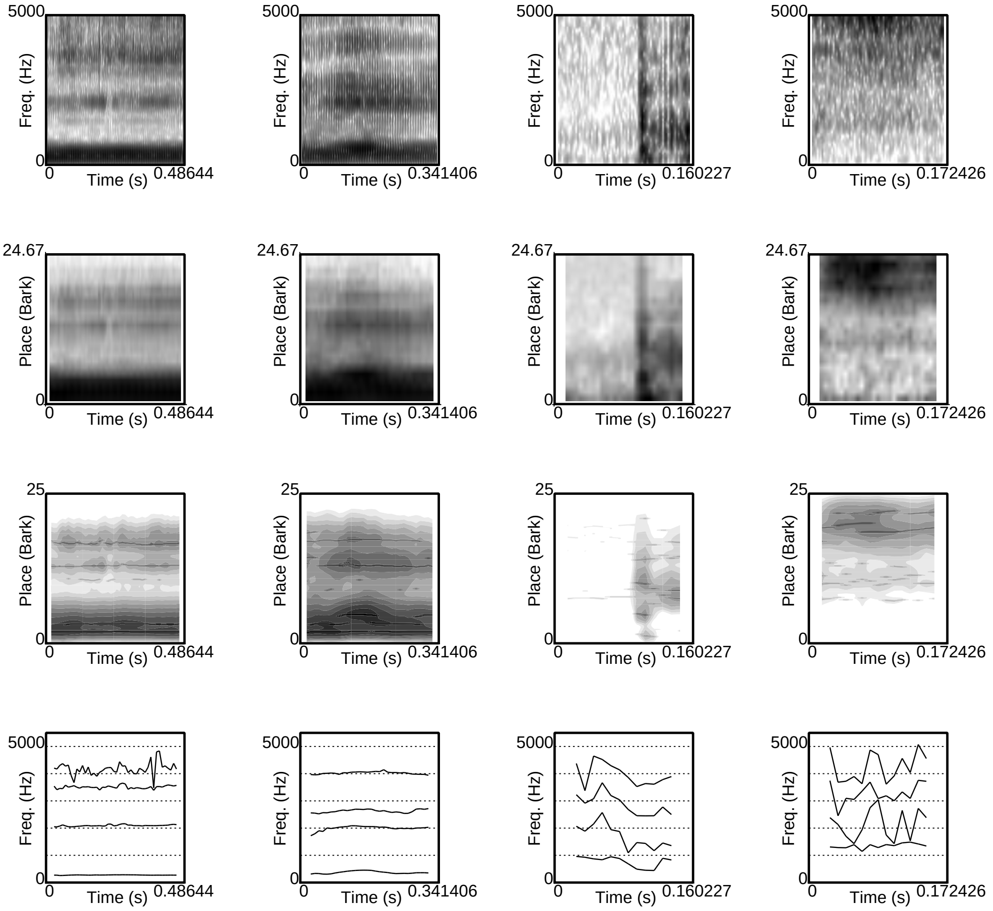


Figure 4.1: Different acoustic representations of four sounds as pronounced by John Wells. Starting with the first row we see respectively spectrograms, Bark-filters, cochleagrams and formant tracks obtained on the basis of the *original* samples.



[i] [e] [p] [s]

Figure 4.2: Different acoustic representations of four sounds as pronounced by John Wells. Starting with the first row we see respectively spectrograms, Barkfilters, cochleagrams and formant tracks obtained on the basis of the *monotonized* samples.

of Jill House were substituted by the corresponding sounds of John Wells. The original [ʋ] of Jill House is used as [w] for both John Wells and Jill House.

The burst in a plosive is always preceded by a period of silence (voiceless plosives) or a period of murmur (voiced plosives). When a voiceless plosive is not preceded by an [a], it is not clear how long the period of silence which really belongs to the sound lasts. Therefore, we have always cut each plosive in such a way that the time span from the beginning to the middle of the burst is equal to 90 ms. In a spectrogram the burst is recognized as a small dark vertical bar at the end of the period of silence or murmur (see e.g. the [p] in spectrograms in Figure 4.1). The middle of the burst was estimated by eye. Among the plosives which were preceded by an [a] or which are voiced (so that the real time of the start-up phase can be found), we found no sounds with a period of silence or murmur which was clearly shorter than 90 ms.

In voiceless plosives, the burst is followed by an [h]-like sound before the following vowel starts. When including this part in the samples, the consequence is that bursts will often not match when comparing two voiceless plosives. However, since aspiration is a characteristic property of voiceless sounds, we retained aspiration in the samples (see Figures 4.3 and 4.4 for both speakers). In the voiced sounds the burst is immediately followed by the vowel. In some cases it was not clear where the burst ended and the vowel started. For the voiced sounds it cannot be guaranteed that nothing from the following vowel is included, although any error here will be minimal. In general when comparing two voiced plosives, the bursts will match (see Figures 4.3 and 4.4 again). When comparing a voiceless plosive and a voiced plosive the bursts will not match.

To keep trills comparable to each other, we always cut three periods, even when the original samples contained more periods. When there were more periods the most regular looking sequence of three periods was cut.

To get a sample of ‘silence’ we cut a small silent part on the IPA tape. This assures that silence has about the same background noise that the other sounds have.

To make the samples as comparable as possible, all vowel and extracted consonant samples are monotonized on the mean pitch of the 28 concatenated vowels. The mean pitch of John Wells was 128 Hertz, the mean pitch of Jill House was 192 Hertz. In order to monotonize the samples the pitch contours were changed to flat lines. Figure 4.1 shows spectrograms of non-manipulated samples while Figure 4.2 shows spectrograms of the corresponding monotonized samples.

The volume was not normalized because volume contains too much segment specific information. For example, it is specific for the [v] that its volume is greater than that of the [f].

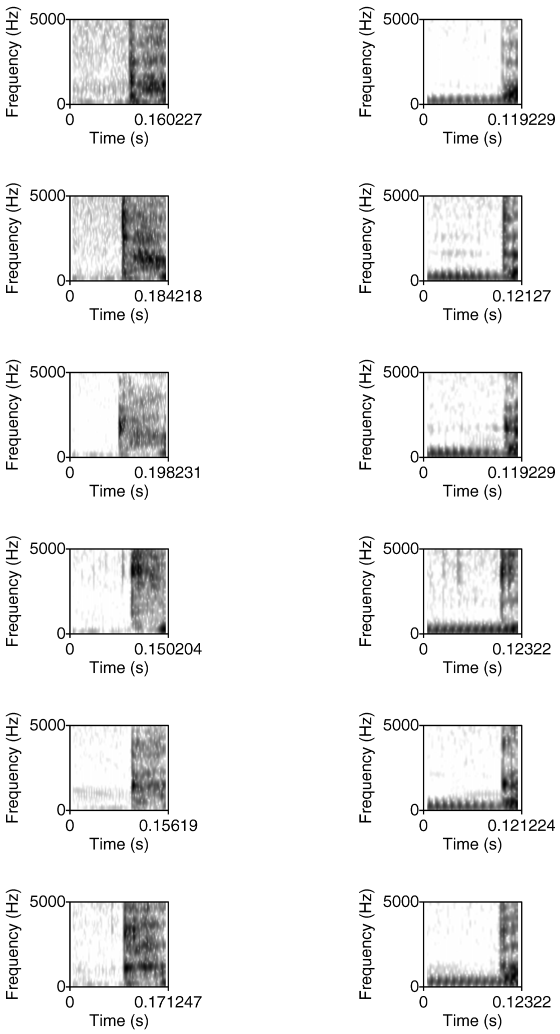


Figure 4.3: Spectrograms of voiceless (left) and voiced (right) plosives as pronounced by John Wells. Starting with the first row spectrograms are given for [p] and [b], [t] and [d], [ʈ] and [ɖ], [c] and [ɟ], [k] and [g], [q] and [ɢ]. When comparing voiceless plosives, the aspiration parts will match, and when comparing voiced plosives the bursts will match. When comparing a voiceless plosive with voiced plosive, the burst of the voiced plosive will partly match the aspiration part of the voiceless plosive.

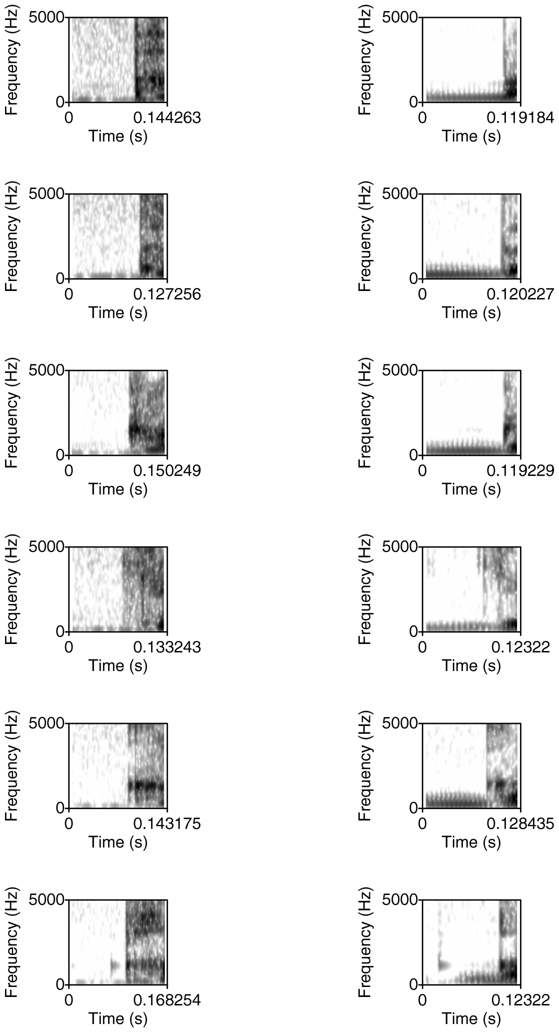


Figure 4.4: Spectrograms of voiceless (left) and voiced (right) plosives as pronounced by Jill House. Starting with the first row spectrograms are given for [p] and [b], [t] and [d], [ʈ] and [ɖ], [c] and [ɟ], [k] and [g], [q] and [ɢ]. When comparing voiceless plosives, the aspiration parts will match, and when comparing voiced plosives the bursts will match. When comparing a voiceless plosive with voiced plosive, the burst of the voiced plosive will partly match the aspiration part of the voiceless plosive.

## 4.3 Representation of segments

On the basis of the samples, manipulated spectrograms can be made or formant tracks can be found. We do not use the most common type of spectrogram with a Hertz-scale, but instead use more perceptually oriented models. In Section 4.3.1 we describe the *Barkfilter*. A Barkfilter has a frequency scale which is roughly linear below 1000 Hz, and roughly logarithmic above 1000 Hz. The logarithms of the intensities are mapped. In Section 4.3.2 we explain the *cochleagram* which is based on the Barkfilter, but may be more similar to human perception. The cochleagram uses the same frequency scale as the Barkfilter, but in the cochleagram, rather than the intensities themselves, the loudnesses as perceived by the ear are given. Besides two spectrogram like representations we also consider the *formant track* representation. Essential for perceiving vowels is that spectral peaks are recognized by the ear. The same applies for the sonorant consonants. These peaks are called *formants*. The formant track representation is discussed in Section 4.3.3. Our brief explanation of the three different representations is based on Rietveld and Van Heuven (1997).

### 4.3.1 Barkfilters

#### 4.3.1.1 Representation

In the most commonly used type of spectrogram the linear Hertz frequency scale is used. The difference between 100 Hz and 200 Hz is the same as the difference between 1000 Hz and 1100 Hz. However, our perception of frequency is non-linear. We hear the difference between 100 and 200 Hz as an octave interval, but the difference between 1000 to 2000 Hz is perceived as an octave as well. Our ear evaluates frequency differences not absolutely, but relatively, namely in a logarithmic manner. Therefore, in the Barkfilter, the Bark-scale is used which is roughly linear below 1000 Hz and roughly logarithmic above 1000 Hz (Zwicker and Feldtkeller, 1967). In the program PRAAT, for a given frequency in Hertz, the corresponding frequencies in Bark are found with a formula of Schroeder et al. (1979):

$$Bark = 7 \times ln\left(\frac{Hertz}{650} + \sqrt{\left(1 + \left(\frac{Hertz}{650}\right)^2\right)}\right) \tag{4.1}$$

Hertz frequencies are plotted against Bark frequencies in Figure 4.11. The graph shows two curves. The upper curve shows Bark values calculated with the formula of Schroeder et al. (1979) (applied when using Barkfilters and cochleagrams, see also Section 4.3.2), the lower curve shows Bark values calculated with the formula of Traunmüller (1990) (applied when using formant tracks, see Section 4.3.3). In the plot the Hertz-scale runs from 20 to 20,000 Hertz, the frequency range which can be perceived by a human being. This corresponds with a frequency range of 0.22 to 28.84 Bark for the Schroeder et al. curve.

In the commonly used type of spectrogram the power spectral density is represented per frequency per time. The power spectral density is the power per unit of frequency as a function of the frequency. In the Barkfilter the power spectral density is expressed in decibels (dB's). "The decibel scale is a way of expressing sound amplitude that is better correlated with perceived loudness" (Johnson, 1997, p. 53). The decibel-scale is a logarithmic scale. Multiplying the sound pressure ten times corresponds with an increase of 20 dB. On a decibel scale intensities are expressed relative to the auditory threshold. The auditory threshold of 0.00002 Pa corresponds with 0 dB (Rietveld and Van Heuven, 1997, p. 199).

A Barkfilter is created from a sound by band filtering in the frequency domain with a bank of filters. In PRAAT the lowest band has a central frequency of 1 Bark per default, and each band has a width of 1 Bark. There are 24 bands, corresponding with the first 24 critical bands of hearing as found along the basilar membrane (Zwicker and Fastl, 1990). A critical band is an area within which two tones influence each other's perceptibility (Rietveld and Van Heuven, 1997, pp. 204–205). Due to the Bark-scale the higher bands summarize a wider frequency range than the lower bands.

In the Figures 4.1 and 4.2 Barkfilters are shown, obtained on respectively non-manipulated and monotonized samples. In this type of spectrogram the Bark-scale is used as frequency scale, while intensities are given in *Decibels*. The frequencies range from 0 to 24.67 Bark. They are divided in 24 equal intervals, where for each interval the mean intensity is given. The sound signal is probed each 0.005 seconds with an analysis window of 0.015 seconds. Here we used the standard settings in the program PRAAT. Other settings may give different results, but since it is not a priori obvious which results are optimal, we restricted ourselves to the default settings.

#### 4.3.1.2 Classification

In Section 4.7 we describe how the distance between two spectrograms is measured in our research. The measure described in that section enables us to calculate the distances between all sounds. For the RND we have 18 vowels and 27 consonants. Also 'silence' is added. This gives a total of 46 sounds. In the NOS data the modern IPA system is used. In the IPA system 28 vowels and 58 pulmonic consonants are given. We added the [w] so we get 59 consonants. 'Silence' is added as well. So we get in total 88 sounds. Because we have two sets of samples (namely of John Wells and Jill House), two distance matrices are obtained for the RND and the NOS. Next the matrices of the two speakers are averaged, resulting in distances which are more general and less speaker dependent. We are aware of the fact that the number of speakers – two – is minimal. Useful future research would be to repeat this experiment on the basis of many more speakers.

Besides averaged matrices for all sounds, we also made matrices containing the averaged distances between vowels only and consonants only. Since it is difficult to appreciate all distances individually we have chosen to visualize the results to give an impression of the results.[2] On the basis of the IPA versions of these matrices we performed multidimensional scaling. Multidimensional scaling was not performed on the RND matrices because the RND sounds are just a subset of the IPA sounds. The multidimensional scaling technique is described in more detail in Section 6.2 and shows us the relations between the sounds in two-dimensional space. This allows us to compare the ordering of the sounds with the way in which they are ordered in the IPA system. The multidimensional scaling plots also show differences between the classifications of the different representations.

Note that for dialect comparison the real sound distances are used, not the multidimensional scaling distances. The multidimensional scaling plots are used here only to visualize the distances and suggest that the spectrogram or formant track distances yield a reasonable measure of pronunciation.

**Vowels** When using multidimensional scaling on the basis of the vowel distances, one dimension explains already 85% of the variance, two dimensions 98% and three dimensions 98% as well. In Figure 4.5 a two-dimensional multidimensional scaling plot is shown. The first dimension (the vertical dimension in the plot) represents the height, and the second dimension (horizontal) the advancement. The positions of the [i], [u], [ɒ] and [a] resemble those in the IPA quadrilateral clearly. We see a clear divison between high and low vowels. Note that the [ə] belongs to the higher vowels of the lower group, while in the IPA quadrilateral this sound is located exactly in the center. This may be explained by the fact that in our calculations information additional to the F1 and F2 were used. When scaling to three dimensions, it appears that the third dimension does not distinguish between spread and rounded vowels as in the IPA quadrilateral, but distinguishes between central vowels on the one hand, and front and back vowels on the other hand.

**Consonants** When using multidimensional scaling on the basis of the consonant distances, one dimension explains 59% of the variance, two dimensions 94% and three dimensions 99%. In Figure 4.6 a two-dimensional multidimensional scaling plot is shown. The first dimension (the vertical dimension in the plot) makes a distinction between voiceless (upper) and voiced (lower) sounds. The second dimension (horizontal) distinguishes between continuous (left) and non-continuous consonants (right). Comparing these results with the IPA table, the

[2] For 28 vowels we get $\binom{28}{2} = 378$ distances, for 59 consonants we get $\binom{59}{2} = 1711$ distances, and for vowels plus consonants plus 'silence' we get $\binom{88}{2} = 3828$ distances.

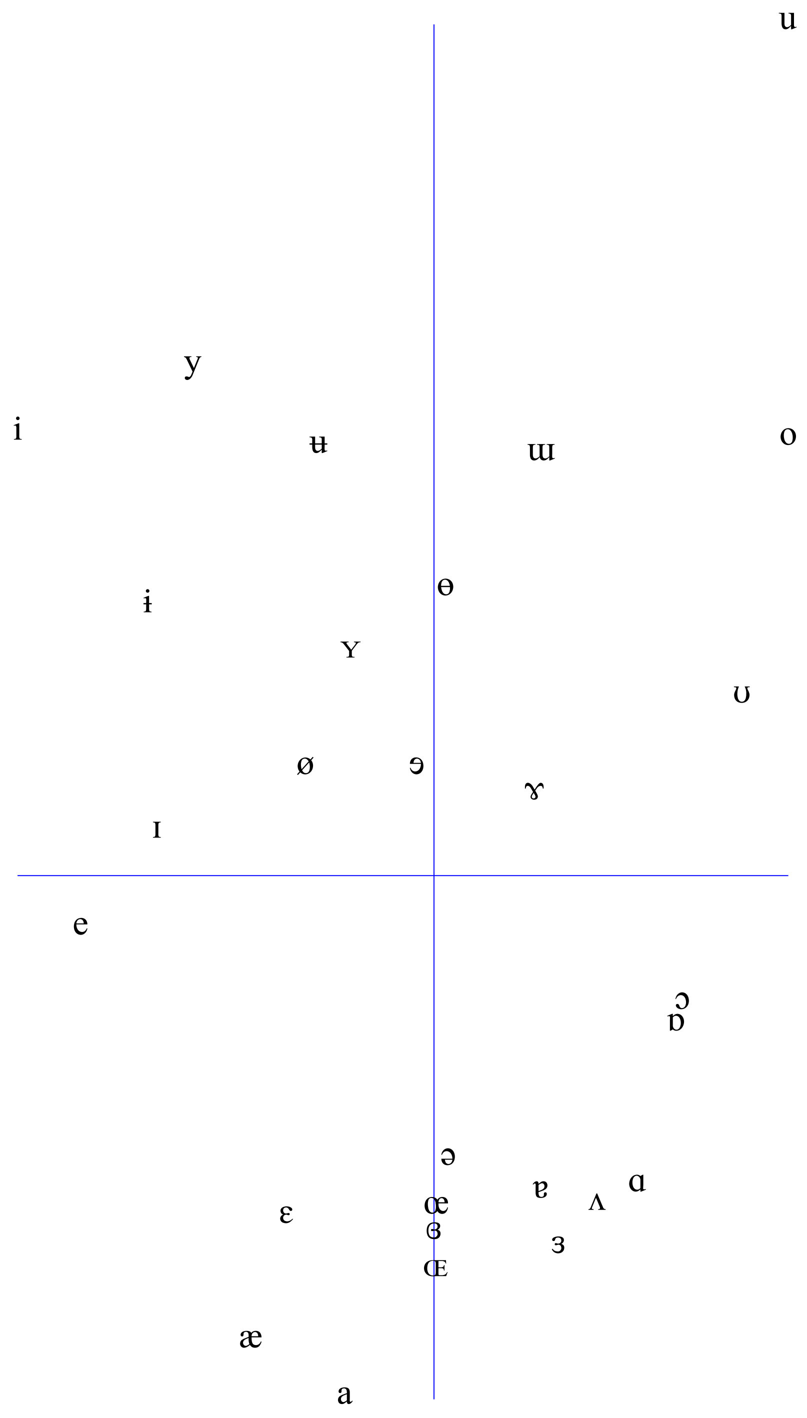


Figure 4.5: Two-dimensional multidimensional scaling plot obtained from the Barkfilter distances between all pairs formed by the 28 vowels. Two dimensions explain 98% of the variance. The first dimension (y-axis) corresponds with height and the second dimension (x-axis) with advancement. We might have expected the schwa [ə] to be placed more highly. A third dimension would distinguish between central vowels on the one hand, and front and back vowels on the other hand.

place of articulation hardly plays any role, in contrast with the manner of articulation which is important. Striking is the position of the [j] between the approximants and the voiced plosives. With respect to the voiced plosives the [j] is most similar to the [g]. We illustrate this relation by two examples. In German, *Morgen* 'morning' is usually pronounced as [mɔɐ̯gən], but in Berlin the same word is also pronounced as [mɔɐ̯jən]. The German word *gemacht* 'made' is mostly pronounced as [gəmaxt]. In Berlin, however, this word is also pronounced as [jəmaxt]. Apart from the [g], among the voiced plosives the [d] is most similar to the [j]. The relation between the [j] and the [d] can easily be illustrated as well. E.g., the Dutch words *goede* 'good' [xudə] and *rode* 'red' [roˑdə] are also pronounced as [xujə] and [roˑjə]. Also striking is the position of the [f] rather close to the voiceless plosives. This relation can be found e.g. in the word *father* [faðəɹ], where the [f] arose from Indo-European [p] (cf. Latin [patəɾ]). Unfortunately, a similar close relation between the [x] and the [k] cannot be found here. Looking at the approximants it is striking that the retroflex variants do not cluster with the other approximants, but are located in the neighborhood of the trills. When scaling to three dimensions the third dimension distinguishes between voiceless plosives, retroflex, velar, uvular, pharyngeal and glottal voiceless fricatives and r-like consonants on the one hand, and (lateral) alveolar, postalveolar and palatal fricatives, the palatal voiced plosive and palatal (lateral) approximants on the other hand. We cannot explain what this distinction is based on.

**All sounds** When using multidimensional scaling on the basis of all sound distances, one dimension explains 76% of the variance, two dimensions 96% and three dimensions 98%. In Figure 4.7 a two-dimensional multidimensional scaling plot is shown. The first dimension (the vertical dimension in the plot) corresponds with intensity. The [a] is loudest and 'silence' is most silent (of course). The second dimension (horizontal) represents clearness. The [ʃ] is the clearest and the [u] is the darkest sound. However, one might expect some consonants (e.g. the [ɸ]) to be located in the darker area. In this context, a sound is clear when is has many harmonic tones, and it is dark when harmonic tones are lacking. The voiceless plosives and the voiced plosives can clearly be identified as different groups. However, the other sounds form a continuum. Drawing a line from [i] to [a], from [a] to [ɒ], from [ɒ] to [u], and from [u] to [i] we recover the IPA vowel quadrilateral. Here the nasals appear as high vowels. Also, note the position of the [r], [ʀ] and [ɾ] in the IPA vowel quadrilateral. A close relation between these liquids and (central) vowels can be illustrated by the fact that e.g. the Dutch word *vier* 'four' is sometimes pronounced as [fiːr] and sometimes as [fiːə]. Here we see that the [r] can correspond with the [ə]. Less easy to explain is the appearance of the voiced fricative [ɣ] on the border of the quadrilateral, close to the [i]. We note that the [u] and [w] are very close, but the [i] and the [j] are not close. Maybe this is due to the fact that the [j] has a lower intensity than the

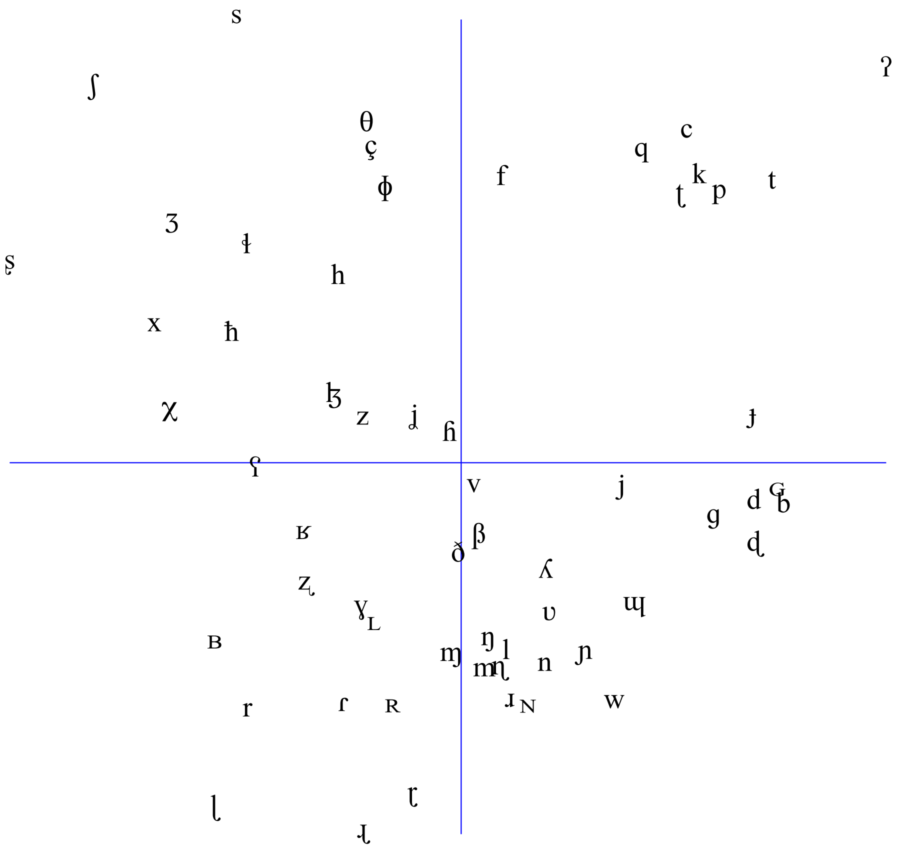


Figure 4.6: Two-dimensional multidimensional scaling plot obtained from the Barkfilter distances between all pairs formed by the 59 consonants. Two dimensions explain 94% of the variance. The first dimension (y-axis) distinguishes between voiceless and voiced consonants, the second dimension (x-axis) between non-continuous and continuous consonants. A third dimension would less easily be interpreted.

[i]. Further the position of 'silence' is very near the glottal stop. When scaling to three dimensions the third dimension distinguishes between low vowels, voiceless plosives, retroflex, velar, uvular, pharyngeal and glottal voiceless fricatives and r-like consonants on the one hand, and high vowels, (lateral) alveolar, postalveolar and palatal fricatives, the palatal voiced plosive and (lateral) approximants on the other hand.

Since the vowel classification is like the IPA quadrilateral, and the consonant classification reflects the different manners of articulation we conclude that the Barkfilter representation is useful for finding segment distances.

## 4.3.2 Cochleagrams

### 4.3.2.1 Representation

The cochleagram represents the behavior of the basilar membrane in the cochlea. The cochlea is the inner part of the ear. Just as in the Barkfilter, in a cochleagram the Bark-frequency scale is used. In the computer program PRAAT Bark values in a Barkfilter are found using the formula of Schroeder et al. (1979) (see Section 4.3.1). In PRAAT the same formula is used for the frequency scale of cochleagrams.

In a Barkfilter for each time and for each frequency the *intensity* is given. In a cochleagram for each time for each frequency the *loudness* is given. When two sounds have the same intensity but different frequencies, they will probably be perceived as differing in loudness. Human aural sensitivity varies with frequency. Loudness is the perceived intensity. Loudnesses are expressed in reference intensities. In a cochleagram the reference intensities are the intensities of a frequency of 1000 Hz. This is the basis for the measurement of loudness in *phon*. If a given sound is perceived to be as loud as a 60 dB sound at 1000 Hz, then it is said to have a loudness of 60 phon. The relation between the reference loudness and the loudness of another given intensity at a specific frequency is determined experimentally.

Since only one pure tone leads to activation of hair cells over a large surface on the basilar membrane, the ear is not able to perceive other neighboring frequencies. One tone is masked by the other. There are two types of masking: lateral masking and forward masking. Lateral masking occurs when at the same time different but neighboring frequencies are recorded. One tone may make other nearby tones (nearly) inaudible. In general a low tone will mask a high tone rather than the opposite. Forward masking appears when tones occur after each other. E.g., after hearing a strong sound our ears may be stunned for a short time. The more successive sounds resemble each other, the stronger the masking will be. In a cochleagram both the lateral and the forward masking is modeled.

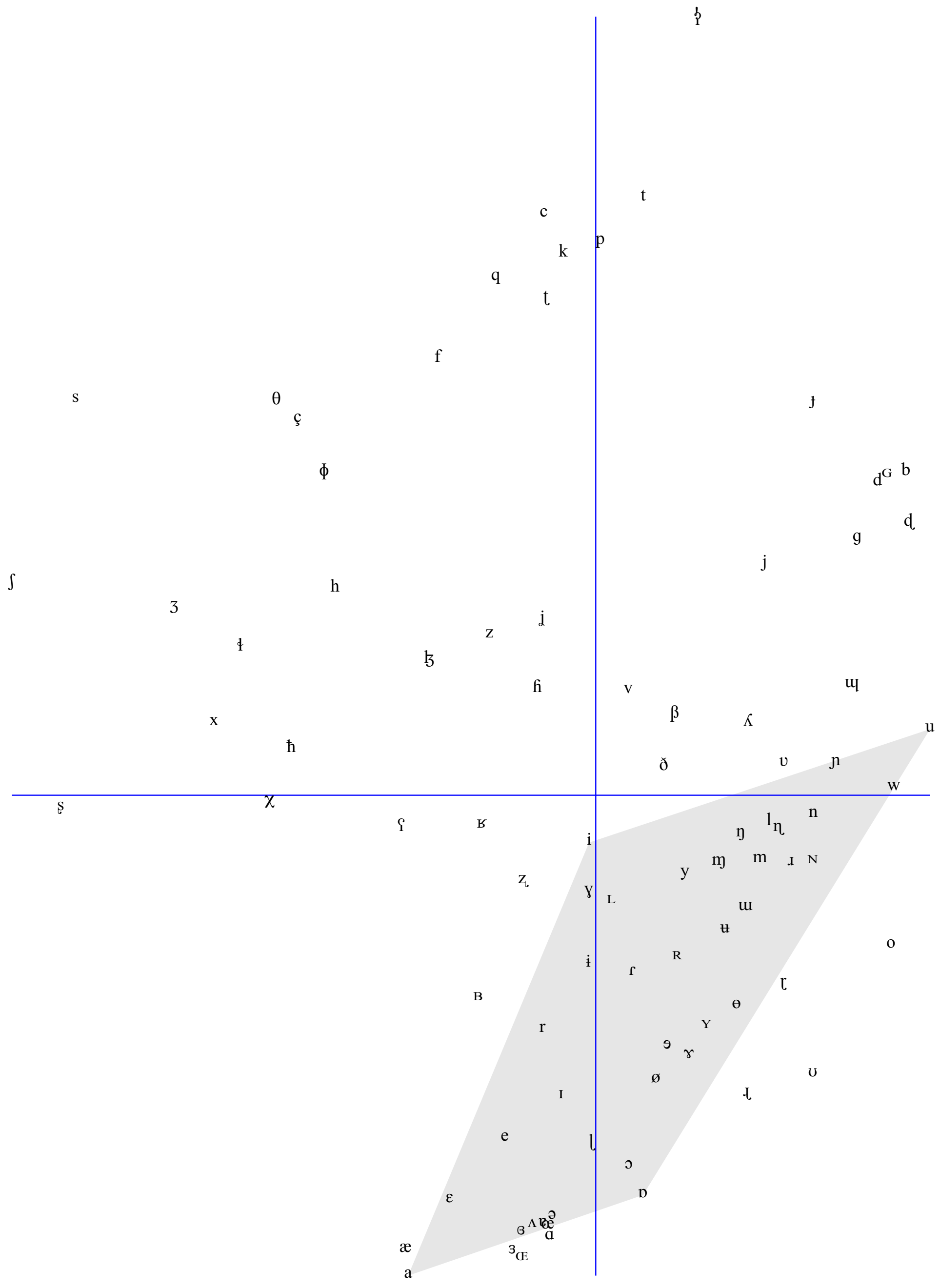


Figure 4.7: Two-dimensional multidimensional scaling plot on the basis of the Barkfilter distances between all pairs formed by the 28 vowels, the 59 consonants and 'silence'. In the plot, the ! is used for 'silence'. Two dimensions explain 96% of the variance. The first dimension (y-axis) represents intensity (lower sounds are louder) and the second dimension (x-axis) clearness (sounds on the right are darker). The shaded area represents a quadrilateral formed on the basis of the vertices of the IPA quadrilateral.

In our research we did not consider forward-masking. In our case all sounds were pronounced in isolation (vowels) or cut from their context (consonants). The effect of forward-masking would mainly be found at the begin of a segment and models the phenomenon that a sound has a gradual onset. For long sounds the effect is relatively greater than for short sounds. Thus the relative influence depends on the absolute length of a sample. However, not all samples have reliable lengths. On the IPA tape the vowels are pronounced in isolation. Therefore, our vowel durations will be longer than the durations of vowels which are pronounced in words. Consonant durations reflect the property of the segment to some extent. However, for trills no more than three periods were cut even if there were more periods (see Section 4.2). Besides cutting both vowels and consonants always involves inaccuracies. Therefore, we did not apply forward-masking.

In the Figures 4.1 and 4.2 cochleagrams are shown, obtained on respectively non-manipulated and monotonized samples. In this type of spectrogram loudnesses are given instead of intensities, expressed in *phon*. Further lateral and forward frequency masking is modeled. The darker lines in the pictures represent formant tracks (see Section 4.3.3.1). The frequencies range from 0 to 25.6 Bark. They are divided in 256 equal intervals, where for each interval the mean loudness is given. The sound signal is probed each 0.01 seconds with an analysis window of 0.03 seconds. The forward-masking time is set at 0.00 seconds. This means that the effect of forward-masking is not regarded in our results. Except for the forward-masking time, here we used the standard settings in the program PRAAT. Other settings may give different results, but just as for the Barkfilter it is not clear which results will be optimal beforehand. Therefore, we restricted ourselves to the default settings.

#### 4.3.2.2 Classification

Just as for the Barkfilter representation, the distances between the IPA sounds are calculated (see Section 4.7). Multidimensional scaling is performed on the basis of vowel distances, the consonant distances and the distances between both vowels and consonants including ‘silence’. Since we followed the same procedure as for the Barkfilter representation, the reader is referred to Section 4.3.1 for more details.

**Vowels** When using multidimensional scaling on the basis of the vowel distances, one dimension already explains 86% of the variance, two and three dimensions 98%. In Figure 4.8 a two-dimensional multidimensional scaling plot is shown. The plot is very similar to the Barkfilter plot (see Figure 4.5), so the conclusion is that it does not matter whether the Barkfilter or the cochleagram representation is used when finding distances between vowels. For the cochleagram vowel plot the same remarks apply as for the Barkfilter vowel plot (see Section 4.3.1).

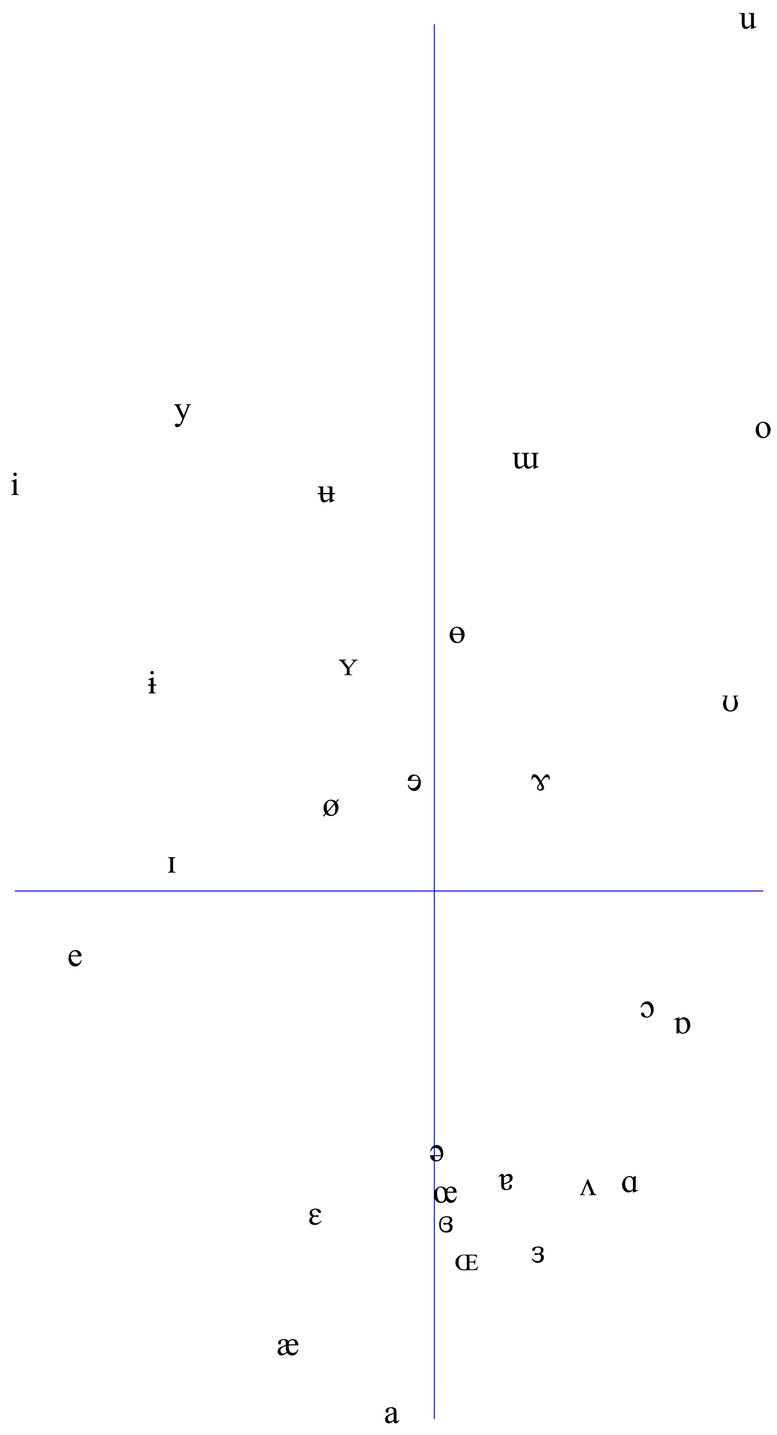


Figure 4.8: Two-dimensional multidimensional scaling plot obtained from the cochleagram distances between all pairs formed by the 28 vowels. Two dimensions explain 98% of the variance. The first dimension (y-axis) corresponds with height and the second dimension (x-axis) with advancement. We might have expected the schwa [ə] to be placed more highly. A third dimension would distinguish between central vowels on the one hand, and front and back vowels on the other hand.

**Consonants** When using multidimensional scaling on the basis of the consonant distances, one dimension explains 81% of the variance, two dimensions 96% and three dimensions 98%. In Figure 4.9 a two-dimensional multidimensional scaling plot is shown. Just as the plot based on the Barkfilter distances (see Figure 4.6) the first dimension (the vertical dimension in the plot) makes a distinction between voiceless (upper) and voiced sounds (lower). The second dimension (horizontal) distinguishes between continuous (left) and non-continuous consonants (right). The plot is very similar to the Barkfilter consonant plot. The only important difference is that the division between voiceless and voiced sounds in the cochleagram plot is sharper than in the Barkfilter plot. Probably this is explained by that fact that the cochlear model uses loudness instead of intensity. Perceptually, the distinction between voiceless and voiced sounds is greater than pure intensities indicate. For further comments see the explanation of the Barkfilter consonant plot (see Section 4.3.1).

**All sounds** When using multidimensional scaling on the basis of all sound distances, one dimension explains 88% of the variance, two dimensions 98% and three dimensions 99%. In Figure 4.10 a two-dimensional multidimensional scaling plot is shown. The first dimension (the vertical dimension in the plot) corresponds with intensity. The [ə] is loudest and 'silence' is most silent (of course). The second dimension (horizontal) represents clearness. The [ʂ] is the clearest and the [b] is the darkest sound. However, one might expect some consonants (e.g. the [ɸ]) to be located in the darker area, just as in the Barkfilter plot. In the plot the distinction between voiceless and voiced sounds is sharper than in the Barkfilter plot, just as we saw for the consonants. Drawing a line from [i] to [a], from [a] to [ɒ], from [ɒ] to [u], and from [u] to [i] we recover the IPA vowel quadrilateral. Just as in the Barkfilter plot the nasals appear as high vowels. In contrast to the Barkfilter most r-like sounds are outside the vowel quadrilateral now. Only the retroflex flap is still in the quadrilateral. Both retroflex approximants are moved to the center area of the quadrilateral. The [u] and the [w] are still closer than the [i] and the [j]. Additionally the position of 'silence' is very near the glottal stop. For the explanation of the third dimension see the Barkfilter representation (Section 4.3.1).

Because the vowel classification is like the IPA quadrilateral, and the consonant classification reflects the different manners of articulation we conclude that the cochleagram representation is useful for finding segment distances. Cochleagrams differ from the Barkfilter representation in virtue of the sharper distinction between voiceless and voiced sounds, and between vowels and r-like sounds.

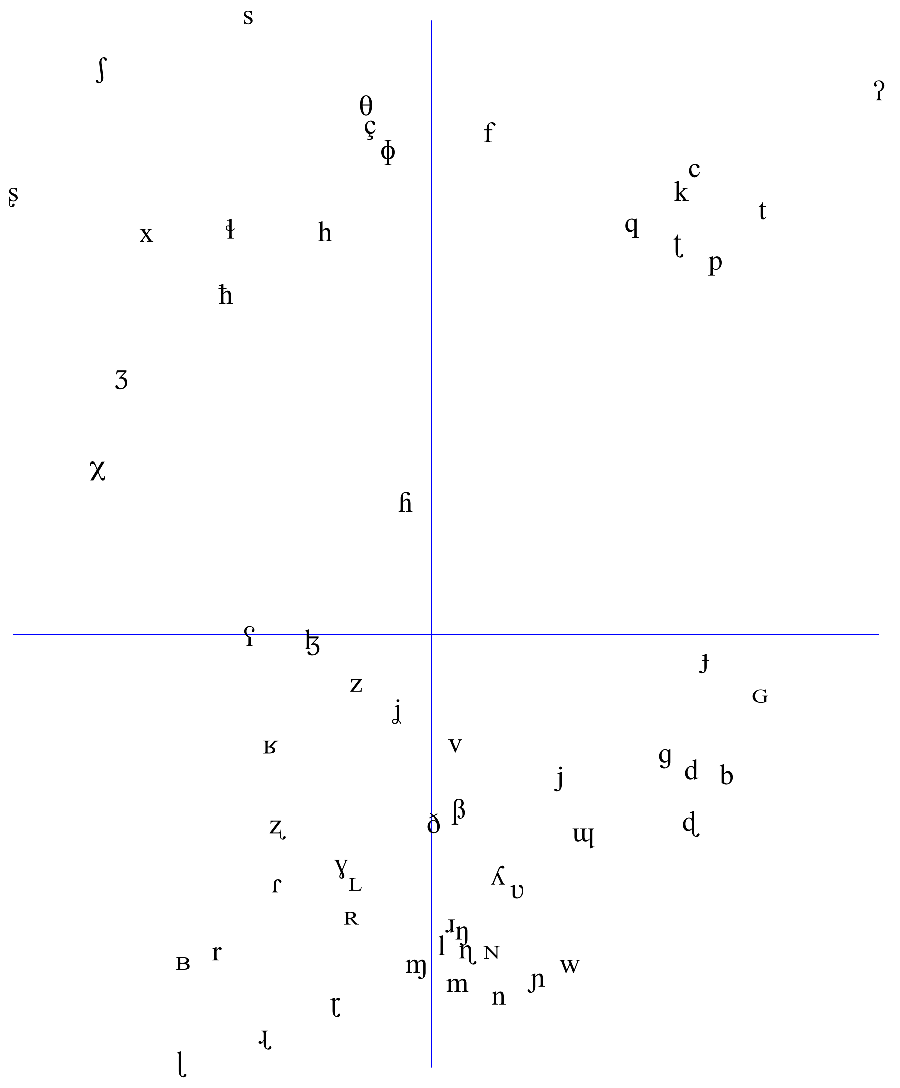


Figure 4.9: Two-dimensional multidimensional scaling plot obtained from the cochleagram distances between all pairs formed by the 59 consonants. Two dimensions explain 96% of the variance. The first dimension (y-axis) distinguishes between voiceless and voiced consonants, the second dimension (x-axis) between non-continuous and continuous consonants. A third dimension would be less easily interpreted.

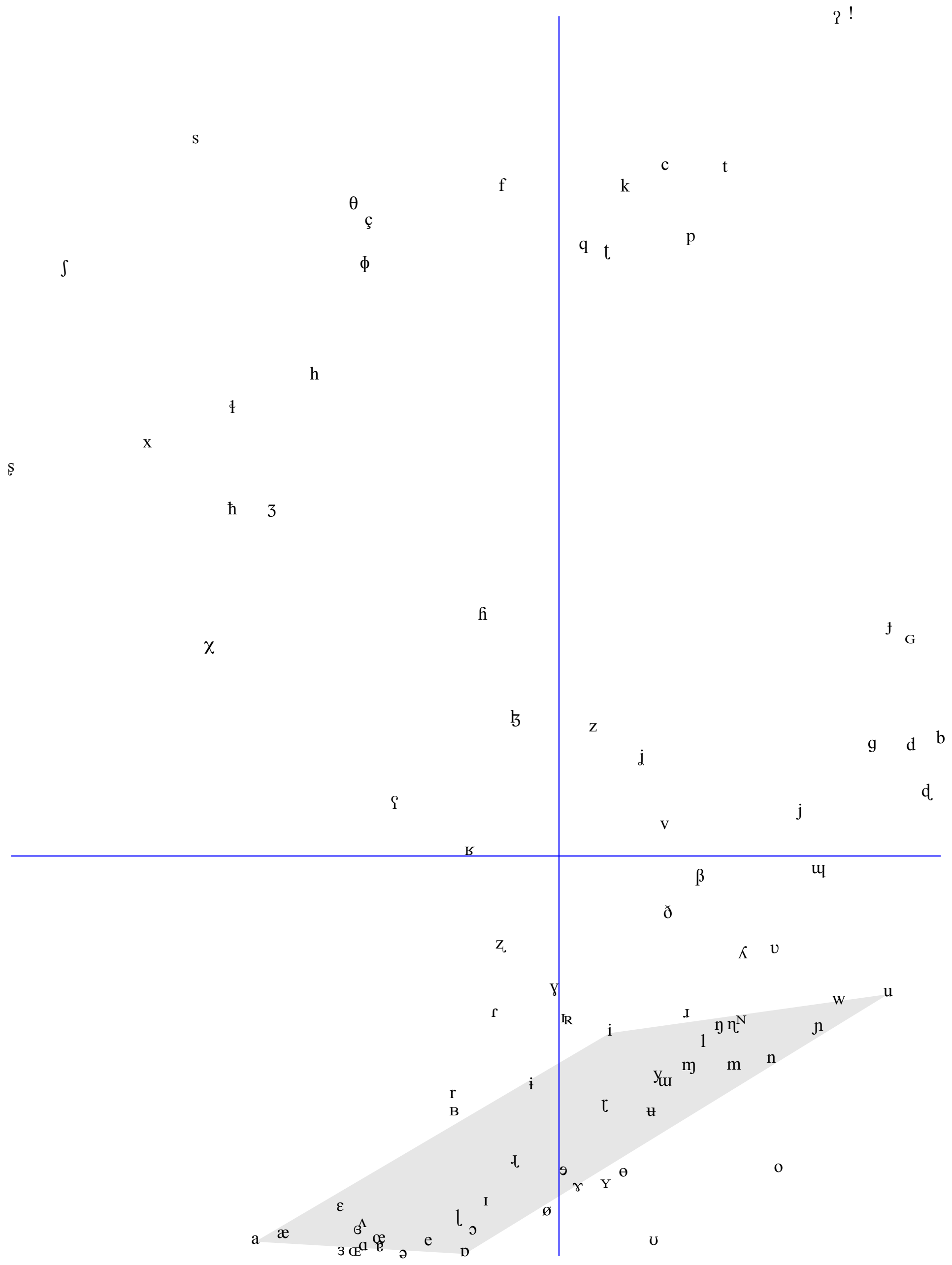


Figure 4.10: Two-dimensional multidimensional scaling plot on the basis of the cochleagram distances between all pairs formed by the 28 vowels, the 59 consonants and 'silence'. In the plot, the ! is used for 'silence'. Two dimensions explain 98% of the variance. The first dimension (y-axis) represents intensity (lower sounds are louder) and the second dimension (x-axis) clearness (sounds on the right are darker). The shaded area represents a quadrilateral formed on the basis of the vertices of the IPA quadrilateral.

### 4.3.3 Formant tracks

#### 4.3.3.1 Representation

Another way to analyse the acoustic signal is to investigate formants. When using a spectrogram with a large analysis window (about 20 ms) the frequency resolution will be high. Individual harmonics will show up as horizontal lines through the spectrogram (see the spectrograms, Barkfilters and cochleagrams in Figures 4.1 and 4.2). The lowest line represents the fundamental frequency or pitch (F0). However, when using a small analysis window (about 3 ms) the frequency resolution will be lower. Individual harmonics get blended together. Instead of lines, bands will show up through the spectrogram. The center frequency at one time in a band is called a formant, the range of center frequencies in the course of time forms a formant track. A formant in the lowest band is called F1, a formant in the next band F2, etc. Formants represent a frequency region that is enhanced by the resonances of the vocal tract.[3]

In the Figures 4.1 and 4.2 formant tracks are shown, obtained on respectively non-manipulated and monotonized samples. When finding formants in the computer program PRAAT, the time step was set to 0.01 seconds with an analysis window of 0.025 seconds. The ceiling of the formant search range should be set to 5000 Hz for males, and to 5500 Hz for females. So for the samples of John Wells the ceiling was set to 5000 Hz, and for Jill House to 5500 Hz. Pre-emphasis starts at 50 Hz. In the manual which can be found in the PRAAT program pre-emphasis is explained as follows:

> “This means that frequencies below 50 Hz are not enhanced, frequencies around 100 Hz are amplified by 6 dB, frequencies around 200 Hz are amplified by 12 dB, and so forth. The point of this is that vowel spectra tend to fall by 6 dB per octave; the pre-emphasis creates a flatter spectrum, which is better for formant analysis because we want our formants to match the local peaks, not the global spectral slope.”

In PRAAT several algorithms can be chosen for finding the Linear Predictive Coding (LPC) coefficients. We chose the algorithm of Burg. This algorithm may initially find formants at very low or high frequencies. However we used in PRAAT the version which removes formants below 50 Hz and formants above 5000 Hz (males) or 5500 Hz (females) minus 50 Hz. In this way the algorithm will identify the traditional F1 and F2. The algorithm of Burg is much more reliable than the Split Levinson algorithm which always finds the requested number of formants in every frame, even if they do not exist. Since we found at least two formants for every frame in every sample when using the more reliable Burg

[3] See also http://www.bsos.umd.edu/hesp/newman/Newman_classes/Newman604/604.html.

algorithm we do not use the Split Levinson algorithm. More about the algorithms can be found in the manual in PRAAT program.

When using formant tracks we had to decide how many formant tracks should be taken into account. It is a well-known fact that in the IPA vowel quadrilateral the height corresponds with the F1 (the lower the F1, the closer the vowel) and that the advancement corresponds with the F2 (the higher the F2, the further sounds are fronted, see Rietveld and Van Heuven (1997, p. 133)). The F2 of rounded vowels is a little lower than the F2 of unrounded vowels. The meaning of higher formants is less clear. At the risk of ignoring information important to dialect recognition we therefore decided to compare sounds only on the basis of the F1 and the F2. Before comparing formant frequencies in the comparison of words the frequencies in Hertz are converted to Bark, which is, as mentioned above, a more faithful scale perceptually. For this purpose we used the formula of Traunmüller (1990) as suggested in standard works about phonetics (Rietveld and Van Heuven, 1997):

$$Bark = \frac{26.81 \times Hertz}{1960 + Hertz} - 0.53 \tag{4.2}$$

The relation between Traunmüller's formula and Schroeder et al.'s (see the formula in (4.1) in Section 4.3.1.1) is shown in Figure 4.11, in which the Hertz frequencies are plotted against the Bark frequencies. The graph shows two curves, the upper one based on the formula of Schroeder et al. (1979), the lower one found by using the formula of Traunmüller (1990). As mentioned in the Sections 4.3.1 and 4.3.2 the Schroeder et al. formula is used for Barkfilters and cochleagrams. In the plot the Hertz-scale runs from 20 to 20000 Hertz, the frequency range which can be perceived by a human being. This corresponds with a frequency range of -0.26 to 23.89 for the Traunmüller curve.

#### 4.3.3.2 Classification

Just as for the Barkfilter representation and the cochleagram representation the distances between the IPA sounds are calculated (see Section 4.7). Multidimensional scaling is performed on the basis of vowel distances, the consonant distances and the distances between both vowels and consonants including 'silence'. Since we followed the same procedure as for the Barkfilter representation, the reader is referred to Section 4.3.1 for more details.

**Vowels** When using multidimensional scaling on the basis of the vowel distances, one dimension already explains 67% of the variance, two dimensions 99% and three dimensions 100%. In Figure 4.12 a two-dimensional multidimensional scaling plot is shown. The first dimension (the horizontal dimension in the plot) represents the advancement, the second dimension (vertical) height. The plot is rather similar to the Barkfilter plot (see Figure 4.5) and the cochleagram plot

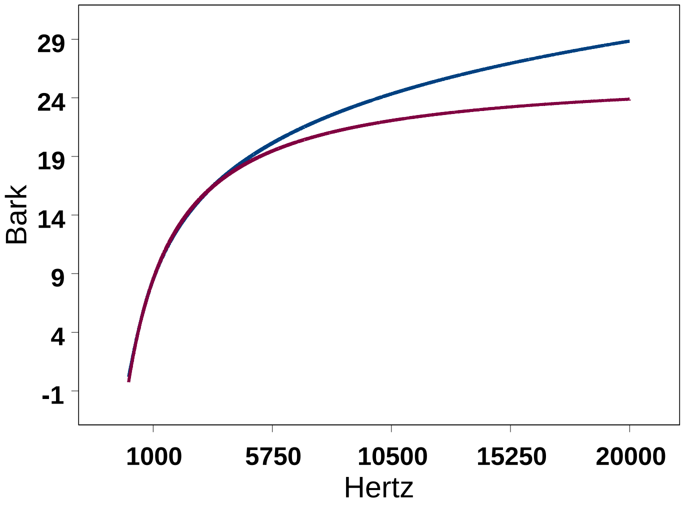


Figure 4.11: Frequencies in Hertz versus frequencies in Bark. The Hertz scale runs from 20 Hz to 20.000 Hz. The upper line shows Bark values calculated with the formula of Schroeder et al. (1979) (applied when using Barkfilters or cochleagrams), the lower line shows Bark values calculated with the formula of Traunmüller (1990) (applied when using formant tracks). Below 1000 Hz both curves are roughly linear, above 1000 Hz they are roughly logarithmic.

(see Figure 4.8). However, when drawing a line from [i] to [a], from [a] to [ɒ], from [ɒ] to [u], and from [u] to [i] we get a triangle rather than a quadrilateral, where the [ɒ] is on the line between the [a] and the [u]. This agrees with results found in Rietveld and Van Heuven (1997, p. 133). They show a triangle based on mean formant values of male speakers derived from Pols (1977). In our plot front vowels have a high F2 and back vowels a low F2. The second dimension represents the height, corresponding with F1. High vowels have a low F1 and low vowels have a high F1. In the plot the [ə] is nearly in the center, while in the Barkfilter plot and the cochleagram plot the sound belonged to the higher vowels of the lower group. The division between high vowels and low vowels is not as sharp as in the Barkfilter plot and the cochleagram plot. When scaling to three dimensions, the third dimension makes a distinction between the [u], [i], [ɪ], [e], [ɛ] and [æ] on the one hand, and the other vowels on the other hand. This may be interpreted as a distinction of high and front vowels versus other vowels, although this is not consistently true.

**Consonants** When using multidimensional scaling on the basis of the consonant distances, one dimension explains 79% of the variance, two dimensions 92% and three dimensions 96%. In Figure 4.13 a two-dimensional multidimensional scaling plot is shown which was obtained on the basis of the consonant distances. The first dimension (the vertical dimension in the plot) distinguishes between voiced (upper) and voiceless sounds (lower) and the second dimension (horizontal) represents the place of articulation very vaguely, albeit in a different way than that used in the IPA pulmonic consonant table. The palatals appear in Figure 4.13 as front consonants (left), and the velar, uvular, pharyngeal and glottal consonants appear as back consonants (right). The [w] which is specified as a voiced labial-velar approximant in the IPA system is found here as a back consonant as well as the [ʋ]. Other consonants are more central. Drawing a line from the [j] to the [w], from the [w] to the [h] and from the [h] to the [j] a similar triangle is found as in the vowel plot. Voiced sounds have a low F1 and voiceless sounds have a high F1. Front consonants have a high F2 and back consonants a low F2. However, this is a simplified sketch, many exceptions can be found if one examines the plot more precisely. The role of manner of articulation is found in the plot. The nasals (upper right), voiced obstruents, (upper central), liquids (central, below the nasals and the voiced obstruents), and voiceless obstruents (low) can be identified. The plot is different from the Barkfilter plot (see Figure 4.6) and the cochleagram plot (see Figure 4.9). Most striking is the fact that there is no sharp separation between plosives and fricatives. We examined the third dimension as well but found no obvious interpretation for this.

**All sounds** When using multidimensional scaling on the basis of all sound distances, one dimension explains 72% of the variance, two dimensions 96% and

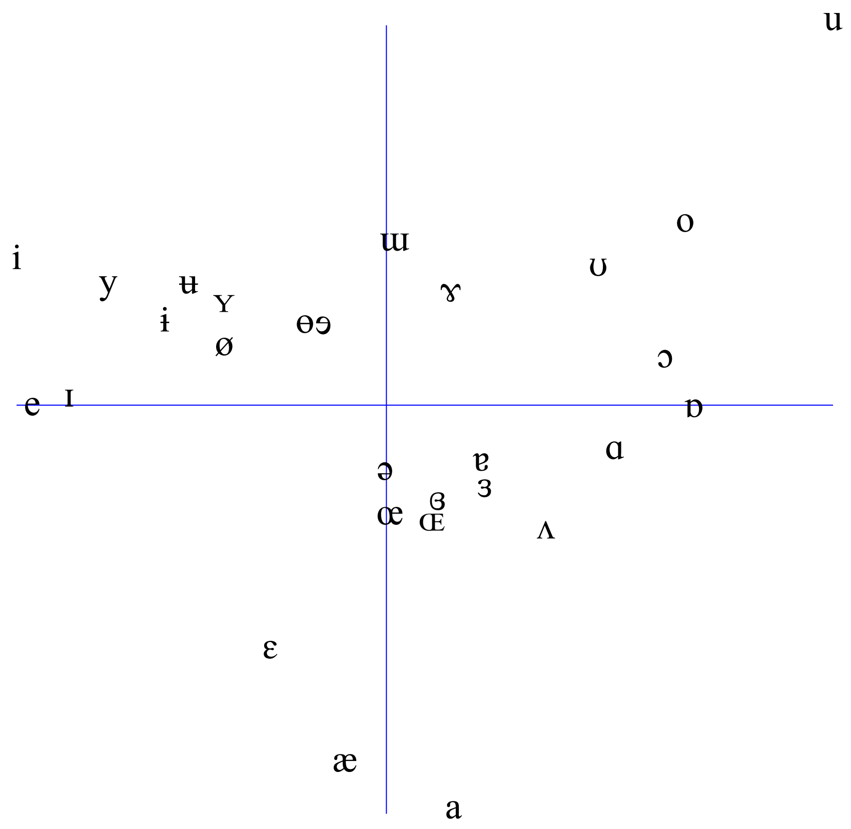


Figure 4.12: Two-dimensional multidimensional scaling plot obtained from the formant track distances between all pairs formed by the 28 vowels. Two dimensions explain 98% of the variance. The first dimension (x-axis) corresponds with advancement and the second dimension (y-axis) with height. Note that the schwa [ə] is located about in the middle. A third dimension would be less easily interpreted. In this plot some symbols are shifted a little bit.

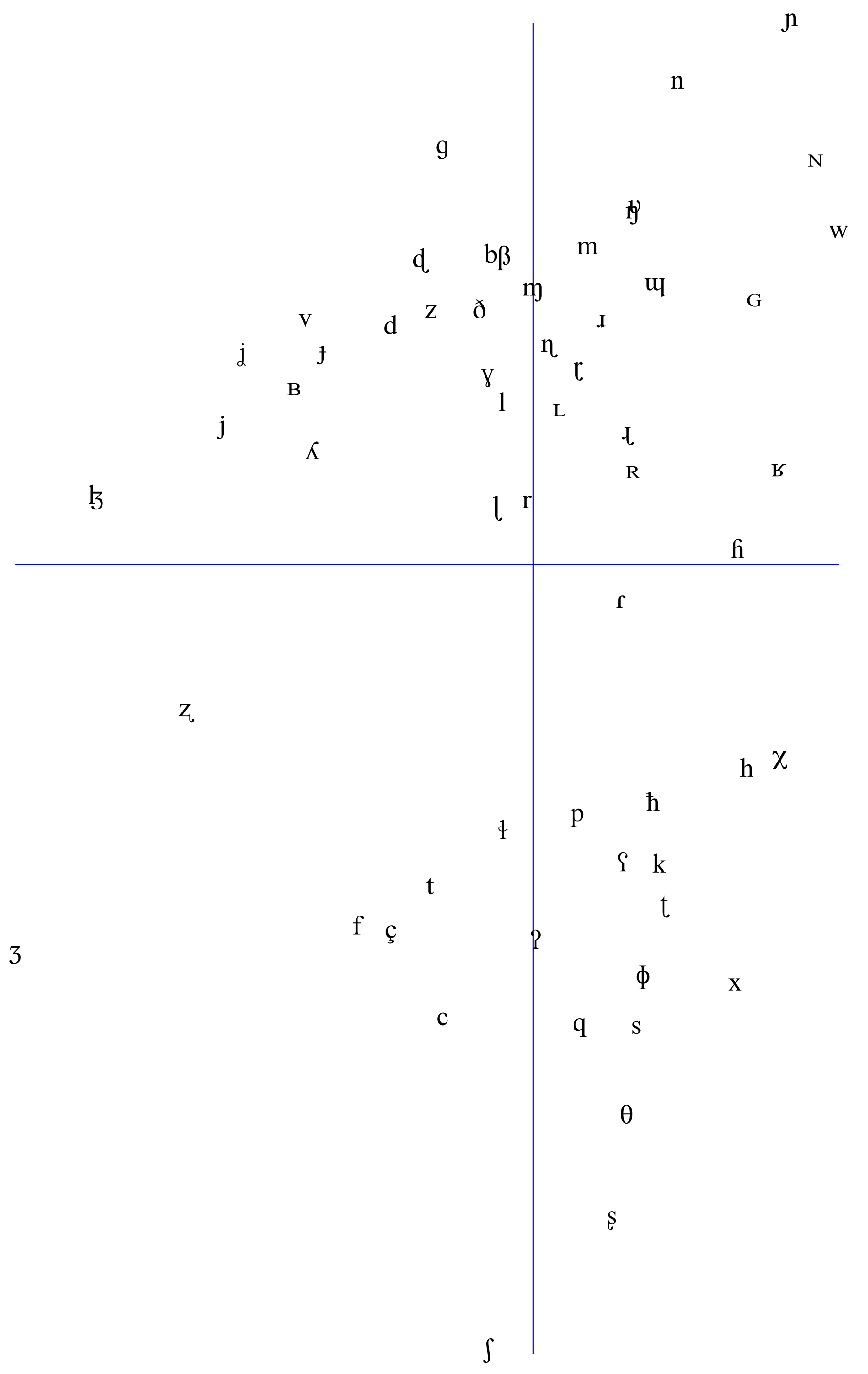


Figure 4.13: Two-dimensional multidimensional scaling plot obtained from the formant track distances between all pairs formed by the 59 consonants. Two dimensions explain 92% of the variance. The first dimension (y-axis) distinguished between voiceless and voiced sounds. For the second dimension (x-axis) we found no obvious interpretation. For a third dimension we found no interpretation as well.

three dimensions 98%. In Figure 4.14 a two-dimensional multidimensional scaling plot is shown on the basis of all sounds. The first dimension (the vertical dimension in the plot) distinguishes between high vowels (high) and low vowels (lower), and between voiced consonants (high) and voiceless consonants (low). The second dimension (horizontal) distinguishes between front vowels (left) and back vowels (right), and between 'front consonants' (left) and 'back consonants' (right). See the separate plots of vowels and consonants for more explanation. When drawing a line from [i] to [a], from [a] to [ɒ], from [ɒ] to [u], and from [u] to [i], we get a triangle again. High vowels and voiced consonants have a lower F1, and low vowels and voiceless consonants a higher F1. Front vowels and 'front consonants' have a higher F2, and back vowels and 'back consonants' a lower F2. In the middle of the triangle we find the r-like sounds, similar to the Barkfilter plot (see Figure 4.7). The nasals are located around the line between the [i] and the [u], however, closer to the [u] than to the [i]. The laterals are located in the corner of the [i]. Most voiced plosives are located above the line between the [u] and the [i], closer to the [i] than to the [u]. The voiceless velar, uvular, and pharyngeal fricatives, and both the voiceless and voiced glottal fricatives are located between the [a] and the [ɒ]. The [i] and the [j] are much closer than the [u] and the [w]. This is the opposite of what we saw in the Barkfilter plot (see Figure 4.7) and the cochleagram plot (see Figure 4.10). The correct closeness of the [i] and the [j] may be explained by the fact the formants are not sensitive to differences in intensity. However, the relatively large distance between the [u] and the [w] can be explained from the fact that we used a sample of Jill House for the [w] in the set of samples of John House (this is justified in Section 4.2). Formants do not neutralize differences in gender. Furthermore, 'silence' is not as close to the glottal stop as when using the Barkfilter representation or the cochleagram representation (see Figure 4.10). We also examined the third dimension but found no clear interpretation for this.

Summarizing, we found that the vowel classification is like the IPA quadrilateral, and the consonant classification reflects the different manners of articulation. Therefore, we conclude that the formant track representation is useful for finding segment distances. For both the vowels and the consonants we found striking differences with respect to the Barkfilter representations and cochleagram representation. For the vowels we found no quadrilateral, but a triangle when connecting the corner points [i], [u], [ɒ] and [a]. This is in accordance with results found in literature (data of Pols (1977) visualized by Rietveld and Van Heuven (1997, p. 133)). For the consonants there is no clear separation between plosives and fricatives. This may be explained from the fact that only formant tracks are used which gives less information than the Barkfilter and cochleagram representation. Therefore, the formant track representation is more accurate for finding vowel distances than for finding consonant distances.

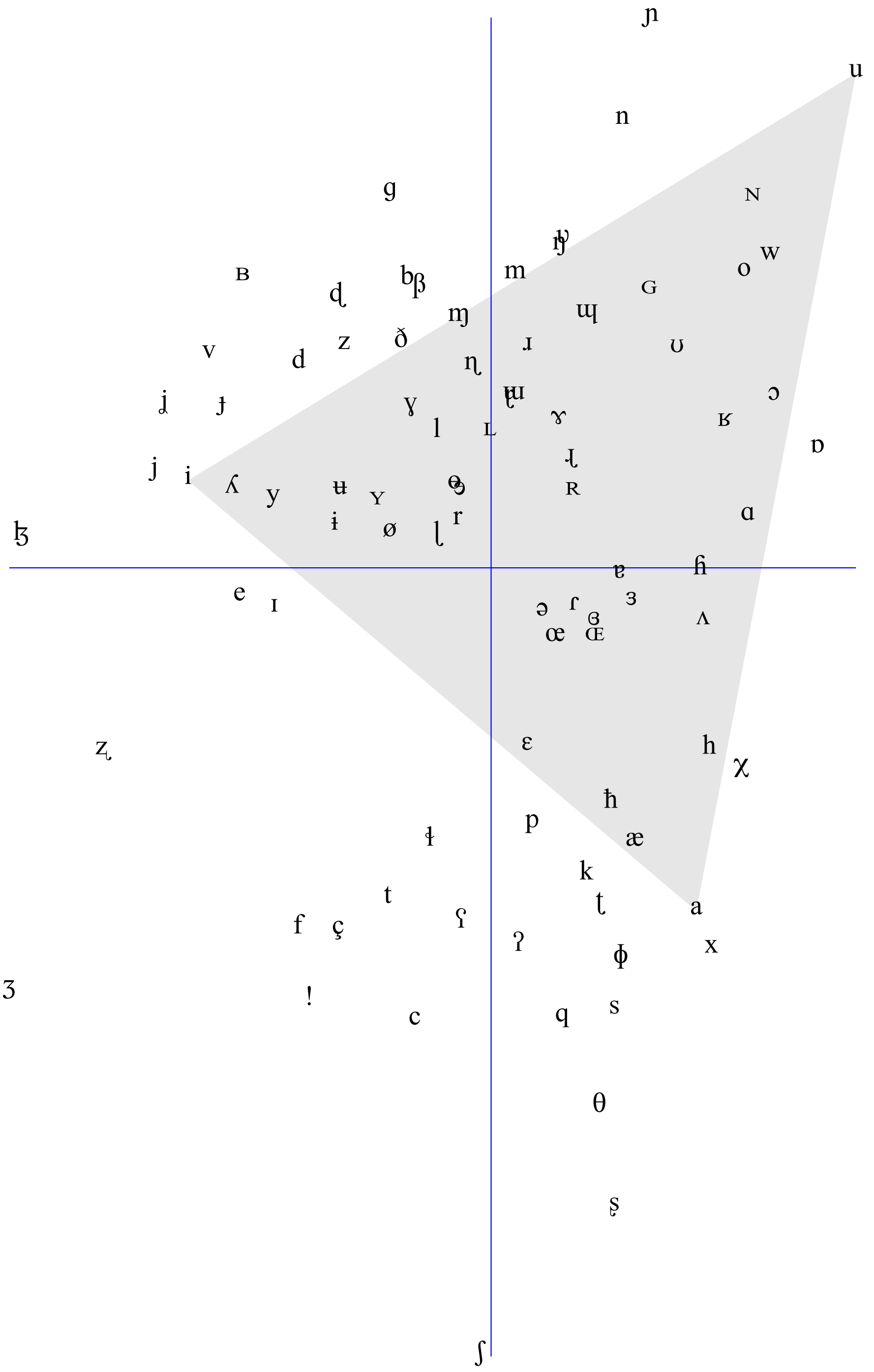


Figure 4.14: Two-dimensional multidimensional scaling plot on the basis of the formant track distances between all pairs formed by the 28 vowels, the 59 consonants and 'silence'. In the plot, the ! is used for 'silence'. Two dimensions explain 92% of the variance. The first dimension (y-axis) represents the F1 (higher sounds have lower F1 values) and the second dimension (x-axis) the F2 (sounds on the left have higher F2 values). The shaded area represents a vowel triangle.

## 4.4 Diphthongs

As mentioned in Section 3.2, there are two possibilities for processing diphthongs. When using the acoustic representations we want to be able to consider both approaches again. In the first approach, a diphthong is considered as nothing more than a sequence of two monophthongs. In the second approach, a diphthong may also be regarded as one sound with a changing color. In the optimal case we should have samples of all diphthongs that may occur in our data. However, on the tape *The Sounds of the International Phonetic Alphabet* no samples of diphthongs can be found. To be able to process diphthongs as one sound nonetheless, we modify the calculations in ways which should have the desired effect.

The distance between a monophthong and a diphthong is calculated as the mean of the distance between the monophthong and the first element of the diphthong and the distance between the monophthong and the second element of the diphthong. So the distance between e.g. [ai] and [ɛ] is calculated as the mean of the distance between [a] and [ɛ] and the distance between [i] and [ɛ]. We will discuss this in more detail. Assume the front vowels [a], [æ], [ɛ], [e] and [i] are on a straight line, just as in the IPA quadrilateral. Assume the distance from [a] to [æ] is 0.5, from [æ] to [ɛ] is 0.5, from [ɛ] to [e] is 1, and from [e] to [i] is 1. What are now the distances from the [ai] to the starting point, the intermediate points and the end point? For each point we calculate the distance to the [a] and to the [i] and take the average of both segments:

| | [a] | [i] | [ai] |
|---|---|---|---|
| [i] | 3.0 | 0.0 | 1.5 |
| [e] | 2.0 | 1.0 | 1.5 |
| [ɛ] | 1.0 | 2.0 | 1.5 |
| [æ] | 0.5 | 2.5 | 1.5 |
| [a] | 0.0 | 3.0 | 1.5 |

We see that for each of the points the distance to the [ai] is the same. In fact the distance is simply equal to $d$([a],[i])/2. This is in accordance with the idea that the color of a [ai] gradually changes from [a] to [æ], from [æ] to [ɛ], from [ɛ] to [e], from [e] to [i]. All (intermediate) points are heard in the diphthong during an infinitesimally small moment in time. However, in our acoustic results the [æ], [ɛ] and [i] do not lay exactly on the line from [a] to [i] (see Figures 4.5, 4.8 and 4.12). The distances of these 'intermediate points' will be greater than $d$([a],[i])/2.

The distance between two diphthongs is calculated as the mean of the distance between the first elements and the distance between the second elements. So the distance between e.g. [au] and [ɛi] is calculated as the mean of the distance between [a] and [ɛ] and the distance between [u] and [i].

## 4.5 Affricates

In the RND data no affricates are used. However, in the NOS data they do appear. On the IPA tape *The Sounds of the International Phonetic Alphabet* two affricates can be found, namely the [k͡p] and the [t͡s]. However, to be able to process many more affricates, we did not use these sample but applied the more general approach as given in Section 3.3. When processing affricates both elements are processed as extra-short, separated elements.

## 4.6 Suprasegmentals and diacritics

The sounds on the tape *The Sounds of the International Phonetic Alphabet* are pronounced without suprasegmentals and diacritics. However, a restricted set of suprasegmentals and diacritics can be processed in our system. Since no features can be changed, only those suprasegmentals and diacritics are taken into account which can be processed by changing the weighting of segments or by averaging sound distances. Suprasegmentals and diacritics which are processed by the first approach are discussed in Section 4.6.1, those which are processed by the second approach in Section 4.6.2.

### 4.6.1 Weighting segments

In Section 3.4.2 we stated that the weighting of extra-short sounds should be halved with respect to the weighting of short sounds. Just as we did with the discrete representations, for the acoustic representations we realize this by changing the transcription beforehand. We retain the extra-short sounds as they are and double all other sounds.

When using discrete representations, we consider two approaches for the processing of *half-long* and *long*. When using acoustic representations, both approaches are regarded again. In fact, for the acoustic representations exactly the same applies as for the discrete phone representations. In the first approach, *half-long* and *long* are not processed since they may sometimes be redundant to some extent. In the second approach, the two length marks are processed by changing the transcription. Half-long sounds are trebled and long sounds are quadrupled. For more details see Section 3.4.2.

In the RND, consonants may also be vocalized. We process vocalized sounds as syllabic sounds. Vocalized (RND) or syllabic sounds (NOS) are marked with the diacritic *syllabic*. Using the acoustic representation, syllabic sounds are processed in the same way as when using the discrete phone representation. We consider two approaches for processing syllabic sounds which corresponds with the two approaches that are regarded when processing *half-long* and *long*. In the first approach *syllabic* is not processed since it may be redundant. In the second approach this diacritic is processed by changing the transcription beforehand. Syllabic sounds are processed as long sounds, i.e. they are quadrupled in the transcription. For more details see Section 3.4.2 again.

For the RND *aspirated* is not processed. However, for the NOS data it is processed. An [h] is inserted after the phone which was noted to be aspirated. This [h] is noted as extra-short, so the significance is halved. For more details see Section 3.4.5.

### 4.6.2 Averaging segments

When using acoustic representations, the diacritics *voiceless*, *voiced*, *apical* and *nasalized* can be processed. When comparing sound $x$ and sound $y$, one or more diacritics may be noted after one or both sounds. To process them, first the distance between $x$ and $y$ is calculated as it is without any diacritics. This is a basic distance and mentioned in the first part of Table 4.1. A counter is set to 1. Next, we check whether the diacritics *voiceless* or *voiced* are used. The possible combinations are listed in the second part of Table 4.1 in the column ‘condition’. If one of the conditions applies, the corresponding distance increase as given in the column ‘distance increase’ is calculated and added to the basic distance. The counter is increased by 1. Subsequently we check whether the diacritic *apical* is used. The possible combinations are listed in the third part of Table 4.1 under ‘condition’. If one of the conditions apply, the corresponding distance increase as suggested under ’distance increase’ is added to the basic distance, and the counter is increased by 1. Finally it is checked whether the feature *nasal* is used. If one of the conditions in the fourth part of Table 4.1 apply, the corresponding distance increase is added to the basic distance and the counter is increased by 1.

If no diacritics were noted, the total distance is equal to the basic distance, and the counter is equal to 1. If all diacritics were found, the largest distance is obtained, and the counter is equal to 4. Now the final distance is equal to the total distance (basic distance plus optionally one or more diacritic increases) divided by the counter.

The idea behind the calculation of the distance increase of the diacritics *voiceless* and/or *voiced* is that a voiced voiceless sound or a voiceless voiced sound is exactly intermediate between a voiceless sound and a voiced sound. In the table $X$ is the voiced counterpart of a voiceless $x$, or the voiceless counterpart of a voiced $x$. The $Y$ is defined analogously to the $X$. For voiced sounds which have no voiceless counterpart (the sonorants), or for voiceless sounds which have no voiced counterpart (the glottal stop) the sound itself is used. Since the RND sounds are a subset of the IPA sounds, there are no voiced counterparts for the [c] and the [h], so respectively the [c] and the [h] are returned. When the diacritic *voiceless* is noted under a voiceless sound, it is most likely that this an error. Instead of this diacritic the diacritic *voiced* is processed. Conversely the diacritic *voiced* under a voiced sound is processed as the diacritic *voiceless*.

The diacritic *apical* is implemented in the NOS system only. The implementation of this diacritic was made to be able to process Romanesque languages as well. Only the [s] and [z] are allowed to be apical in our system. The /s/ "of standard Spanish is an apical-alveolar sound" (Pountain, 2001, p. 299). The tip of the tongue is often "retroflexed or turned back as it touches the alveolar ridge" (Dalbor, 1969, p. 91). In some Sardinian dialects the same apical-alveolar sound is found. When comparing a non-apical sound with an apical sound, the distance increase is equal to the distance between the non-apical sound and the [ʂ] (if the apical sound was a [s]) or the [ʐ] (if the apical sound was a [z]).

The thought behind the way in which the diacritic *nasal* is processed is that a nasal sound is about intermediate between its non-nasal version and the [n]. So when comparing a non-nasal sound with a nasal sound, we quantify the effect of the diacritic *nasal* by calculating the distance between the non-nasal sound and the [n].

## 4.7 Comparison of segments

In this section, we explain the comparison of segments in order to get distances between segments that will be used in the Levenshtein distance (see Section 5.1). In a Barkfilter or cochleagram, the intensities or loudnesses of frequencies are given for a range of times. A spectrum contains the intensities or loudnesses of frequencies at one point in time. In a formant track representation, the formants are given for a range of times. A formant bundle contains the formants for one point in time. The smaller the time step, the more spectra or formant bundles in the acoustic representation. Per acoustic representation we consistently used the same time step for all samples.

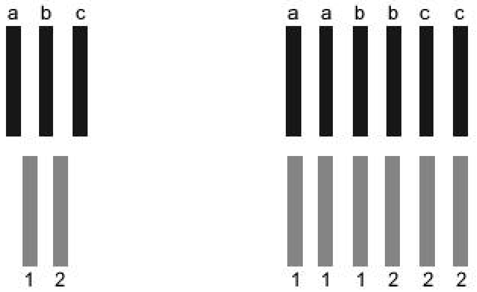
It appears that the duration of the segment samples varies. This may be explained by variation in speech rate. Duration is also a sound-specific property. E.g., a plosive is shorter than a vowel. The result is that the number of spectra of formant bundles per segment may vary, although for each segment sample the same time step was used. Since we want to normalize the speech rate and regard segments as linguistic units, we see to it that two segments get the same number of spectra or formant bundles when they are compared to each other.

When comparing one segment of $m$ spectra or formant bundles with another segment of $n$ spectra or formant bundles, each of the $m$ elements is duplicated $n$ times, and each of the $n$ elements is duplicated $m$ times. So both segments get a length of $m \times n$. Below two segments are schematically visualized, one with 3 elements (black bars) and one with 2 elements (grey bars). Now both get a

| | Diacritic | Condition | Distance increase | Counter increase |
|---|---|---|---|---|
| 1 | basis | *x* vs. *y* | *d*(*x*,*y*) | 1 |
| 2 | voice | *x̥* vs. *y* | *d*(*X*,*y*) | 1 |
| | | *x* vs. *y̥* | *d*(*x*,*Y*) | |
| | | *x̥* vs. *y̥* | *d*(*X*,*Y*) | |
| | | *x̬* vs. *y* | *d*(*X*,*y*) | |
| | | *x* vs. *y̬* | *d*(*x*,*Y*) | |
| | | *x̬* vs. *y̬* | *d*(*X*,*Y*) | |
| | | *x̥* vs. *y̬* | *d*(*X*,*Y*) | |
| | | *x̬* vs. *y̥* | *d*(*X*,*Y*) | |
| 3 | apical | [s̺] vs. *y* | *d*([ʂ],*y*) | 1 |
| | | [z̺] vs. *y* | *d*([ʐ],*y*) | |
| | | *x* vs. [z̺] | *d*([s],[ʐ]) | |
| | | *x* vs. [s̺] | *d*([z],[ʂ]) | |
| | | [s̺] vs. [z̺] | *d*([ʂ],[ʐ]) | |
| | | [z̺] vs. [s̺] | *d*([ʐ],[ʂ]) | |
| | | [s̺] vs. [s̺] | *d*([ʂ],[ʂ]) | |
| | | [z̺] vs. [z̺] | *d*([ʐ],[ʐ]) | |
| 4 | nasal | *x̃* vs. *y* | *d*([n],*y*) | 1 |
| | | *x* vs. *ỹ* | *d*(*x*,[n]) | |
| | | *x̃* vs. *ỹ* | *d*([n],[n]) | |

Table 4.1: When $x$ and $y$ are sounds with the diacritics, first the basis distance is calculated (1) and a counter is set to 1. Next for each valid condition under 'Condition' (2,3,4) the basis distance is increased with the corresponding distance under 'Distance increase' and the counter is increased by 1. The final distance is the total distance divided by the counter. $X$ and $Y$ are the voiceless or voiced counterparts of the voiced or voiceless $x$ and $y$.

length of 6 when each of the 3 elements are duplicated 2 times, and each of the 2 elements are duplicated 3 times.



For finding the distance between two sounds the Euclidean distance is calculated between each pair of corresponding spectra or formant bundles, one from the first, and one from the second sound. Assume a spectrum or formant bundle $e1$ and $e2$ with respectively $n$ frequencies or formants, then:

$$d(e1, e2) = \sqrt{\sum_{i=1}^{n}(e1_i - e2_i)^2} \tag{4.3}$$

The distance between two segments is equal to the sum of the spectrum or formant bundle distances divided by the number of spectra or formant bundles. In this way we found that the greatest distance occurs between the [a] and 'silence' (Barkfilter, Cochleagram) or between the [u] and the [ʃ] (formants). We regard the maximum distance as 100%. Other segment distances are divided by this maximum and multiplied by 100. This gives segment distances expressed in percentages. Word distances and dialect distances which are based on them may also be given in terms of percentages.

In Section 3.7 we explained that segment distances obtained on the basis of feature definitions may be used in two ways. First, the distances can be used directly, i.e. linearly. Second the logarithms of the distances can be taken. The latter reflects the fact that in perception small differences in pronunciation may play a relatively strong role in comparison with larger differences. For the acoustic segment distances we consider both the linear and logarithmic approach again. Since the logarithm of 0 is not defined, and the logarithm of 1 is 0, distances are increased by 1 before the logarithm is calculated. To obtain percentages, we calculate $ln(distance + 1)/ln(maximum\ distance + 1)$.

## 4.8 Correlation between systems

In this section, we compare the different acoustic systems in order to check for striking differences between them. In Section 4.8.1 we compare the different representations by calculating the correlation coefficient between the distances, just as we did for the different feature representations (see Section 3.8).

We also compare the acoustic systems with the discrete systems. In Section 4.8.2, we correlate the acoustic based distances with the distances obtained on the basis of feature representations (see Section 3.1) in order to examine to what extent acoustic distances differ from feature based systems.

We calculate correlations on the basis of both the set of RND sounds and the set of IPA sounds. Again we have 18 vowels and 27 consonants for the RND, and 28 vowels and 59 consonants for the IPA. When we correlate on the basis of both vowels and consonants, we get respectively 46 sounds for the RND and 88 sounds for the IPA.

For finding the significance of a correlation coefficient we used the Mantel test, just as in Chapter 3. The Mantel test was also used to determine whether one correlation coefficient is significantly higher than another. The Mantel test is explained in Section 3.8.2. As significance level we again choose $\alpha = 0.05$.

### 4.8.1 Acoustic vs. acoustic

The Tables 4.2 and 4.3 show the correlation coefficients between matrices of segment distances obtained on the basis of the Barkfilter, the cochleagram and the formant track representation. Correlations are given for both the RND and the IPA segment distances. Results are given for vowels, consonants and all sounds. When all segments are used, 'silence' is also included. All correlations are significant for $\alpha = 0.05$.

For both the RND and the IPA the Barkfilter and the cochleagram distances correlate significantly more strongly than the other pairs of representations for vowels, consonants and all segments. This outcome is not surprising since these representations are most similar. Looking at the vowels, the correlation between the formant-track distances and the Barkfilter distances is not significantly weaker or stronger than the correlation between the formant track distances and the cochleagram distances. However, when looking at the consonants and all sounds, the formant-track distances yield correlations significantly stronger with the cochleagram distances than with the Barkfilter distances. We observed this but cannot explain it.

Since the correlations between the Barkfilter distances and cochleagram distances are rather high ($0.94 \leq r \leq 1.00$), dialect distances based on Barkfilter distances and cochleagram distances are expected to be similar. The correlations between the formant track distances and the Barkfilter distances are lower ($0.45 \leq r \leq 0.78$) just as the correlations between the formant track distances and the cochleagram distances ($0.51 \leq r \leq 0.78$). So we expect that the use of formant track segment distances in dialect comparison will give significantly different results than when using Barkfilter distances or cochleagram distances.

| | | | vow. | cons. | all |
|---|---|---|---|---|---|
| Bark. | vs. | Coch. | 1.00 | 0.94 | 0.96 |
| Bark. | vs. | Form. | 0.77 | 0.64 | 0.45 |
| Coch. | vs. | Form. | 0.77 | 0.74 | 0.51 |

Table 4.2: Correlation coefficients among RND segment distances between vowels (vow.), consonants (cons.) and all segments obtained on the basis of the Barkfilter (Bark.), the cochleagram (coch.) and the formant track (form.) representation. When all segments are used, 'silence' is also included.

| | | | vow. | cons. | all |
|---|---|---|---|---|---|
| Bark. | vs. | Coch. | 1.00 | 0.94 | 0.95 |
| Bark. | vs. | Form. | 0.78 | 0.59 | 0.51 |
| Coch. | vs. | Form. | 0.78 | 0.70 | 0.57 |

Table 4.3: Correlation coefficients among IPA segment distances between vowels (vow.), consonants (cons.) and all segments obtained on the basis of the Barkfilter (Bark.), the cochleagram (coch.) and the formant track (form.) representation. When all segments are used, 'silence' is also included.

### 4.8.2 Acoustic vs. features

We also examined the correlations between the acoustic distances and the feature-based distances. Results are given in Tables 4.4 and 4.5. Almost all correlations between the acoustic representations and the feature representations are significant for $\alpha = 0.05$. Only the lowest correlations which are found between the A & B representations (Manhattan and Euclidean distance) and the formant track representation for the RND consonants are not significant.

Looking at the vowel correlations, we observe that the Barkfilter distances correlate strongest with the A & B distances, regardless which feature bundle metric is used. The three A & B correlations are not significantly higher than the corresponding ones of the two other feature systems, however. Just as for the Barkfilter distances, the cochleagram distances correlate strongest with the A & B distances. However, the correlations for the three feature bundle metrics do not differ significantly among the different feature systems. The formant track distances correlate strongest with the V & C distances for most metrics. However the correlations for all three feature bundle metrics are for the most part not significantly higher than the same metrics applied to different feature systems. In the feature system of A & B height has a greater weight than advancement (see Table 3.12). Examining Figures 4.5 and 4.8, we observe that for the Barkfilter and the cochleagram representation height is weighted more strongly than advancement as well. In the feature system of V & C advancement has a greater

| | | Bark. | | Coch. | | Form. | |
|---|---|---|---|---|---|---|---|
| | | vow. | cons. | vow. | cons. | vow. | cons. |
| H & H | M. | 0.39 | 0.29 | 0.38 | 0.36 | 0.56 | 0.35 |
| | E. | 0.41 | 0.29 | 0.39 | 0.35 | 0.58 | 0.34 |
| | P. | 0.31 | 0.35 | 0.30 | 0.40 | 0.49 | 0.35 |
| V & C | M. | 0.54 | 0.35 | 0.52 | 0.37 | 0.72 | 0.25 |
| | E. | 0.55 | 0.36 | 0.54 | 0.37 | 0.74 | 0.26 |
| | P. | 0.36 | 0.40 | 0.34 | 0.43 | 0.62 | 0.27 |
| A & B | M. | 0.67 | 0.19 | 0.65 | 0.19 | 0.64 | 0.02 |
| | E. | 0.67 | 0.21 | 0.66 | 0.19 | 0.66 | 0.03 |
| | P. | 0.71 | 0.21 | 0.70 | 0.22 | 0.61 | 0.09 |

Table 4.4: Correlations among segment distances as specified by three feature systems and three acoustic systems on the basis of the distances between the RND segments. Feature bundle distances are found by calculating the (M)anhattan distance, (E)uclidean distance or (P)earson correlation coefficient.

| | | Bark. | | Coch. | | Form. | |
|---|---|---|---|---|---|---|---|
| | | vow. | cons. | vow. | cons. | vow. | cons. |
| H & H | M. | 0.40 | 0.31 | 0.39 | 0.35 | 0.43 | 0.26 |
| | E. | 0.40 | 0.30 | 0.39 | 0.33 | 0.43 | 0.25 |
| | P. | 0.24 | 0.38 | 0.24 | 0.41 | 0.24 | 0.30 |
| V & C | M. | 0.52 | 0.36 | 0.51 | 0.39 | 0.60 | 0.30 |
| | E. | 0.51 | 0.36 | 0.50 | 0.38 | 0.59 | 0.30 |
| | P. | 0.36 | 0.42 | 0.34 | 0.47 | 0.50 | 0.32 |
| A & B | M. | 0.68 | 0.23 | 0.67 | 0.22 | 0.58 | 0.07 |
| | E. | 0.68 | 0.25 | 0.66 | 0.24 | 0.58 | 0.08 |
| | P. | 0.72 | 0.27 | 0.71 | 0.27 | 0.55 | 0.12 |

Table 4.5: Correlations among segment distances as specified by three feature systems and three acoustic systems on the basis of the distances between the IPA segments. Feature bundle distances are found by calculating the (M)anhattan distance, (E)uclidean distance or (P)earson correlation coefficient.

weight than height (see Table 3.6). Examining Figure 4.12 we also find for the formant track representation that advancement has a greater weight than height. All acoustic representations correlate worst with the H & H system, although not significantly lower than with other systems. The lower correlations may be explained by the unnatural way in which height is defined (see Section 3.1.2.1).

Looking at the consonant correlations we observe that the H & H distances in most cases and the V & C distances in all cases correlate significantly better with the Barkfilter distances than the A & B distances do. The V & C correlations are higher than the H & H correlations, but they are not significantly higher. The H & H distances and the V & C distances correlate in all cases significantly better with the cochleagram distances than the A & B distances do. Just as for the Barkfilter the V & C correlations are higher than the H & H correlations, but again they are not significantly higher. The higher correlations of both the Barkfilter distances and the cochleagram distances with the V & C distances may be explained by the categorical way in which manner of articulation is defined in the system of V & C. The worse correlation with the A & B system may be explained by the fact that in this feature system, manner of articulation is defined as a scale. The H & H distances and the V & C distances correlate in all cases significantly better with the formant track distances than the A & B distances do. Using the RND consonants, the formant track distances correlate strongest with the V & C distances, but not significantly more strongly than with the H & H distances. When using the IPA consonants, the formant track distances correlate strongest with the H & H distances, but not significantly more strongly than with the V & C distances. The higher correlations of the V & C distances and the H & H distances may again be explained by the fact that manner of articulation is defined as a scale in the A & B system. The difference between the RND and the IPA may be explained by the fact that in the system of H & H all RND consonants are uniquely defined, but all IPA consonants are not.

As explained in Section 3.1.3, the V & C feature system is perceptually based. We expect that the V & C distances will correlate more strongly with the cochleagram distances than with the Barkfilter distances since the cochlear model is a more exact model of the cochlea than the Barkfilter model. For the vowels we see exactly the opposite: the Barkfilter distances correlate more strongly with the V & C distances than the cochleagram distances, although no significant differences between correlation coefficients were found. For the consonants we find what we expected: the cochleagram distances correlate more strongly with the V & C distances than the Barkfilter distances. However, the differences between the correlations coefficients are not significant. It should be interesting to correlate the complete set of vowels *and* consonants with perceptually based distances. However, vowels and consonants are separated in the V & C system, so unfortunately, this was not possible.

Although almost all correlation coefficients are significant, they are not extremely high. For the vowels the highest correlation for the Barkfilter distances is 0.72 (with respect to A & B, Pearson, IPA), for the cochleagram distances 0.71 (A & B, Pearson, IPA) and for the formant track distances 0.74 (V & C, Euclidean, IPA). For the consonants the highest correlation for the Barkfilter distances is 0.42 (with respect to V & C, Pearson, IPA), for the cochleagram distances 0.47 (V & C, Pearson, IPA) and for the formant track distances 0.35 (H & H, Manhattan and Pearson, RND). The vowel distances correlate more strongly than the corresponding consonant distances in most cases. Especially when regarding the lower consonant correlations, we expect that the use of acoustic segment distances in dialect comparison will give results that are different with respect to feature-based results.

## 4.9 Conclusions

In this chapter we presented the use of acoustic representations for finding distances between segments. In contrast to most feature systems acoustic representations are based on physical measurements. We examined the Barkfilter, cochleagram and formant track representation, which are more perceptually oriented models. We performed multidimensional scaling on the acoustic distances and scaled them to two dimensions. For all representations we obtained a vowel classification which is like the IPA quadrilateral, and the consonant classification reflects the different manners of articulation. Therefore, we conclude that the three representations are useful for finding segment distances. The Barkfilter and the cochleagram representations correlate significantly more strongly than any other pairs of representations. The results obtained on the basis of the formant track representation are more different. With the formant track representation a vowel triangle is obtained, and for the consonants no clear separation between plosives and fricatives was found.

When correlating distances obtained by the feature representations with distances obtained by the acoustic representations, it appears that, for the vowels, both the Barkfilter distances and the cochleagram distances correlate strongest with the A & B distances. The formant track distances correlate strongest with the V & C distances for most metrics. The correlation coefficients were for the most part not significantly higher than other comparable correlation coefficients. All acoustic representations correlate worst with the H & H system, but not significantly lower than with other systems. The lower correlations can be explained by the unnatural way by which height is defined in the system of H & H. Therefore, for vowels we prefer A & B and V & C to H & H. However we made no choice between A & B and V & C. On the one hand, the Barkfilter and cochleagram representations contain more information, on the other hand, the formant track representation may be limited to information which is relevant in perception.

For the consonants, the Barkfilter distances and the cochleagram distances correlate strongest with the V & C distances, but only significantly better than with the A & B distances. The formant track distances correlate strongest with the V & C distances (RND) or H & H distances (IPA). The correlation coefficients were only significantly higher than the comparable ones of A & B. This suggests that manner of articulation should not be represented as a scale (as in A & B), but as different categories (as in V & C and H & H). Therefore, for consonants we find V & C and H & H preferable to A & B, but cannot make a choice between V & C and H & H.

When correlating feature-based segment distances with acoustically-based distances all correlation coefficients are significant. For both vowels and consonants they are not extremely high, although for vowels higher correlation coefficients were found than for consonants. Therefore, in Chapter 7 the use of both feature-based and acoustically-based segment distances will be validated.

# Chapter 5

# Measuring dialect distances

In the Chapters 3 and 4 we described how distances between phonetic segments are found. When the segments are aligned, we are able to find distances between words, and in turn between language varieties. The way in which distances are found between words and between language varieties is the topic of this chapter. The central algorithm in this chapter and in this research is the Levenshtein distance, a method that allows distances between words to be measured. This algorithm may be applied to both transcriptions of words and to the representations of the acoustic signals of word samples. The application of the Levenshtein distance to transcriptions of words is described in Section 5.1. This approach uses the phonetic segment distances as measured in the Chapters 3 and 4. In Section 5.2 we describe the application of the Levenshtein distance to acoustic word samples. In this approach a transcription is only used for finding the number of phonetic segments per word. The segment distances as measured in the Chapters 3 and 4 are not used.

## 5.1 Levenshtein distance using transcriptions

### 5.1.1 Sequence comparison

Sequence comparison is used in many different fields. Kruskal (1999) gives an overview. First, Kruskal mentions the application to molecular biology, where sequence comparison is used for the comparison of macromolecules. An example that is more related to our research is the application of sequence comparison to speech and speaker recognition. Sequence comparison is also used for correction of typing errors on a computer or keypunch machine, for the comparison of comparable computer files and for error control of codes which are transferred by e.g. radio or telegraph. Levenshtein (1966) 'presented the earliest known use of a distance function that is appropriate in the presence of insertion and deletion errors' (Kruskal, 1999, p. 5). Sequence comparison is also applied in gas chromatography, a physical method used to separate and/or analyse complex mixtures. A 'mixture is swept by a continuous stream of nonreactive carrier gas through a long, densely packed column of special material' (p. 6). Components with strong attraction to one part of the column move more slowly than those with weak attraction. 'The components emerge at different times over a period of minutes or hours' (p 6). A chromatogram shows different peaks in time. The peaks correspond to the intensities of the different components in the sample mixture. Chromatograms are sequences that are compared to each other. Also related to our research is the application of sequence comparison to bird song. In 'some bird species, song is an important means of communication, which is learned by the young from their elders, and it has dialect-like variation from place to place' (p. 7). Another application of sequence comparison is found in the comparison of stratigraphic sequences, tree rings and varves ('annual layers of sediment, generally clay, in which is it possible to count the years', p. 7). More related to our research, sequence application is applied to collation of different versions of the same text. Furthermore, sequence comparison is found when 'computer processing handwritten material such as signatures and line drawings' (p. 8). Comparison of "brain waves" in response to a stimulus may also be application of sequence comparison.

In our research we apply sequence comparison to the comparison of different pronunciations corresponding to different language varieties in order to measure the distance between them. Kruskal mentions several methods which require that sequences have the same length. Examples are Hamming distance (the number of positions in which the corresponding elements are different), Manhattan (or city-block) distance, and Euclidean distance (see Section 3.6.2.5). However different pronunciations will not have the same length in many cases. Also, the correspondences made in the methods just mentioned may not always be correct. E.g. *afternoon* may be pronounced as [ˈæəftəˌnʉ·n] in the dialect of Savannah, Georgia and as [ˌæftərˈnu·n] in the dialect of Lancaster, Pennsylvania.[1] Assume we compare both pronuncations using the Hamming distance. When ignoring diacritics this is done as follows:

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| æ | ə | f | t | ə | n | ʉ | n |
| æ | f | t | ə | r | n | u | n |
| | 1 | 1 | 1 | 1 | | 1 | |

We get a cost of 5. However, we see that elements which correspond to one another, are unfortunately not regarded as corresponding elements when calculating the Hamming distance. The consequence is that the distance calculated between the two pronunciations is too high.

Both the length and the correspondence problem are solved when using the *Levenshtein distance.* This algorithm is able to deal with different lengths calcu-

[1] The data is taken from the *Linguistic Atlas of the Middle and South Atlantic States* (LAMSAS) and available via `http://hyde.park.uga.edu/lamsas/`.

lating the distance on the basis of probable correspondences. Although Levenshtein applied his algorithm to error control of codes which are transfered by e.g., radio or telegraph, the algorithm may be applied to all cases mentioned at the beginning of this section. For the comparison of genetic macromolecules and the recognition of human speech the algorithm has actually been used. In speech recognition the algorithm is often referred to as (*dynamic*) *time-warping* or *dynamic programming*. The Levenshtein distance was first applied by Kessler (1995) to dialect comparison. He used the algorithm for the comparison of Irish Gaelic varieties. In our research we have used the algorithm for the comparison of Dutch and Norwegian varieties. See Bolognesi and Heeringa (2002) for an application to Sardinian dialects.

### 5.1.2 Minimum cost

Fundamental to the idea of the Levenshtein distance is the notion of string-changing operations. To determine the extent to which two strings differ from each other, an inventory of what operations can change one string into another should be made. The operations available are:

- Deletions
  Delete an element from the string.

- Substitutions
  Replace an element from the one string by an element of the other string.

- Insertions
  Add an element to the string.

*In*sertions and *del*etions are also referred to as *indels*. To each of the operations weights are assigned. In the simplest case all operations have the same weight, e.g. 1 (Gusfield, 1999, p. 218). When applying these weights to the comparison of word pronunciations, we judge that roughly speaking, substitutions will be about equally noticeable as indels from the perceptual point of view. We illustrate the use of the operations with an example. As mentioned in Section 5.1.1 *afternoon* may be pronounced as [ˈæəftəˌnʉˑn] in the dialect of Savannah and as [ˌæftərˈnuˑn] in the dialect of Lancaster. Ignoring diacritics for this moment the one pronunciation can be changed into the other as follows:

| | | |
|---|---|---|
| æəftənʉn | delete ə | 1 |
| æəftnʉn | delete f | 1 |
| æətnʉn | subst. t/r | 1 |
| æərnʉn | insert f | 1 |
| æfərnʉn | insert t | 1 |
| æftərnʉn | subst. ʉ/u | 1 |
| æftərnun | | |
| | | 6 |

However, the procedure which is followed here is round about. It can be done much more efficiently in this way:

| | | |
|---|---|---|
| æəftənʉn | delete ə | 1 |
| æftənʉn | insert r | 1 |
| æftərnʉn | subst. ʉ/u | 1 |
| æftərnun | | |
| | | 3 |

Both examples illustrate that is is possible to change one pronuncation to the other in many ways, often resulting in different costs. We are interested in the set of operations with the least cost that change a pronunciation $w_1$ into a pronunciation $w_2$. This is equal to the Levenshtein distance $d(w_1, w_2)$. Given that there are many different sets of operations mapping $w_1$ to $w_2$, it is not obvious how to determine the minimal set, and it is even less obvious how to determine it efficiently. The Levenshtein algorithm, however, accomplishes both these tasks.

### 5.1.3 Operation weights

Pronunciations are compared on the basis of their segments. When using the phone representation (see Section 3.1.1) the cost of substitutions and indels is set to 1, just as in the examples in Section 5.1.2. When using feature representations (see Sections 3.1.2, 3.1.3 and 3.1.4) or acoustic representations (see Sections 4.3.1, 4.3.2 and 4.3.3) the weights gradually vary. For substitutions, the weight is equal to the distance between the corresponding segments calculated according to the chosen segment representation. For indels the weight is equal to the distance between the segment to be inserted or deleted and 'silence'. This weight also depends on the segment representation chosen and is defined for each of the feature representations and acoustic representations separately. Gradual substitutions and indels may be based on both linear and logarithmic segment distances (see Section 3.7 and Section 4.7).

### 5.1.4 Allowed matches

To accord with syllabification in words, the Levenshtein distance should be based on an alignment with plausible matches. In our implementation of the algorithm the basic rule is that a vowel may normally only match with a vowel and a consonant normally only with a consonant. However, the [w] and the [j] may also match with vowels, and the [u] and the [i] may also match with consonants.

For some representations, vowels can be compared to consonants, for other representations this is impossible. Checking the different representations we see the following:

- Using the phone representation, sounds have no real definitions. If they are equal, the distance is 0, otherwise 1. The comparison of a vowel with a consonant is possible in principle (see Section 3.1.1).
- In the feature system of Hoppenbrouwers & Hoppenbrouwers (H & H) all features apply for both vowels and consonants, which basically offers the possibility of comparing a vowel with a consonant (see Section 3.1.2).
- In the feature systems of V & C and A & B it is not possible to compare a vowel with a consonant. Since we defined the [i], [u], [j] and [w] as both vowel and consonant, these sounds are exceptions (see Sections 3.1.3 and 3.1.4).
- When using an acoustic representation the comparison of a vowel with a consonant can be easily made (see Chapter 4).

For those representations where it is possible to compare a vowel with a consonant, we will not allow all vowel-consonant matches. And it is indeed not likely that a [p] will change into an [a]. On the other hand, it is not unusual that e.g. an [r] matches with an [ə]. For example two possible pronunciations for the Dutch word *vier* 'four' are [fiːr] and [fiːə]. Here we want the ending [r] and the ending [ə] to match with each other. Therefore, we allow the match of a schwa with a sonorant.

### 5.1.5 No swap operation

A phenomenon which can be found in dialect data is metathesis. For example the equivalent of 'wasp' is pronounced as [ʋɛsp] in Standard Dutch and also in the dialect of Amsterdam, and as [ʋɛps] in the dialects of Utrecht and Den Haag.[2] Using only substitutions and indels the minimum cost is found with the following alignments:

[2] The pronunciations are taken from a data set compiled by Renée van Bezooijen, University of Nijmegen, in 2000.

| | | | | |
|---|---|---|---|---|
| ʊ | ɛ | s | p | ∅ |
| ʊ | ɛ | ∅ | p | s |
| | | 1 | | 1 |

| | | | | |
|---|---|---|---|---|
| ʊ | ɛ | ∅ | s | p |
| ʊ | ɛ | p | s | ∅ |
| | | 1 | | 1 |

| | | | |
|---|---|---|---|
| ʊ | ɛ | s | p |
| ʊ | ɛ | p | s |
| | | 1 | 1 |

Actually the s and the p in the first transcription should correspond with the s and the p in the second transcription. In that way no segment distances are found. When using the phone representation it seems reasonable to assign a weight of 1 to the swap operation, the same weight as assigned to substitutions and indels. However when using gradual weights (see Section 5.1.3), the swap operation should also be weighted gradually. Some segments may be easily swapped, for example a plosive and a non-plosive (as in our example), or a vowel and a consonant, but for other segments this may be (nearly) impossible. Once these gradual weights are found, they should be scaled so that they are in the right proportion to the weights of substitutions and indels. We have not yet succeeded in finding gradual and correctly scaled weights for the swap operations. This is an interesting topic for future work.

### 5.1.6 Calculation of distance

In this section we explain the calculation of the Levenshtein distance on the basis of both pronunciations of *afternoon*. We call [æftərnun] S1 and [æəftənʉn] S2. The number of segments in S1 is $m$ and in S2 $n$. We see that $m = 8$ and $n = 8$. The Levenshtein distance calculates the minimum cost needed to change S1 into S2. For this we use a matrix `dist` of size $(m+1, n+1)$. The rows are numbered from $0 \ldots m$ and the columns from $0 \ldots n$. The cell `dist`[0,0] gets the value 0. We traverse the matrix `dist` row by row, assigning values to the other cells. We begin with row 0, and within each row, we always begin with column 0 (only in the zeroth row do we start with the first column). We call the current row number $i$ and the current column number $j$. For each cell in the matrix, we always have to look at three possibilities (to obtain a minimum):

1. *Deletion* of the i-th segment from S1. We determine *weight*($S1_i$,ø). We take the sum of this weight and the value in the cell above the current one: `dist`$[i-1, j]$. This sum is assigned to the temporary variable `upper`. This operation is only considered when $i > 0$.

2. *Substitution* of the i-th segment of S1 by the j-th segment of S2. We look up *weight*($S1_i$,$S2_j$). We take the sum of this weight and the value in the cell above and to the left of the current one: `dist`$[i-1, j-1]$. This sum is kept in the temporary variable `upperleft`. This operation is only taken into account when $i > 0$ and $j > 0$.

```
function Levenshtein_distance(S1,S2)
begin
  for i:=0 to m do begin
    for j:=0 to n do begin
      upper=upperleft=left:=maxint;

      if i>0
        then upper:=dist[i-1,j]+weight(S1[i],ø);

      if i>0 and j>0
        then upperleft:=dist[i-1,j-1]+weight(S1[i],S2[j]);

      if j>0
        then left:=dist[i,j-1]+weight(ø,S2[j]);

      dist[i,j]:=min(upper,upperleft,left);
      if dist[i,j]=maxint then dist[i,j]:=0;
    end
  end

  Levenshtein_distance:=dist[m,n];
end
```

Figure 5.1: Levenshtein algorithm in pseudo-code. The algorithm works dynamically, so that, for each $p_1$, $p_2$ prefix pair of S1, S2, it determines the least cost of operations mapping $p_1$ to $p_2$. The number of segments in S1 is $m$ and in S2 $n$.

3. *Insertion* of the j-th segment in S2. We compute $weight(\o,S2_j)$. We take the sum of this weight and the value in the cell left of the current one: `dist`$[i, j-1]$. The sum is retained in the temporary variable `left`. This operation is only considered when $j > 0$.

Now, we take the minimum of the three values, `upper`, `upperleft` and `left`, and the current cell takes it as value:

$$\texttt{dist}[i,j] \leftarrow minimum(\texttt{upper},\texttt{upperleft},\texttt{left})$$

In this way we ensure that paths arise only by adding minimally to minimal-cost cells. This guarantees that the least distance is computed. Once we have traversed the entire matrix, and computed values for all cells, then the distance – the least cost of operations mapping from S1 to S2 – is found in the cell `dist`$[m, n]$. This is the Levenshtein distance between the strings. The algorithm in pseudo-code is shown in Figure 5.1.

The matrix below shows the application of the procedure to our example. Initially `dist`[0,0] gets the value 0. In most other cells four values are given. The variables `upper`, `upperleft`, `left` are given respectively in the upper right, upper left and lower left of a cell. The minimum of these three variables is given in the lower right of a cell. Note that in the 0-th row only the variable `left` (insertions) could be calculated and in the 0-th column only the variable `upper` (deletions) could be calculated. The final distance between the two pronunciations is the lowerright value in cell `dist`[8,8]: 3. In Section 5.1.7 we explain how the cheapest path from `dist`[8,8] to `dist`[0,0] can be recovered.

| | | ∅ | | æ | | ə | | f | | t | | ə | | n | | ʉ | | n | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | | 1 | | 2 | | 3 | | 4 | | 5 | | 6 | | 7 | | 8 | |
| ∅ | 0 | | | | | | | | | | | | | | | | | | |
| | | | **0** | 1 | **1** | 2 | **2** | 3 | **3** | 4 | **4** | 5 | **5** | 6 | **6** | 7 | **7** | 8 | **8** |
| æ | 1 | | 1 | 0 | 2 | 2 | 3 | 3 | 4 | 4 | 5 | 5 | 6 | 6 | 7 | 7 | 8 | 8 | 9 |
| | | | **1** | 2 | **0** | 1 | **1** | 2 | **2** | 3 | **3** | 4 | **4** | 5 | **5** | 6 | **6** | 7 | **7** |
| f | 2 | | 2 | 2 | 1 | 1 | 2 | 1 | 3 | 3 | 4 | 4 | 5 | 5 | 6 | 6 | 7 | 7 | 8 |
| | | | **2** | 3 | **1** | 2 | **1** | 2 | **1** | 2 | **2** | 3 | **3** | 4 | **4** | 5 | **5** | 6 | **6** |
| t | 3 | | 3 | 3 | 2 | 2 | 2 | 2 | 2 | 1 | 3 | 3 | 4 | 4 | 5 | 5 | 6 | 6 | 7 |
| | | | **3** | 4 | **2** | 3 | **2** | 3 | **2** | 3 | **1** | 2 | **2** | 3 | **3** | 4 | **4** | 5 | **5** |
| ə | 4 | | 4 | 4 | 3 | 2 | 3 | 3 | 3 | 3 | 2 | 1 | 3 | 3 | 4 | 4 | 5 | 5 | 6 |
| | | | **4** | 5 | **3** | 4 | **2** | 3 | **3** | 4 | **2** | 3 | **1** | 2 | **2** | 3 | **3** | 4 | **4** |
| r | 5 | | 5 | 5 | 4 | 4 | 3 | 3 | 4 | 4 | 3 | 3 | 2 | 2 | 3 | 3 | 4 | 4 | 5 |
| | | | **5** | 6 | **4** | 5 | **3** | 4 | **3** | 4 | **3** | 4 | **2** | 3 | **2** | 3 | **3** | 4 | **4** |
| n | 6 | | 6 | 6 | 5 | 5 | 4 | 4 | 4 | 4 | 4 | 4 | 3 | 2 | 3 | 3 | 4 | 3 | 5 |
| | | | **6** | 7 | **5** | 6 | **4** | 5 | **4** | 5 | **4** | 5 | **3** | 4 | **2** | 3 | **3** | 4 | **3** |
| u | 7 | | 7 | 7 | 6 | 6 | 5 | 5 | 5 | 5 | 5 | 5 | 4 | 4 | 3 | 3 | 4 | 4 | 4 |
| | | | **7** | 8 | **6** | 7 | **5** | 6 | **5** | 6 | **5** | 6 | **4** | 5 | **3** | 4 | **3** | 4 | **4** |
| n | 8 | | 8 | 8 | 7 | 7 | 6 | 6 | 6 | 6 | 6 | 6 | 5 | 5 | 4 | 4 | 4 | 3 | 5 |
| | | | **8** | 9 | **7** | 8 | **6** | 7 | **6** | 7 | **6** | 7 | **5** | 6 | **4** | 5 | **4** | 5 | **3** |

## 5.1.7 Tracing backwards

Once the distance between S1 and S2 is computed, it is possible to find the corresponding alignment(s) which show the mapping of S1 to S2. For this purpose it is easy to set pointers when traversing the matrix for the first time. For each cell this is done as follows:

1. If the variable `upper` is equal to the minimum value, set a pointer from the current cell to the cell above.

2. If the variable `upperleft` is equal to the minimum value, set a pointer from the current cell to the cell leftabove.

3. If the variable `left` is equal to the minimum value, set a pointer from the current cell to the cell left.

When $k$ variables are equal to the minimum, there are at least $k$ paths from `dist`[0,0] to the current cell which results in the minimum cost for that (sub)sequence. The matrix below shows the pointers for our example.

| | | ∅ | æ | ə | f | t | ə | n | ʉ | n |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| ∅ | 0 | 0 | ← | ← | ← | ← | ← | ← | ← | ← |
| æ | 1 | ↑ | ↖<br>0 | ← 1 | ← | ← | ← | ← | ← | ← |
| f | 2 | ↑ | ↑ | ↖ | ↖<br>1 | ← | ← | ← | ← | ← |
| t | 3 | ↑ | ↑ | ↖ ↑ | ↖ ↑ | ↖<br>1 | ← | ← | ← | ← |
| ə | 4 | ↑ | ↑ | ↖ | ↖ ↑<br>← | ↑ | ↖<br>1 | ← | ← | ← |
| r | 5 | ↑ | ↑ | ↑ | ↖ | ↑ | ↑<br>2 | ↖ | ↖<br>← | ↖<br>← |
| n | 6 | ↑ | ↑ | ↑ | ↖ ↑ | ↖ ↑ | ↑ | ↖<br>2 | ↖<br>← | ↖ |
| u | 7 | ↑ | ↑ | ↑ | ↖ ↑ | ↖ ↑ | ↑ | ↑ | ↖<br>3 | ↖ ↑<br>← |
| n | 8 | ↑ | ↑ | ↑ | ↖ ↑ | ↖ ↑ | ↑ | ↑ | ↖ ↑ | ↖<br>3 |

The optimum alignment is found by tracing backwards. We start at `dist`[$m$,$n$] and follow along the arrows to obtain a path all the way to `dist`[0,0]. The alignment is read off from the path in reverse order. The arrows have the following meaning:

1. A *vertical* arrow in row $i$ means: delete $S1_i$ and place $\begin{bmatrix} S1_i \\ \emptyset \end{bmatrix}$ in the alignment.

2. A *diagonal* arrow in row $i$ and column $j$ means: substitute $S1_i$ by $S2_j$ and place $\begin{bmatrix} S1_i \\ S2_j \end{bmatrix}$ in the alignment.

3. A *horizontal* arrow in column $j$ means: insert $S2_j$ and place $\begin{bmatrix} \emptyset \\ S2_j \end{bmatrix}$ in the alignment.

When $k$ arrows are found in a cell, there are at least $k$ paths from `dist`[0,0] to the current cell. In our example there is only one path which gives the minimal cost. The shaded cells make up this path. This path corresponds with the following alignment:

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| æ | ə | f | t | ə | ∅ | n | ʉ | n |
| æ | ∅ | f | t | ə | r | n | u | n |
| | 1 | | | | 1 | | 1 | |

### 5.1.8 Normalization of length

When computing the distance between two sequences, in general the distance between longer sequences will be greater than the distance between shorter sequences. The longer the sequences, the greater the chance of differences between them. If we used these distances directly, then longer words would contribute disproportionally to the estimation of distances between varieties, which does not accord with the idea that words are linguistic units. Therefore, we normalize the distance by a factor that is related to the length of the sequences. Assume four different string pairs are aligned in the following way:

| | | |
|---|---|---|
| $a_1$ | $a_2$ | $a_3$ |
| $b_1$ | $b_2$ | $b_3$ |
| 1 | 1 | 1 |

(1)

| | | |
|---|---|---|
| $a_1$ | $a_2$ | $\emptyset$ |
| $b_1$ | $b_2$ | $b_3$ |
| 1 | 1 | 1 |

(2)

| | | |
|---|---|---|
| $a_1$ | $a_2$ | $\emptyset$ |
| $\emptyset$ | $\emptyset$ | $b_3$ |
| 1 | 1 | 1 |

(3)

| | | |
|---|---|---|
| $a_1$ | $a_2$ | $\emptyset$ |
| $\emptyset$ | $b_2$ | $b_3$ |
| 1 | 1 | 1 |

(4)

Since substitutions and indels have the same weight for each of the string pairs we get a cost of 3 for all string pairs. To do justice fully to the fact that all sequence operations have the same weight, in our opinion the costs of the different string pairs should also be the same after normalizing over the length of the sequences. For (1) the distance can simply be divided by the length of either of the two sequences, so the normalization factor becomes 3. This factor should also be used for the other string pairs. In (2) this factor is equal to the length of the longer sequence. In (3) this factor is equal to the sum of the lengths of both sequences. In (4) the normalization factor can be found less easily. The length of both sequences is equal to 2 (there is no longer sequence since the two sequences have the same length) and the sum of the sequences is equal to 4. Taking the mean of these two values we find the normalization factor: $(2 + 4)/2 = 3$. The different string pairs show that the way in which the normalization factor is found on the basis of the lengths of the sequences is not always the same. However, when dividing by the length of the alignment, for *all* cases the same normalization factor is found, namely 3.

As described in Section 5.1.7 the minimum cost of changing one sequence into another may correspond with more than one path in the matrix. As a hypothetical example we consider the subsequence [æft] of S1 and the subsequence [æəf] of S2. The value lowerright in cell `dist`[3,3] (see Section 5.1.6) gives the minimum cost that is needed to change one subsequence into the other: 2. Examining the

1. *bee*

| | | | | |
|---|---|---|---|---|
| German: | b | i | n | ə |
| Dutch: | b | ɛ | i | |
| | | 1 | 1 | 1 |

| | | | | | |
|---|---|---|---|---|---|
| German: | b | | i | n | ə |
| Dutch: | b | ɛ | i | | |
| | | 1 | | 1 | 1 |

2. *rabbit*

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| German: | k | a | n | i | n | ç | ə | n |
| Dutch: | k | o | n | ɛ | i | | | n |
| | | 1 | | 1 | 1 | 1 | 1 | |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| German: | k | a | n | | i | n | ç | ə | n |
| Dutch: | k | o | n | ɛ | i | n | | | |
| | | 1 | | 1 | | | 1 | 1 | 1 |

3. *kanari*

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| English: | k | ə | n | ɛ | ə | r | i |
| Frisian: | k | ə | n | ɑ | r | j | ə |
| | | | | 1 | 1 | 1 | 1 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| English: | k | ə | n | ɛ | ə | r | i | |
| Frisian: | k | ə | n | ɑ | | r | j | ə |
| | | | | 1 | 1 | | 1 | 1 |

Figure 5.2: Three word pairs with two alignments each. The longer alignment on the right is judged as the better one. Diacritics are not taken into account.

pointer matrix (see Section 5.1.7) it appears that in cell `dist`[3,3] two pointers are given, one pointing to the cell above and left, and one pointing to the cell above. The result is that there are two possible paths corresponding with the following alignments:

| | | |
|---|---|---|
| æ | f | t |
| æ | ə | f |
| | 1 | 1 |

| | | | |
|---|---|---|---|
| æ | ∅ | f | t |
| æ | ə | f | ∅ |
| | 1 | | 1 |

We judge the alignment to the right as the better one since in this alignment the two f's appear as corresponding segments. We get the impression that the longest alignment has always the greatest number of matches. Shorter and longer alignments for more pairs of different pronunciations are given in Figure 5.2.[3] Both alignments give the minimum cost. The examples confirm our conjecture. In the longer alignments more matches are found than in the shorter ones. However, is an alignment with a greater number of matches always better than an alignment with a smaller number of matches? To answer this question, consider that distances should approach human perception as close as possible. Therefore, an alignment should reflect the way in which people perceive differences between pronunciations rather than reflecting the way in which one pronunciation changed into the other in history. From this point of view the longer alignments in the examples 1, 2 and 3 are the better ones. We suppose that in perception people will try to match the common sounds in two different pronunciations, so we prefer the longer alignments.

We normalize by dividing the distance by the length of the longer alignment. This gives the average of the weights used. In our hypothetical example in which

[3] The pronunciations are taken from a data set compiled by Renée van Bezooijen, University of Nijmegen, in 2000.

two pronunciations of the word *aft* are compared, the distance is equal to 2 and the length of the longer alignment is equal to 4. The total cost of 2 is now divided by the length of 4 which gives a average weight of 0.5. When the weights represent percentages (as for the acoustic distances, see Section 4.7), dividing the distance by the length gives the average weight as a percentage. In that case the word distance is expressed as a percentage. In our example, the weights 0 and 1 may be replaced by 0% and 100%. This results in a word distance of 50%. In our example in which two pronunciations of the word *afternoon* are compared the distance is equal to 3 and the length of the alignment is equal to 9. The word distance expressed as a percentage is equal to $(3 \times 100\%)/9 = 33\%$.

### 5.1.9 Calculation of length

The normalization length is taken to be equal to the length of the alignment. In the previous section we showed that different alignments corresponding with different paths in the matrix `dist` may give the same minimum cost. The different alignments or paths may have different lengths. We explained that we prefer to divide the minimum cost by the length of the longer alignment or path. The way in which the maximum length is calculated is comparable to the way in which the minimum distance is found. The calculation of distance and length is done in the same software module. We use a matrix `length` with the same size as `dist`: $(m+1, n+1)$. The rows are numbered from $0 \dots m$ and the columns from $0 \dots n$. The matrix `length` is traversed the same way as and simultaneously with the matrix `dist`. For each cell in the matrix, we regard three possibilities:

1. If `upper` is equal to the minimum cost of the (sub)sequence we assign the value of `length`$[i-1, j]$ increased by 1 to a temporary variable `Upper`. Otherwise `Upper` becomes negative. This operation is only used when $i > 0$.

2. If `upperleft` is equal to the minimum cost of the (sub)sequence we assign the value of `length`$[i-1, j-1]$ increased by 1 to a temporary variable `UpperLeft`. Otherwise `UpperLeft` becomes negative. This operation is only possible when $i > 0$ and $j > 0$.

3. If `left` is equal to the minimum cost of the (sub)sequence we assign the value of `length`$[i, j-1]$ increased by 1 to a temporary variable `Left`. Otherwise `Left` becomes negative. This operation is only possible when $j > 0$.

Increasing the value of a previous cell by 1 represents one step in the path. In contrast to the procedure in the distance calculation we take the maximum of the three values, `Upper`, `UpperLeft` and `Left`, and the current cell takes it as value:

$$\texttt{length}[i,j] \leftarrow \mathit{maximum}(\texttt{Upper},\texttt{UpperLeft},\texttt{Left})$$

Once we have traversed the entire matrix, and computed values for all cells, then the length of the longest alignment which gives the minimum cost is found in `length`$[m, n]$.

The matrix below shows the application of the procedure to our example. In the matrix, the cells of the path which gives the minimum cost are shaded. Initially `length` gets the value 0. For each cell in the matrix the variables `Upper` (upper right in the cell), `UpperLeft` (upper left) and `Left` (lower left) are given. However, they are only given when their corresponding counterparts (`upper`, `upperleft` and `left`) are equal to the minimum cost (compare also the pointer matrix in Section 5.1.7). Otherwise a negative value is assigned which is not given in the matrix. The maximum of these three variables is given in the lower right of a cell. The final length is the lower right value of cell `length`[8,8]: 9.

| | | ∅ | æ | ə | f | t | ə | n | ʉ | n |
|---|---|---|---|---|---|---|---|---|---|---|
| | | 0 | 1 | 2 | 3 | 4 | 5 | 6 | 7 | 8 |
| ∅ | 0 | **0** | 1 **1** | 2 **2** | 3 **3** | 4 **4** | 5 **5** | 6 **6** | 7 **7** | 8 **8** |
| æ | 1 | 1<br>**1** | 1<br>**1** | 2 **2** | 3 **3** | 4 **4** | 5 **5** | 6 **6** | 7 **7** | 8 **8** |
| f | 2 | 2<br>**2** | 2<br>**2** | 2<br>**2** | 3<br>**3** | 4 **4** | 5 **5** | 6 **6** | 7 **7** | 8 **8** |
| t | 3 | 3<br>**3** | 3<br>**3** | 3 3<br>**3** | 3 4<br>**4** | 4<br>**4** | 5 **5** | 6 **6** | 7 **7** | 8 **8** |
| ə | 4 | 4<br>**4** | 4<br>**4** | 4<br>**4** | 4 5<br>5 **5** | 5<br>**5** | 5<br>**5** | 6 **6** | 7 **7** | 8 **8** |
| r | 5 | 5<br>**5** | 5<br>**5** | 5<br>**5** | 5<br>**5** | 6<br>**6** | 6<br>**6** | 6<br>**6** | 7<br>7 **7** | 8<br>9 **9** |
| n | 6 | 6<br>**6** | 6<br>**6** | 6<br>**6** | 6 6<br>**6** | 6 7<br>**7** | 7<br>**7** | 7<br>**7** | 7<br>8 **8** | 8<br>**8** |
| u | 7 | 7<br>**7** | 7<br>**7** | 7<br>**7** | 7 7<br>**7** | 7 8<br>**8** | 8<br>**8** | 8<br>**8** | 8<br>**8** | 9 9<br>9 **9** |
| n | 8 | 8<br>**8** | 8<br>**8** | 8<br>**8** | 8 8<br>**8** | 8 9<br>**9** | 9<br>**9** | 9<br>**9** | 9 9<br>**9** | 9<br>**9** |

## 5.1.10 Dialect distances

Once we are able to calculate the distance between two pronunciations of a word, we can also find the distance between two varieties. When for $k$ words the pronunciations are given for both variaties, we get $k$ word pairs. For each pair we calculate the Levenshtein distance. In this way we get $k$ Levenshtein distances. Now the distance between the two varieties is equal to the average of the $k$ Levenshtein distances. When the word distances represent percentages, the distance between both varieties represents a percentage as well. Because in our research varieties are mostly dialects, the distance between two varieties is referred to as *dialect distance*. When having $n$ dialects, for each possible pair of dialects the average Levenshtein distance can be calculated. The corresponding distances can be arranged in a $n \times n$ matrix.

#### 5.1.10.1 Missing transcriptions

When comparing two varieties on the basis of $k$ word pairs, it may happen for one or more of the pairs, for one or both varieties, that no pronunciation is given. Since we work with average distances, we simply discount the effect of missing transcriptions. This has the same effect as estimating the word distance to be average word distance. The use of average distance has the advantage that it allows us to examine distances between $n$ dialects for each of the $k$ words individually, even if for some varieties pronunciations are missing.

#### 5.1.10.2 Multiple transcriptions

Sometimes several transcriptions are given for the same word in one variety. Assume in variety 1 the equivalent for 'house' is pronounced as [hys] and [hus]. In variety 2 the same two transcriptions are found. The two varieties are equal in the sense that both [hys] and [hus] are possible. When simply calculating the mean of all possible pairs ([hys] vs. [hys], [hys] vs. [hus], [hus] vs. [hys] and [hus] vs. [hys]) we incorrectly get a word distance which is higher than 0. A better approach would be to calculate the mean of all plausible pairs, i.e. pairs with elements that probably correspond to each other. In that case we get the mean distance of the pairs [hys] vs. [hys] and [hus] vs. [hus] which is equal to 0. In this section we propose a procedure that is based on the idea that the mean word distance should be based on the most natural word pairs.

Our current implementation is able to deal with at most ten different pronunciations per word per variety. We illustrate the way in which we process them by an example. Assume word W1 in dialect 1 has the following transcriptions:

$$a\ b\ c$$

and word W2 in dialect 2 has the following transcriptions:

$$x\ y$$

Dialect 1 has three transcriptions, and dialect 2 has two transcriptions. We duplicate each of the three transcriptions in dialect 1 two times, and each of the two transcriptions in dialect 2 three times. For dialect 1 we get:

$$a\ a\ b\ b\ c\ c$$

and for dialect 2 we get:

$$x\ x\ x\ y\ y\ y$$

We see that the number of variants of W1 and W2 is the same, namely 6. We want to find the 6 most likely pairs of variants. This is done by a heuristics. For finding 6 pairs we perform 6 iterations. Within each iteration we find the pair of variants (one of dialect 1 and one of dialect 2) with the the smallest distance. The members of the pair may not already be used in previous formed pairs. The final distance between W1 and W2 is the sum of the word distances corresponding with the 6 pairs divided by 6.

The procedure may be described in more general terms. Assume for word W1 $m$ transcriptions are given and for word W2 $n$ transcriptions. Each of the $m$ transcriptions is duplicated $n$ times, and each of the $n$ transcriptions is duplicated $m$ times. In this way for both W1 and W2 we get $m \times n$ variants (which are not all unique). The variants of W1 are indexed as $1 \leq p \leq m \times n$, and the variants of W2 as $1 \leq q \leq m \times n$. Now we have to find the $m \times n$ most natural word pairs. The way in which these are found is described by the pseudo-code in Figure 5.3.

The algorithm starts with assigning the value 0 to a variable `sum`. Next $m \times n$ iterations are performed. Within each iteration we search for the word pair $(p, q)$ which has the smallest distance. This distance is added to `sum`. The final word distance is equal to the average of the distances corresponding with the formed word pairs, which is equal to `sum`$/(m \times n)$.[4]

## 5.2 Levenshtein distance using acoustic word samples

Once we are able to find distances between sound samples, it is not a big step to extend the methodology so that the distances between word samples can be found as well. We do not use the acoustic representations of separate segments, but complete acoustic representations of whole words. Transcriptions are only used to find the number of segments of words. This number is used to normalize the speech rate. Kruskal and Liberman (1999), Hunt et al. (1999) and Ten Bosch

[4] An alternative approach of processing multiple variants is given by Nerbonne and Kleiweg (2003).

```
function word_distance(W1,W2);
begin
  sum:=0;

  repeat
    for all possible word pairs (p,q) do
      if (Levenshtein_distance(W1(p),W2(q))<smallest) and
           p and q not used in previous formed pairs
        then smallest:=Levenshtein_distance(W1(p),W2(q))
        else {nothing};
    end;

    sum:=sum+smallest;
  until m*n word pairs are found

  word_distance:=sum/(m*n);
end
```

Figure 5.3: Algorithm in pseudo-code for finding the most natural word pairs. The algorithm assumes that the $m$ pronunciations of word W1 are multiplied $n$ times and the $n$ pronunciations of word W2 are multiplied $m$ times before the algorithm is called.

(2000) present methods with which pronunciations are compared on the basis of the acoustic signal.

Kruskal and Liberman (1999) describe the development of continuous time-warping and 'formulate discrete analogues to all concepts and definitions involved' (p. 127). A continuous function in multidimensional space is called a *trajectory*. For trajectories it holds that 'variation in speed appears concretely as compression and expansion with respect to the time axis' (p. 125). Among other things time-warping makes it possible to 'measure how different two sequences are in a way that is not sensitive to compression-expansion but is sensitive to other differences' (p. 125). Kruskal and Liberman 'formalize the notion of a time-warping as a "linking" that connects the time scales of the two trajectories or sequences' (p. 129). The distance between two trajectories is defined 'as the minimum possible length of any linking between them' (p. 129). The chief application of time-warping has been in speech processing which makes the methodology interesting for dialect comparison as well.

Hunt et al. (1999) present a syllable-based speech recognition system in which unknown syllables are acoustically recognized by matching them against stored syllable templates. Syllables are represented as a sequence of acoustic-parameter vectors, each vector corresponding to one time frame. A Levenshtein algorithm finds the optimum frame-to-frame correspondence between the template syllable and the unknown syllable and calculates the distances between them over that optimum frame correspondence.

Ten Bosch (2000) describes research in which an Automatic Speech Recognition (ASR) based distance measure is used to find the acoustic distances between dialects. Words are represented as a series of frames where each frame contains acoustic features. Words are compared by aligning the frames by a Viterbi alignment procedure, a technique roughly comparable to how phonetic segments are aligned when using transcriptions. Alignment is done by matching the frames with trained ASR Hidden Markov Models (HMMs). More about the Viterbi algorithm and HMMs can be found in Manning and Schütze (1999).

The advantage of comparing words directly on the basis of acoustic samples is that no transcriptions need to be made. It is time consuming to make phonetic transcriptions and, furthermore, the quality of the transcriptions varies greatly, depending on the skills of the transcriber. In this section we present the methodology for the comparison of word samples (almost) without the use of transcriptions.

In Section 5.2.1 we describe some necessary manipulations that should be applied to the samples first. When comparing sounds, we examine several representations of the acoustic signal. The same representations are used here and discussed in Section 5.2.2. In Section 5.2.3 we explain how we normalize different speech rates. In Section 5.2.4 we describe how distances between word samples are actually found using the Levenshtein distance. These sections contain material published in Heeringa and Gooskens (2003).

### 5.2.1 Preprocessing

The voices of different speakers will have different pitches. Most obvious is the difference in pitch between male and female voices. Furthermore, the intonation per speaker may vary individually, in a way unindicative of variety. When two speakers read the same text aloud, the one may stress different words than the other. To make samples of different speakers as comparable as possible, all word samples are monotonized, i.e. manipulated to have the same pitch for all times in the sample. When there are male speakers and female speakers, we found the mean pitch of the men and the mean pitch of the women first. Next, all samples were monotonized on the average of the two means with the program PRAAT. Figure 5.4 shows spectrograms of non-manipulated word samples, while Figure 5.5 shows spectrograms of the corresponding monotonized word samples.

Just as for the sound samples, the volume was not normalized because volume contains a good deal of sound specific information. For example it is specific for the [v] that its volume is greater than that of the [f].

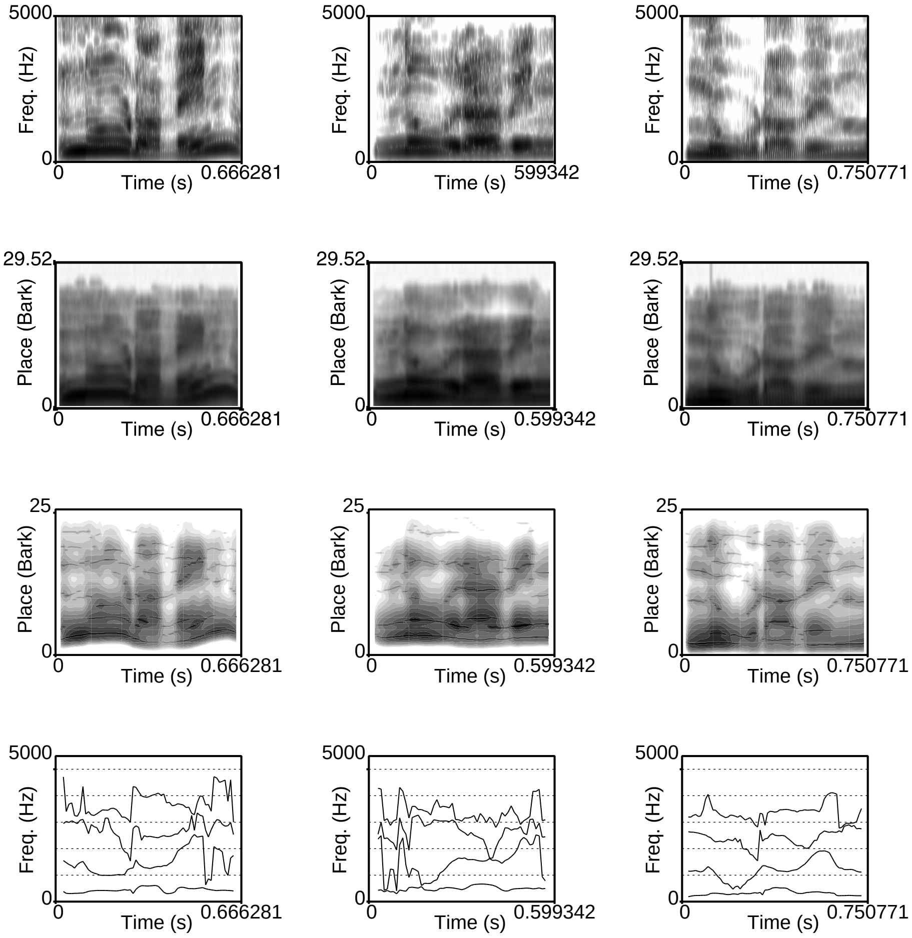
### 5.2.2 Representation of words

When comparing words, we do not use the type of spectrogram most commonly used which has a Hertz-scale, but the more perceptual models which we also used for the comparison of segments. The Barkfilter is described in Section 4.3.1 and the cochleagram is described in Section 4.3.2. Formant tracks represent the prominent frequency tracks in the spectrogram. In this more reduced representation speaker-specific information may be filtered away to a greater extent. The formant track representation is described in Section 4.3.3. When using these representations for the comparison of words, mostly the same parameter values are used as when comparing segments. Parameters which deserve particular attention are discussed below.

Just as in the comparison of segments forward-masking is not taken into account when using the cochleagram representation (see Section 4.3.2.1). The same word pronounced at different speech rates gives different sample sizes. When applying forward-masking (in PRAAT a default value of 0.03 seconds is given), the effect on smaller samples is relatively larger than on larger samples. This shows that forward-masking depends on speech rate. Because we suppose that speech rate is speaker-dependent, we want to reduce its influence as much as possible in the comparison of dialects. Therefore, we do not apply forward-masking.

When using the formant tracks in the comparison of segments, only two tracks are used to get results comparable to the IPA vowel quadrilateral, which also reflects F1 and F2 only. For word comparison this restriction need not to be maintained since no comparison with the IPA vowel quadrilateral will be made. The number of formants may vary over time in a word and per word. In the PRAAT program, we maintain the default value for the maximum number of formants which may be found: 5. Next, we find the minimum number of formants examining all times of all words which are taken into consideration. After that, on the basis of this minimum number of formants the word samples are compared. In the samples we use (see Section 7.2.3) for each word sample at each time sample, at least three formants could be found. Therefore, the comparison of word samples here is based on (the first) three formant tracks. Furthermore, in the PRAAT program the ceiling of the formant search should be set to 5000 Hz for males, and to 5500 Hz for females. Because the samples on the basis of which the formants are determined were monotonized to the average of the mean pitch of the males and the females (see Section 5.2.1), we set this ceiling to 5250 Hz.

To illustrate the differences between the several representations when applied to word samples, we show visualizations of three Norwegian pronunciations of the word *nordavinden* 'the northwind' using spectrograms, Barkfilters, cochleagrams and formant tracks (Figures 5.4 and 5.5). The pronunciations of the dialects of Bjugn, Halden and Larvik are given. The recordings were made by Jørn Almberg (see Section 7.2 for more details about the recordings). The pictures in Figure 5.4

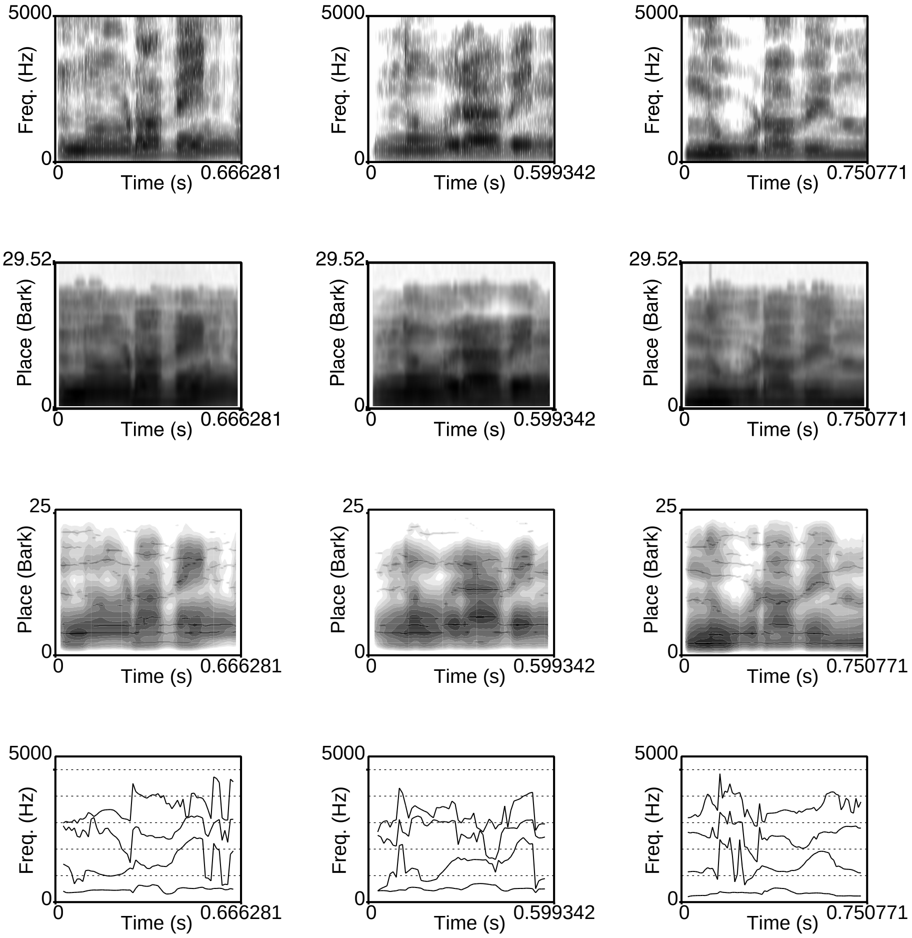


Bjugn [$^2$nuːɽɑˌʋiɲˑ]

Halden [$^2$nuːɾɑˌʋinː]

Larvik [$^2$nuɽɑˌʋinˑn̩]

Figure 5.4: Different acoustic representations of three Norwegian pronunciations of *nordavinden* ‘the northwind’. Starting from the first row we see respectively spectrograms, Barkfilters, cochleagrams and formant tracks obtained on the basis of the *original* samples.



Bjugn
[$^2$nuːɽɑˌʋiɲˑ]

Halden
[$^2$nuːɾɑˌʋinː]

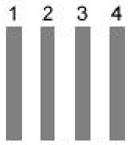
Larvik
[$^2$nuɽɑˌʋinˑn̩]

Figure 5.5: Different acoustic representations of three Norwegian pronunciations of *nordavinden* ‘the northwind’. Starting from the first row we see respectively spectrograms, Barkfilters, cochleagrams and formant tracks obtained on the basis of the *monotonized* samples.

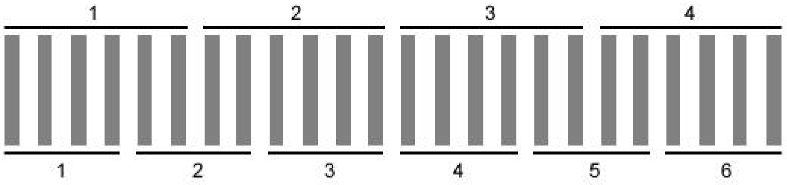
are made on the basis of the original (not manipulated) samples, and those in Figure 5.5 on the basis of the monotonized samples. The monotonized samples are used for dialect comparison.

### 5.2.3 Speech rate

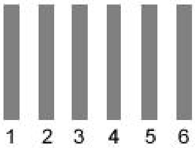
When comparing word samples, we have to allow for the fact that different speech rates give different sample sizes. We perform a rough normalization by using the number of segments per word according to the phonetic transcription. Assume that the acoustic representation of a word sample consists of $l$ spectra or formant bundles. If the number of segments of this word pronunciation according to the phonetic transcription is $m$, and we want to represent each segment by $n$ spectra or formant bundles, then we represent the complete word sample by $m \times n$ elements. Changing the representation of $l$ elements into a representation of $m \times n$ elements is realized in two steps. First we duplicate each of the $l$ elements $m \times n$ times. This gives $l \times m \times n$ elements in total. Second we regard the $l \times m \times n$ elements as $m \times n$ groups, each consisting of $l$ elements, and fuse the elements in each group to one element by averaging them. The result is a representation of $m \times n$ elements. We illustrate this by an example. Assume we have a word sample of $l = 4$ elements:



If this word pronuncation is transcribed as a sequence of $m = 2$ segments, and we want to represent each segment by $n = 3$ elements, then we represent the complete word sample by $2 \times 3 = 6$ elements. We change the representation of 4 elements into a representation of 6 elements. For this purpose first we duplicate each of the 4 elements 6 times. This gives 24 elements in total:



Second we treat the 24 elements as 6 groups, each consisting of 4 elements, and fuse the elements in each group to one element by averaging them. The result is a representation of 6 elements:



In our research we chose $n = 20$, i.e. 20 spectra or formant bundles per segment. A higher value gives nearly the same results, but the computing time increases greatly. We are aware of the fact that our way of normalizing speech rate is a rough approach, but we hypothesize that it is refined enough to capture significant variation.

### 5.2.4 Comparison of words

The Levenshtein distance calculates the cost of changing one sequence or string into another. It determines how the one sequence or string can be changed into the other in the easiest way by inserting, deleting or substituting elements. A detailed description of the algorithm was given in Section 5.1. When finding the distance between different pronunciations on the basis of their transcriptions, the elements are the phonetic segments. However, when using the acoustic signal, the elements are spectra or formant bundles.

The cost of a substitution of spectra or formant bundles is the vector (Euclidean) distance between them. Assume a spectrum or formant bundle $e1$ and $e2$ with respectively $n$ frequencies or formants, then:

$$d(e1, e2) = \sqrt{\sum_{i=1}^{n}(e1_i - e2_i)^2} \tag{5.1}$$

For the calculation of insertions and deletions we used definitions of ‘silence’. We defined a ‘silence spectrum’ as a spectrum for which the intensities of all frequencies are equal to 0. A ‘silence formant bundle’ is defined as a bundle for which all frequencies are equal to 0. This means that in absolute silence there are no vibrations.

As alternative, ‘silence’ can be defined as a spectrum or formant bundle which is sampled from background noise. At first sight this may seem to be a better approach because background noise is found in all the recordings. However in the recordings which we used (see Section 7.2.3) the background noise differs by dialect recording. Since the background noise was very low for each recording, we used no sampled ‘silence’, but ‘silence’ as defined above. Our definition of ‘silence’ approximates real ‘silence’ very closely without favoring one particular recording.

When the algorithm has calculated the sum of the operations, this sum is divided by the length of the corresponding longest alignment. The longest alignment inherently has the greatest number of matches.

When comparing two varieties on the basis of $k$ word pairs, it may appear that for one or more of the pairs for one or both varieties, no translation is given. In that case, the distance for that word pair is ignored, which is equivalent to taking the average of the distances of all word pairs for which translations in both dialects are available. Finally, in the implementation we only use one pronunciation per word.

# Chapter 6

# Analysing dialect distances

As mentioned in Chapter 5 we may use Levenshtein distance to find the distance between two pronunciations of the same word. The distance between two varieties is equal to the average of a sample of Levenshtein distances of corresponding word pairs. When we have $n$ varieties, then the average Levenshtein distance is calculated for each possible pair of varieties. For $n$ dialects $n \times n$ pairs can be formed. The corresponding distances are arranged in a $n \times n$ matrix which is comparable to distance tables published by auto clubs and often found in pocket calendars that show the distances between the main towns.

The distance at each variety with respect to itself is found in the distance matrix on the diagonal from upperleft to lowerright. These values are always zero and therefore give no real information, so that only $n \times (n-1)$ pairs are interesting. Furthermore, the Levenshtein distance is symmetric. This means that the distance between word 1 and word 2 is equal to the distance between word 2 and word 1. The result is that distance between variety 1 and variety 2 is equal to the distance between variety 2 and variety 1 as well. Therefore, the distance matrix is symmetric. We need to use only one half which contains the distances of $(n \times (n-1))/2$ word pairs.

To interpret the $(n \times (n-1))/2$ varieties, they can be visualized on a map. On the map each pair of points is connected by a line. Darker lines correspond to similar language varieties, lighter lines to more distant varieties. Very distant relations result in lines too faint to be seen (in the interest of overall contrast). An example of such a map can be found in Figure 9.4. The map shows distances between 360 Dutch varieties and is discussed in Section 9.2. On the map, dialect groups can already be distinguished to some extent. This way of visualizing is related to the beam maps of Séguy and Goebl (see Section 2.3.1). However, on the beam maps only neighboring points are connected while on our map all points are in principle connected showing all $(n \times (n-1))/2$ distances.

Another way of interpreting the distances is to examine the results of classification methods which are applied to the distances. Classification results show relations between elements in a way which is easy to understand. We used cluster analysis and multidimensional scaling, two common techniques that complement each other. The result of cluster analysis is a dendrogram, a tree where the varieties are the leaves. The technique is described in Section 6.1. The result of multidimensional scaling is a map, where the distance between kindred varieties is small, and between different dialects great. This technique is explained in Section 6.2.

## 6.1 Cluster analysis

Jain and Dubes (1988, p. 55) define cluster analysis as 'the process of classifying objects into subsets that have meaning in the context of a particular problem.' The goal of clustering is to identify the main groups in complex data. In this section, we discuss a set of cluster methods that are referred to as SAHN (Sequential, Agglomerative, Hierarchical, Nonoverlapping) clustering methods by Sneath and Sokal (1973). Sequential means that the objects are processed one by one instead of simultaneously. Agglomerative procedures starts with placing each object in its own cluster and gradually merges smaller clusters in larger clusters until all objects are in one single cluster. A hierarchical classification is a nested sequence of partitions. Nonoverlapping means that for every split in the hierarchy each object belongs to exactly one cluster. SAHN clustering methods are suitable for classification of language varieties because they show both groupings and distances. The distances are reflected to some extent by the hierarchical structure.

### 6.1.1 Johnson's algorithm

The general scheme used for SAHN clustering is called Johnson's algorithm. Jain and Dubes (1988) mention that the scheme was suggested by King (1967) and formalized by Johnson (1967). We will demonstrate the algorithm by an example. Assume we get the following matrix, which shows the linguistic distances between some Dutch dialects[1]:

| | Grouw | Haarlem | Delft | Hattem | Lochem |
|---|---|---|---|---|---|
| Grouw | | 42 | 44 | 46 | 47 |
| Haarlem | | | 16 | 36 | 38 |
| Delft | | | | 38 | 40 |
| Hattem | | | | | 21 |
| Lochem | | | | | |

[1] In Chapter 9 linguistic distances between 360 Dutch dialects are calculated. Our small $5 \times 5$ table is a subtable of the large $360 \times 360$ table.

The value of each cell $(i, i)$ is of course equal to 0 (the distance of a variety with respect to itself). Because the matrix is symmetric we do not need the distances in the left lower half.

Clustering with Johnson's algorithm is an 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. To keep the example simple we calculate the distance from $k$ to a newly formed cluster $[ij]$ as the mean of the distance between $i$ and $k$ and the distance between $j$ and $k$. So for each $k$ we calculate:

$$d_{k[ij]} = \frac{d_{ki} + d_{kj}}{2}$$

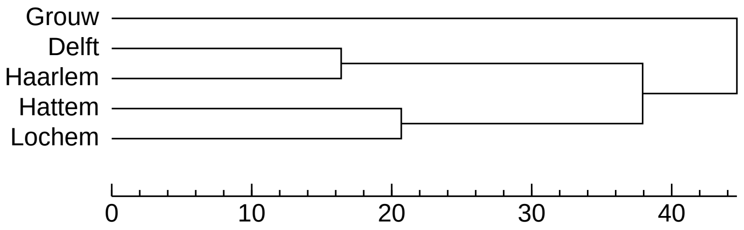
In the distance matrix the shortest distance is found between Haarlem and Delft. Both Haarlem and Delft are removed from the matrix, and a new cluster Haarlem & Delft is inserted. To iterate, we have to assign a distance from the newly formed cluster to all other points. For example, the distance between Grouw and Haarlem & Delft is calculated as follows:

$$\begin{aligned} d_{Grouw,\ [Haarlem\ \&\ Delft]} &= \frac{d_{Grouw,\ Haarlem} + d_{Grouw,\ Delft}}{2} \\ &= \frac{42\ +\ 44}{2} \\ &= 43 \end{aligned}$$

After calculating the distances between Hattem and Haarlem & Delft and between Lochem and Haarlem & Delft as well, we get the following matrix (new values are in bold type):

| | Grouw | Haarlem & Delft | Hattem | Lochem |
|---|---|---|---|---|
| Grouw | | **43** | 46 | 47 |
| Haarlem & Delft | | | **37** | **39** |
| Hattem | | | | 21 |
| Lochem | | | | |

In each iteration the matrix is reduced in size. The iterations are repeated until no elements are left which can be fused to a new cluster. The final result is a complete hierarchical grouping of varieties. This grouping is visualized as a dendrogram, a tree in which the leaves are the varieties and the lengths of the branches correspond with the distances. In our example we get the following dendrogram:



In Figure 6.1 Johnson's algorithm is given in pseudo-code. On the basis of $n$ elements $n-1$ clusters are obtained. The elements are numbered from 1 to $n$, and the clusters from $n+1$ to $n+n-1$. Therefore the variable $k$ that gives the cluster index is set to the number of elements initially. The input of the procedure is $DistanceMatrix$ which contains the distances. The output is $Cluster$, containing for each of the $n-1$ clusters its subclusters and the distance between both subclusters. On the basis of $Cluster$ the dendrogram is constructed.

```
procedure cluster(DistanceMatrix,Cluster);
begin
  k:=number of elements;

  while elements or clusters are left that can be fused do begin
    k:=k+1;

    find pair (i,j) in DistanceMatrix that has smallest distance;
    store subclusters i and j in Cluster[k];
    distance between subclusters of Cluster[k]:=distance between i and j;

    delete rows and columns of i and j in DistanceMatrix;
    insert a row and a column of cluster k in the DistanceMatrix;
    calculate distances from cluster k to all remaining points;
  end;
end
```

Figure 6.1: Johnson's algorithm in pseudo-code.

### 6.1.2 Matrix updating algorithms

Each time two clusters are fused to a new cluster, the distances from the newly formed cluster to all other points (or clusters) need to be calculated. In our example, the distance from a new cluster $ij$ to point $k$ was calculated as the mean of the distance between $i$ and $k$ and the distance between $j$ and $k$. The way in which the distances between a newly formed cluster and the remaining points is calculated is called a *matrix updating algorithm.* Sneath and Sokal

(1973, pp. 218–219) mention six matrix updating algorithms. Jain and Dubes (1988, p. 80) mention the same updating algorithms and added a seventh, Ward's method.

Assume points (or clusters) $i$ and $j$ are fused to one cluster $ij$. Then for calculating the distance from cluster $ij$ to a point (or cluster) $k$ the following data are (partly) needed: $n_i$ (number of varieties in cluster $i$), $n_j$ (number of varieties in cluster $j$), $n_k$ (number of varieties in cluster $k$), $d_{ij}$ (distance between $i$ and $j$), $d_{ki}$ (distance between $k$ and $i$) and $d_{kj}$ (distance between $k$ and $j$). Now the seven matrix updating algorithms are defined as follows:

1. Single-link (nearest neighbor):

$$d_{k[ij]} = minimum(d_{ki}, d_{kj})$$

2. Complete-link (furthest neighbor):

$$d_{k[ij]} = maximum(d_{ki}, d_{kj})$$

3. Unweighted Pair Group Method using Arithmetic averages (UPGMA):

$$\begin{array}{llllllll} d_{k[ij]} = & (n_i & / & (n_i + n_j)) & \times & d_{ki} & + \\ & (n_j & / & (n_i + n_j)) & \times & d_{kj} & \end{array}$$

4. Weighted Pair Group Method using Arithmetic averages (WPGMA):

$$d_{k[ij]} = (\tfrac{1}{2} \times d_{ki}) + (\tfrac{1}{2} \times d_{kj})$$

5. Unweighted Pair Group Method using Centroids (UPGMC):

$$\begin{array}{lrllllll} d_{k[ij]} = & (n_i & / & (n_i + n_j)) & \times & d_{ki} & + \\ & (n_j & / & (n_i + n_j)) & \times & d_{kj} & - \\ & ((n_i \times n_j) & / & (n_i + n_j)^2) & \times & d_{ij} & \end{array}$$

6. Weighted Pair Group Method using Centroids (WPGMC):

$$d_{k[ij]} = (\tfrac{1}{2} \times d_{ki}) + (\tfrac{1}{2} \times d_{kj}) - (\tfrac{1}{4} \times d_{ij})$$

7. Ward's method (minimum variance):

$$\begin{array}{lrllllll} d_{k[ij]} = & ((n_k + n_i) & / & (n_k + n_i + n_j)) & \times & d_{ki} & + \\ & ((n_k + n_j) & / & (n_k + n_i + n_j)) & \times & d_{kj} & - \\ & (n_k & / & (n_k + n_i + n_j)) & \times & d_{ij} & \end{array}$$

When finding the distance between a new cluster and an existing cluster, single-link finds the closest pair of elements in the two clusters and complete-link the most distant pair. UPGMA and WPGMA assess the dissimilarity between the new cluster and the existing cluster by the distance between the means. Instead of using means UPGMC and WPGMC use centroids, i.e. the hypothetical points at the centers of clusters $i$, $j$ and $k$. Ward's method assigns a new distance in the way that results in the smallest increase in the within-cluster sum of squares, i.e., the sum of the squared distances between each point and the resultant cluster centroid (Wilks, 1995).

Note the use of the terms *unweighted* and *weighted. Weighted clustering* was introduced by Sokal and Michener (1958). In this approach clusters that merge get equal weights regardless of the number of elements in each cluster. In that case elements in small clusters are weighted relatively more heavily than elements in larger clusters. In *Unweighted clustering* each element in a cluster gets equal weight, regardless of the number of elements in that cluster (Sneath and Sokal, 1973, p. 228). Although this terminology was adopted by Jain and Dubes (1988) and others, it may be confusing since in unweighted clustering we weight clusters the merge by their size, and in weighted clustering we do not.

### 6.1.3 Experimentation

In our research, we have to decide which matrix updating algorithm should be used. Because both single-link and complete-link take only one cluster into account when merging two clusters, we did not use them. Weighting clusters by their size when fusing them seems more reasonable to us than weighting them equally, so we prefer unweighted clustering. Using the centroid-based methods, we sometimes get results in which the distance between two clusters is smaller than between the subclusters in (one of) the two clusters. When comparing dialects such results are not natural. So only two matrix updating algorithm are left: UPGMA and Ward's method.

Wilks (1995) indicates that Ward's method tends to create clusters of equal size, which is not always reasonable. In our research we found that varieties which appear as outliers in a UPGMA dendrogram are neatly ordered under a group of moderate size in a dendrogram obtained by Ward's method. We will compare on the basis of the following matrix:

$$\begin{array}{c|cccc} & \mathrm{a} & \mathrm{b} & \mathrm{c} & \mathrm{d} \\ \hline \mathrm{a} & & 1 & 2 & 2 \\ \mathrm{b} & & & 2 & 2 \\ \mathrm{c} & & & & ? \\ \mathrm{d} & & & & \end{array} \tag{6.1}$$

We applied both UPGMA and Ward's method to the matrix, and experimented with different values for the '?', the distance between objects c and d. We found that for the values 0 through 1.9 (with a step size of 0.1) similar dendrograms are obtained. Setting the distance between objects c and d to 2.0, dendrograms are obtained which show that objects c and d are equally distant to the cluster containing objects a and b. However, in the dendrogram generated by the Ward's method objects c and d form a cluster. For values varying from 2.1 through 2.3 (with a step size of 0.1 again), in the UPGMA dendrogram object d is further apart from objects a and b than object c. In the dendrogram generated by Ward's method object c and d still form one cluster, which is unexpected and counterintuitive. For values 2.4 and higher the two methods will give similar results. Results for the values 1.0, 2.0, 2.2 and 2.4 are found in Figure 6.2.

A useful quantitative method for validating cluster results is developed by Sokal and Rohlf (1962). They proposed to calculate the *cophenetic correlation coefficient*, which is a measure of the agreement between the distances as implied by the dendrogram and those of the original distance matrix. This approach is also described in Sneath and Sokal (1973, pp. 277–284) and Jain and Dubes (1988, pp. 66–68 and 166–170). The *cophenetic correlation coefficient*, abbreviated as CPCC by Farris (1969), measures the correlation between the original distances and the cophenetic distances. Because dialect distances are numeric data, we used the Pearson correlation coefficient (see Section 3.6.2.5). Cophenetic values are the distances as suggested by the dendrogram. For finding the cophenetic distance between objects $i$ and $j$ we have to find the least significant (smallest) cluster in which both objects are first present. The cophenetic distance between $i$ and $j$ is equal to the distance between the subclusters of this cluster. Once we have a correlation coefficient, we can calculate to what extent the cophenetic distances explain the variance in the original distances. The variance is found by taking the square of the cophenetic correlation coefficient. The variance is expressed as a percentage when multiplied by 100. In this thesis this variance is given as a percentage for most dendrograms.

Using the CPCC we examined the difference between UPGMA and Ward's method further. For the '?' in the matrix (the distance between c and d) the values 0.0 through 3.0 are filled in with a step size of 0.1, and for each value the CPCC is calculated for both UPGMA and Ward's method. In Figure 6.3 the CPCC is plotted against *distance*(c,d). From 0.0 through 2.0 we see that the CPCC of the UPGMA is equal to 1.00. For values higher than 2.0 the CPCC decreases. For Ward's method the CPCC increases from 0.0 through 1.0, decreases from 1.0 through 2.1, increases from 2.1 through 2.3, and increases from 2.3. UPGMA and Ward's method are equal for *distance*(c,d)=1 where they have both a CPCC of 1.00. The greatest difference between UPGMA and Ward' method is found for *distance*(c,d)=2.0, where UPGMA has CPCC=1.00 and Ward's method CPCC=0.9439. From 2.1 through 2.3 Ward's method gives counterintuitive results, suggesting that c and d form a group which is not justified. For values 2.4

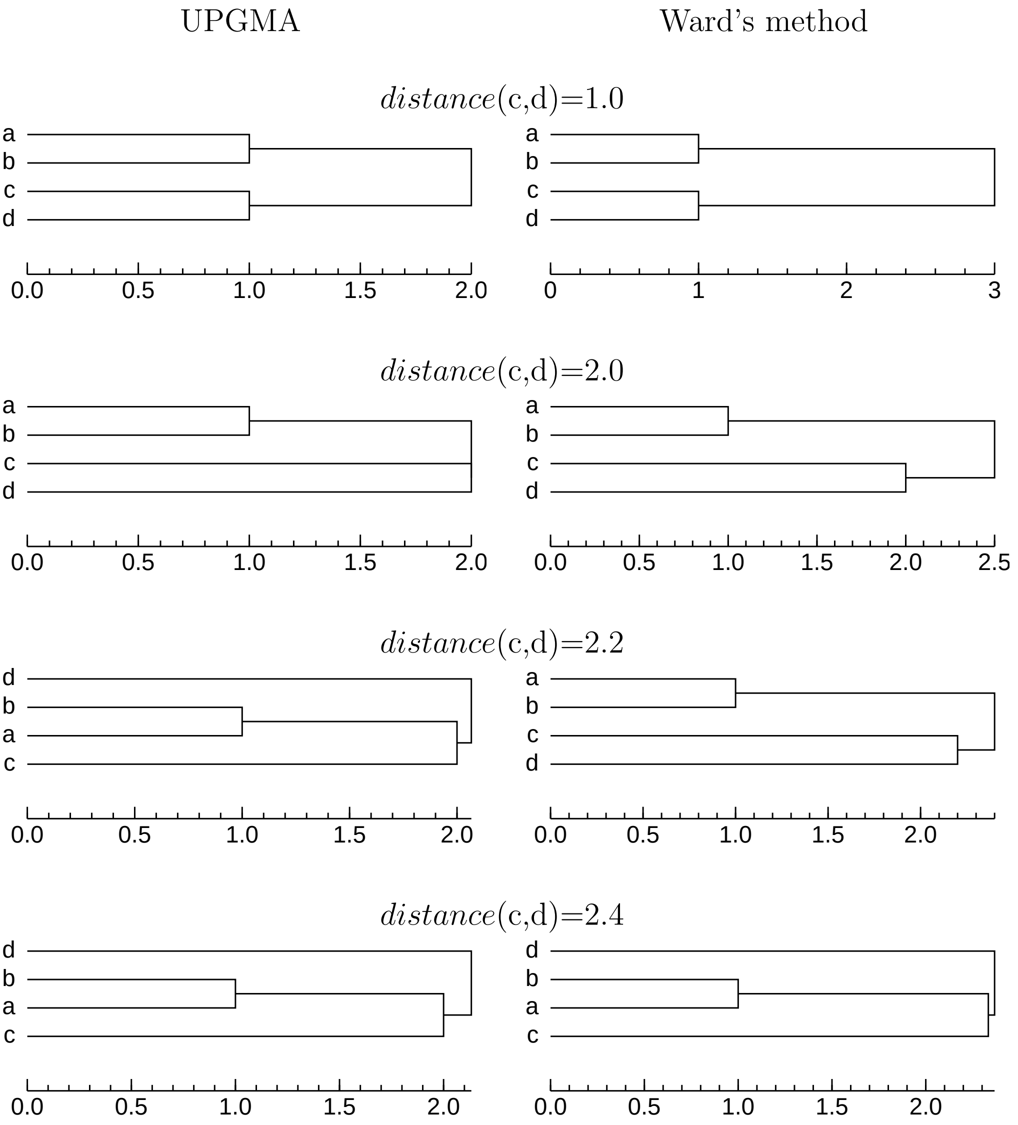


Figure 6.2: Results obtained from the matrix given in (6.1) using UPGMA and the Ward's method. In the matrix different values have been used as distance between objects c and d. When the distance between c and d is equal to 2.2, in the UPGMA dendrogram object d is further from objects a and b than object c. In the dendrogram generated by Ward's method object c and d form one cluster, which is perhaps counterintuitive with regard to seeking dialectological groups.

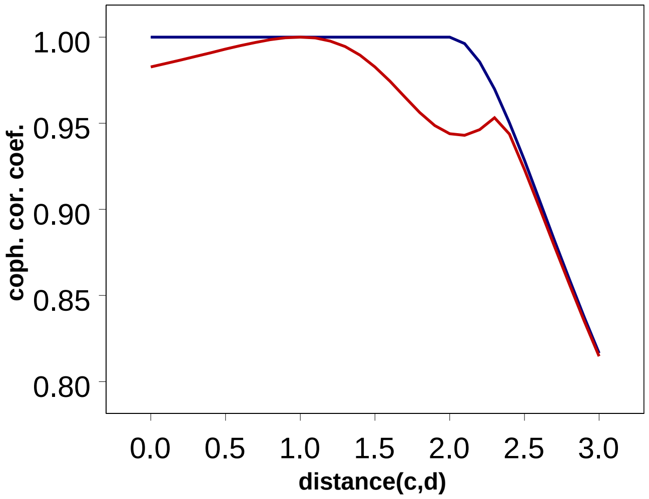


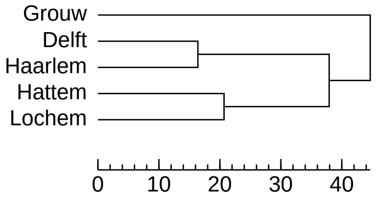
Figure 6.3: Comparison of UPGMA (upper line) and Ward's method (lower line) for the matrix given in (6.1) for 0 ≤ *distance*(c,d) ≤ 3.0 with a step size of 0.1. To obtain a clear graph, points are connected by lines. For *distance*(c,d)=1.0 both methods have CPCC=1.00. Greatest difference was found for *distance*(c,d)=2.0, where UPGMA has CPCC=1.00 and Ward's method CPCC=0.9439. From 2.1 through 2.3 Ward's method gives counterintuitive results which is reflected in this graph by lower CPCC's with respect to the CPCC's of UPGMA.

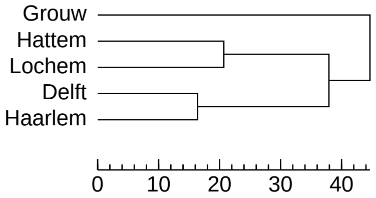
and higher both UPGMA and Ward's method run parallel, where for both the CPCC decreases. The UPGMA still performs better than the Ward's method. Only when *distance*(c,d)=16, both methods have the same CPCC when using four decimals.

The graph shows that the cophenetic correlation coefficient gives results which accord with our findings at the beginning of this section. On the basis of this we tend to judge UPGMA preferable to Ward's method. Therefore, we will use UPGMA troughout the rest of this thesis. We are aware that more situations are possible which should be tested. However, it is beyond the scope of this thesis to find and discuss them all.

### 6.1.4 Ordering of clusters

In Section 6.1.1 we show a dendrogram on the basis of five varieties. Below the dendrogram on the left is the same dendrogram as shown in Section 6.1.1, and the dendrogram on the right is an alternative. In the right one the clusters Delft & Haarlem and Hattem & Lochem are swapped. In the dendrogram Delft and Haarlem can be swapped in the same way as Hattem and Lochem. Thus the same clustering can be visualized by different dendrograms. The question arises which of them is better. Examining the two dendrograms below we prefer the left one since the distance between Grouw and the cluster Delft & Haarlem ($(42 + 44)/2 = 43$, see the distance matrices in Section 6.1.1) is smaller than the distance between Grouw and the cluster Hattem & Lochem ($(46+47)/2 = 46.5$). In our implementation of the graphic display of the cluster algorithm the branches are ordered so that the more related varieties or clusters are located near each other in the dendrogram.





Assume we have a clustering that contains two clusters. The first cluster contains subclusters a and b (but a and b may also be leaves), and the second cluster contains subclusters c and d (c and d may again be leaves). Figure 6.4 shows that the clustering can be visualized in four different ways. So we have to decide which one is better. For this purpose we examine the distance between a, b, c and d and mirror one or both subclusters if necessary. Assume the subcluster which contains a and b is called $i$, the subcluster containing c and d is called $j$ and the (sub)cluster containing $i$ and $j$ is called $C$. Now the procedure is as follows:

```
if minimum(b,c,C)
  then {nothing}
  else if minimum(a,c,C)
         then mirror subgroup i
         else if minimum(b,d,C)
                then mirror subgroup j
                else if minimum(a,d,C)
                       then mirror subgroup i and mirror subgroup j
```

The function *minimum* checks whether the distance for the pair of given elements is smaller than for each other possible pair of elements in (sub)cluster $C$. The result of this procedure is a dendrogram in which closest clusters (or terminals) are located near each other. This procedure was applied to the dendrograms in this thesis.

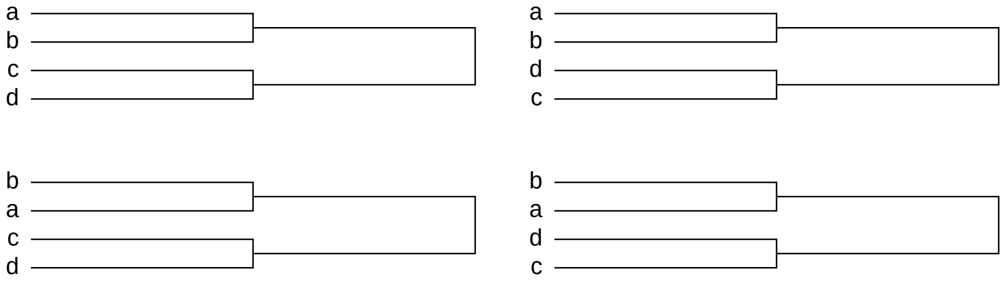


Figure 6.4: Four dendrograms, each of which is generated on the basis of the same distances using the same cluster method. We prefer the top left visualization in cases where $d(b, c) < d(a, c)$, $d(b, c) < d(b, d)$ and $d(b, c) < d(a, d)$.

### 6.1.5 Application

In Section 6.1.3 we mentioned that we use the UPGMA clustering method throughout this thesis. Examples of dendrograms can be found in Figures 8.3 and 9.5. The dendrogram in Figure 8.3 is based on distances between 55 Norwegian varieties, and the dendrogram in Figure 9.5 is based on 360 Dutch varieties. The two dendrograms are discussed in respectively Section 8.2.1 and Section 9.3.1.

For a dendrogram with $n$ varieties the $k$ most significant groups can be found, where $1 \leq k \leq n - 1$. The choice of a suitable value for $k$ depends on the number of varieties. When each group of the partition gets a unique color, the groups can be identified on a map. On such a map the most important groups in the dendrogram can easily be found. When neighboring points belong to different groups, the exact border between the points is found on the basis of triangulation. With this technique the two points are blown up to small areas until they touch each other (see Section 6.2.4). Figures 8.4 and 9.6 show maps based on the dendrograms in respectively Figure 8.3 and Figure 9.5. The two maps are discussed in respectively Section 8.2.2 and 9.3.2.

The groups as found in a dendrogram can also be represented geographically by a *composite cluster map*.[2] On this type of map groups are separated by borders which are represented by lines. Darker lines separate distant groups, lighter lines more similar groups. When creating this map, in the first step the border between the two most significant groups is drawn. In the second step, two borders are drawn which separate the three most significant groups. The first border, which was already drawn in the previous step, is drawn again, resulting in a darker color. The second border is drawn for the first time, so it will be lighter than the other one. In the $i$-th step, the $i - 1$ borders of the $i$ most

[2] Composite cluster maps were introduced by Peter Kleiweg, see also `http://www.let.rug.nl/~kleiweg/ccmap/`.

significant groups are drawn again (they get darker), and a new border is added. If the cluster contains $n$ varieties, we start with drawing borders which separate the 2 most signficant groups, and end with drawing borders which separate the $n$ most significant groups. Figure 9.7 shows a composite cluster map based on the dendrogram in Figure 9.5. The map is discussed in Section 9.3.3.

Comparing the color area map with the composite cluster map, the benefit of the composite cluster map is that the weigth of borders between groups is visualized. On the other hand, composite cluster maps have the disadvantage that they cannot show that varieties which are geographically separated by varieties of other groups, belong to the same group. In the color area map varieties of the same group get simply the same color.

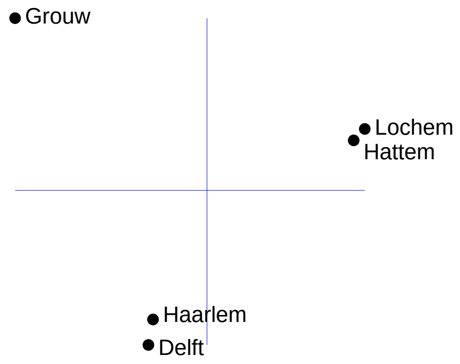
## 6.2 Multidimensional scaling

On the basis of geographic coordinates, the distances between locations can be determined. The reverse is also possible: on the basis of the known distances, an optimal coordinate system can be determined with the coordinates of the locations in it. The latter is realized by a technique known as 'multidimensional scaling' (MDS). In a multidimensional scaling plot, strongly related dialects are close to each other, while strongly different dialects are located far away from each other. MDS has its origins in psychometrics. Different persons are judged as similar if they tend to give similar responses to the same stimuli. MDS helps to understand the results of similar experiments (Oh and Raftery, 2001). Togerson (1952) proposed the first MDS method and coined the term.

### 6.2.1 Basic idea

The purpose of multidimensional scaling (MDS) is to provide a visual representation of the pattern of distances among a set of elements. On the basis of distances between a set of elements a set of points is returned so that the distances between the points are approximately equal to the original distances. The result is that on the plot like concepts are plotted nearby and unlike concepts are distant.

In Section 6.1 we use a small distance matrix which contains the distances between five varieties. On the basis of these distances with multidimensional scaling the varieties are plotted on a map, where the distances between the elements reflect the original distances as close as possible. This gives the following result:



In the original distance matrix small distances were found between Haarlem and Delft and between Hattem and Lochem. In the MDS plot, Haarlem and Delft, and Hattem and Lochem appear as two close clusters. The original distance matrix shows large distances between Grouw and the four other dialects, between Haarlem and Hattem, Haarlem and Lochem, Delft and Hattem and Delft and Lochem. These large distances are clearly reflected in the MDS plot. The x-axis represents the first dimension, and the y-axis the second dimension. If required the axes may be swapped. MDS values may also be used inversely. Both swapping axes and using values inversely are allowed since they do no change the distances between the elements on the plot.

### 6.2.2 Algorithms

Having three elements a, b and c, it is not difficult to place them in two-dimensional space so that the distances between them are correctly rendered. First a and b are placed with the right distance between them, and next c is placed so that is has the right distance with respect to both a and b. However, when adding a fourth element d, it is more difficult and it may be impossible to locate it so that the distances with respect to a, b and c are reflected perfectly. MDS assigns coordinates so that the Euclidean distances between the assigned points reflect the original distances as closely as possible. Normally, MDS is used to scale to two dimensions since three or more dimensions are difficult to display on paper. However, MDS can also be used to scale to three or more dimensions. In our research we used MDS routines as implemented in the statistical R package.[3] The program provides three MDS procedures: Classical Multidimensional Scaling, Kruskal's Non-metric Multidimensional Scaling and Sammon's Non-Linear Mapping.

As mentioned above, Togerson (1952) proposed the first MDS method which is known as Classical Multidimensional Scaling. The method is also described in

[3] The program R is a free public domain program and available via `http://www.r-project.org/`.

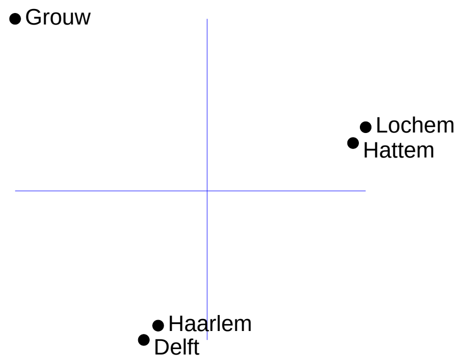
Togerson (1958) and is a *metric* procedure. The MDS plot in the example above is obtained on the basis of this procedure.

Both Kruskal's Non-metric Multidimensional Scaling and Sammon's Non-Linear Mapping are *non-metric* procedures, i.e. the ranks of the distances are used. Kruskal's Non-metric Multidimensional Scaling was the first non-metric multidimensional scaling procedure. In Shepard (1962), Kruskal (1964) and Kruskal and Wish (1978) the procedure is explained. Sammon's Non-Linear Mapping is described by Sammon (1969). In both procedures the MDS coordinates are found by an iterative algorithm. The algorithm starts with an initial configuration. Usually random values are assigned to coordinates of each of the elements. In R, however, the initial coordinates are found with Classical Multidimensional Scaling. Next, a range of steps is repeated until the optimal coordinates are found. First, the Euclidean distances between the elements on the basis of their coordinates are calculated. These distances are compared to the original distances using a *stress* function. The smaller the stress value, the closer the correspondence. The function is discussed below. Next, the coordinates are adjusted to reduce the stress. The most optimal coordinates are found when the stress can drop no further.

As mentioned above the degree of correspondence between the Euclidean distances between the MDS coordinates and the original distances is measured by a *stress* function. Assume $d_{ij}$ is the original distance between elements $i$ and $j$, and $D_{ij}$ is the Euclidean distance between elements $i$ and $j$ as found on the basis of the coordinates. When using Kruskal's Non-metric Multidimensional Scaling the stress is calculated with the following formula:

$$STRESS = \sqrt{\frac{\sum_{i<j}(f(d_{ij}) - D_{ij})^2}{\sum_{i<j} {D_{ij}}^2}} \tag{6.2}$$

In this formula, $f(d_{ij})$ is a weakly monotonic transformation of the original distances. The function maps the original distances to values that best preserve the rank order. The transformation is found via *monotonic regression* (Jain and Dubes, 1988, pp. 49–50). Monotone regression is a step-function which is constrained to always increase from left to right. First, a monotonic transformation of the original distances is performed. Next, linear regression is applied to these transformed original distances and the coordinate-based distances. Subsequently, on the basis of the regression formula, original distances are predicted on the basis of the coordinate-based distances. In the formula, such a predicted value is noted as $f(d_{ij})$. Using monotone regression, the correlation between $f(d_{ij})$ and $D_{ij}$ will be maximized. Monotone regression is also known as *isotonic regression.* Therefore, in R, Kruskal's Non-metric Multidimensional Scaling is also known as isoMDS. The denominator of the fraction is a constant scaling factor that assures that stress values are between 0 and 1. Applying Kruskal's Non-metric Multidimensional Scaling to the distances of the five dialects in our example we get the following plot:

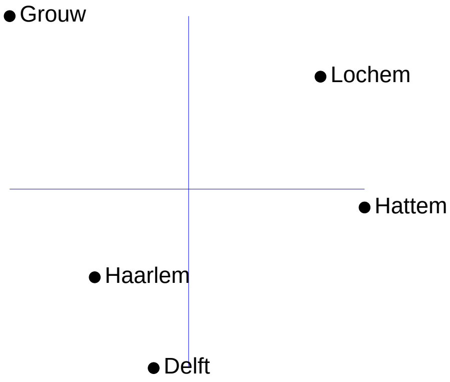


The x-axis represents the first dimension, and the y-axis the second dimension. Compared to the plot obtained on the basis of Classical Multidimensional Scaling the distance between Haarlem and Delft has become a bit smaller, and the distance between Hattem and Lochem has become a bit greater. Otherwise the two plots are very similar.

When using Sammon's Non-Linear Mapping another stress function is used. The formula is:

$$STRESS = \frac{\displaystyle\sum_{i<j} \frac{(d_{ij} - D_{ij})^2}{d_{ij}}}{\displaystyle\sum_{i<j} d_{ij}} \tag{6.3}$$

The main difference with stress in Kruskal's Non-metric Multidimensional Scaling is that the squared differences between the original distances and the coordinate-based distances are weighted by the original distances. Because of this normalization the preservation of small distances will be emphasized. The whole sum in the numerator is divided by the sum of the original distances in the denominator in order to scale the stress to a value between 0 and 1. The following plot show the result of Sammon's Non-Linear Mapping when applied to the distances of the five varieties in our example:



The x-axis represents the first dimension, and the y-axis the second dimension. When comparing the plot to the plots obtained by Classical Multidimensional Scaling and Kruskal's Non-metric Multidimensional Scaling we see that the distance between Haarlem and Delft and between Hattem and Lochem has become relatively much larger, although still two clusters can be recognized. The plot shows clearly that small distances are emphasized and greater distances are weakened. Using a larger set of points it appears that groups which are sharply distinguished on plots obtained by Classical Multidimensional Scaling and Kruskal's Non-metric Multidimensional Scaling form a continuum on a plot obtained by Sammon's Non-Linear Mapping.

### 6.2.3 Experimentation

R provides us with stress values only for Kruskal's Non-metric Multidimensional Scaling and Sammon's Non-Linear Mapping. However, stress values calculated with different formulas are not comparable. In order to compare the results of the different MDS procedures, we need a measure of fitness which can applied to each MDS result. This was found in ALSCAL, an MDS program using the alternating least square algorithm. A description of the algorithm is given by Takane et al. (1977) and Norušis (1997).[4] In ALSCAL the squared Pearson's correlation coefficient is calculated between the original distances and the Euclidean MDS coordinate-based distances. A higher correlation coefficient indicates that the multidimensional scaling values are a good representation of the original distances. The square of this correlation coefficient is equal to the variance of the original distances as explained by the chosen number of dimensions. We extended the R procedure so that Pearson's correlation coefficient $r$ between the original distances and the final Euclidean distances on the plot is calculated by default. On the basis of this correlation coefficient the $r^2$ value was calculated and given for each plot in this thesis.

[4] The ALSCAL program is a free public domain program and available via: `http://forrest.psych.unc.edu/research/alscal.html`. ALSCAL is also included in the statistical package SPSS.

The $r^2$ value may help us to decide which of the MDS procedures available in R should be used. Calculating the $r$ and $r^2$ values for the three MDS plots of our example, we get the following outcomes:

| Method | $r$ | $r^2$ |
|---|---|---|
| Classical Multidimensional Scaling | 0.989 | 97.7% |
| Kruskal's Non-metric Multidimensional Scaling | 0.990 | 98.0% |
| Sammon's Non-Linear Mapping | 0.935 | 87.4% |

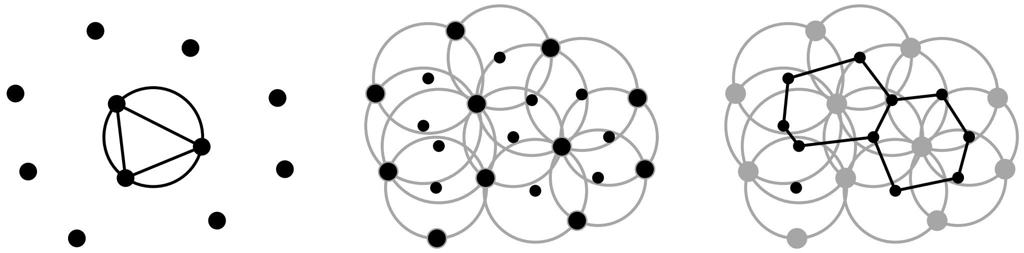
We used a three-digit precision to distinguish the values from each other. All correlation coefficients are significant, but none is significantly higher than the others.[5] Because the data set of our example contains only five varieties, no firm conclusions can be drawn. However, when applying the procedures to other data sets, in general Kruskal's Non-metric Multidimensional Scaling gets the highest $r$ and $r^2$ value. Therefore, we will use this procedure throughout this thesis. Note that the non-metric Kruskal MDS performs better than the metric classical MDS as well although all distances we measured (between segments or between dialects) are metric data. We cannot explain this. Sammon's Non-Linear Mapping sometimes outperforms Classical Multidimensional Scaling, but the opposite was also observed many times.

### 6.2.4 Application

In Section 6.2.3 we mentioned that we use Kruskal's Non-metric Multidimensional throughout this thesis. An example of a three-dimensional multidimensional scaling plot can be found in Figure 8.5. The plot is based on distances between 55 Norwegian varieties. The plot is discussed in Section 8.3.1. On the basis of three dimensions of a three-dimensional solution, each variety can be represented by a color. If we let the first dimension be the intensity of red, the second the intensity of green and the third the intensity of blue, each variety gets an unique color. This approach can be used to create a dialect map. Each dialect point gets a color according to its MDS values. Colors can be assigned to the MDS dimensions so that a color scheme is obtained that is as similar as possible to existing dialect maps. Space between points can be colored in two ways. In the first approach dialect points are blown up to small areas. The areas are found by using the *Delaunay triangulation* (Krämer, 1995). Triangles connect points so that the circumcircle (circle that passes through all three points) does not contain any other point (see left picture below). For each circle that connects the three points of a triangle the center can be found (see below the picture in the middle). The centers of circles corresponding with adjacent triangles are connected. In this

[5] For finding significances we used the Mantel test which is explained in Section 3.8.2. As significance level we choose $\alpha = 0.05$.

way a pattern of polygons arises known as *Voronoi polygons*, *Thiessen polygons* or *Dirichlet tessellation* (see right picture below). The same technique for finding polygons is also used by Goebl (1982) (see also Goebl (1993) and Figure 2.1 and 2.2).



Sometimes MDS plots clearly show that varieties are language islands in the continuum. In that case we do not derive a Voronoi cell from these varieties. On the map they are marked with a diamond, where only the diamond is colored on the basis of the MDS values. Varieties which fit in the continuum are marked with a black dot. An example of such a map is found in Figure 9.32. The map is based on MDS values of a three-dimensional solution which was obtained on the basis of distances between 360 Dutch varieties. The map is discussed in Section 9.5.2.

In the second approach space between points is colored by interpolation. Assume point $a$ is yellow and point $b$ is blue. When no other points are located between points $a$ and $b$, an unknown point exactly in the middle of both points will be green. An unknown point closer to point $a$ will be more yellow and a point closer to point $b$ will be more blue. In our research we used the most simple interpolation procedure, which is known as *inverse distance weighting.* Assume we have a map with $n$ points. Each point has geographic coordinates $(x_i, y_i)$ and MDS coordinates $(R_i, G_i, B_i)$ where $1 \leq i \leq n$. Now for an intermediate point with geographic coordinates $(x_p, y_p)$ we want to calculate MDS coordinates $(R_p, G_p, B_p)$ where $1 \leq p \leq m$ and $m$ is the number of intermediate points. These points form a regular grid over the area. Obviously, $m$ determines the density. A higher $m$ will result in a map on which colors more gradually change. $R_p$ is found as follows:

$$R_p = \frac{\sum\limits_{i=1}^{n} R_i \times \dfrac{1}{\delta(i,p)^s}}{\sum\limits_{i=1}^{n} \dfrac{1}{\delta(i,p)^s}} \tag{6.4}$$

where $\delta(i,p) = (x_i - x_p)^2 + (y_i - y_p)^2$ and $s = 2$.

Coordinates $G_p$ and $B_p$ are found in an analogous way. When $s = 0.5$, then $\delta(i,p)$ is just the Euclidean distance between $i$ and $p$. Higher values for $s$ give higher color contrasts. In our research we used $s = 2$ which results in color contrasts which are strong enough to be seen on the one hand, and realistic on the other hand.

Just as when applying triangulation dialect islands are excluded from interpolation. As they would be in triangulation they are marked with a diamond on the map, where only the diamond is colored on the basis of the MDS values. Varieties which fit in the continuum are marked with a black dot. The color of the space immediately around this dot will nearly reflect the color of the variety itself. Examples of this type of map are given in Figures 8.6 and 9.33. Figure 8.6 is based on the MDS values which are represented by the multidimensional scaling plot in Figure 8.5. The map is discussed in Section 8.3.2. Figure 9.33 is based on the same MDS values as the map in Figure 9.32 and described in Section 9.5.2.

Triangulation takes less computation time than interpolation. When the network of data points has a high density, interpolation may hardly be needed. On the other hand, interpolation does justice to the idea that the dialect landscape may be regarded as a continuum.

# Chapter 7

# Validating Norwegian dialect distances

From the previous chapters it is clear that a great number of alternative methods is available for comparing dialects. Many of the alternatives are refinements of one another, leading to the question which methods are most suitable in general. In this chapter validation work is reported, which gives an answer to this question.

Section 7.1 starts with an overview of the alternative methods we validate in this chapter. The methods will be validated on the basis of the Norwegian NOS data. This data source is described in Section 7.2. Since measurements are valid only if they are reliable, the reliability of the measurements which are obtained by the word-based methods (frequency method and Levenshtein distance) is checked. The reliability or consistency checking is explained in Section 7.3. Subsequently all methods are validated in Section 7.4. The results of the dialectometric methods are compared to perceptual distances, as found on the basis of a perception experiment. On the basis of reliability checking and validation work we find the optimal comparison method in Section 7.5. We apply this method to the NOS data and show results.

## 7.1 Overview of methods

In this chapter we validate the different methods with which distances between varieties can be calculated. We examine dialect distance measurements varying several dimensions:

- *Comparison method*
  We examined the corpus frequency method (see Section 2.3.2), the frequency per word method (see Section 2.3.3) and the Levenshtein distance (see Sections 2.3.4, 5.1 and 5.2). The advantage of the frequency per word method compared to the corpus frequency method is that words are re-

garded as linguistic units, and the Levenshtein distance improves on the frequency per word method in that the order of segments in a word is taken into account.

- *Data source*
  All comparison methods are applied to phonetic transcriptions. However, the Levenshtein distance was applied not only to transcriptions (see Section 5.1) but to acoustic word samples as well (see Section 5.2).

- *Transcription segment representation*
  When using transcriptions, speech segments may be represented as phones in the simplest case (see Section 3.1.1). In more refined methods segments are represented by features or acoustically. We examined the feature systems of Hoppenbrouwers & Hoppenbrouwers (H & H, see Section 3.1.2), Vieregge & Cucchiarini (V & C, see Section 3.1.3) and Almeida & Braun (A & B, see Section 3.1.4). These discrete representations are used in combination with the corpus frequency method, the frequency per word method and the Levenshtein distance. To obtain good acoustic representations of canonical segments we examined the Barkfilter (see Section 4.3.1), the cochleagram (see Section 4.3.2) and formant track representations (see Section 4.3.3). Acoustic representations are only used in combination with the Levenshtein distance.

- *Acoustic word representation*
  Above we used acoustic samples of individual segments. In the following section we consider whole word recordings. When using the Levenshtein distance based on acoustic word samples, we experimented with three representations, namely the Barkfilter, the cochleagram and the formant track representation (see Section 5.2.2).

- *Number of length gradations*
  For all variants of comparison methods there is a distinction made between extra-short and non-extra-short sounds which we implemented by changing the transcriptions and weighting non-extra-short sounds at least two times as heavily as extra-short sounds. For the processing of half-long and long we examined two approaches. In the first approach half-long and long are processed by changing feature values or simply ignored when using phone or acoustic segment representations. In this case only *two* degrees of length are represented by weighting segments, namely extra-short and non-extra-short. In the second approach half-long and long are processed by weighting half-long segments three times and long segments four times as heavily as an extra-short sound. In this case *four* degrees of length are represented by weighing segments, namely extra-short, short, half-long and long (see Sections 3.4.2 and 4.6.1).

- *Representation of diphthongs*
  When using transcriptions a diphthong may be processed as the sequence of two segments or as one segment which has a gradual changing color (see Section 3.2 and Section 4.4.

- *Comparison of feature histograms or feature bundles*
  In feature-based measures the distance between feature histograms (corpus frequency method and frequency per word method) or feature bundles (Levenshtein distance) can be determined via Manhattan distance, Euclidean distance, or via a measure based on Pearson's $r$ (see Section 3.6.2.5).

- *Scaling of segment distances*
  Discrete and acoustic segment distances can be used in two different ways in combination with Levenshtein distance. First they can be used unchanged, i.e. linearly. Alternatively, the logarithms of the distances can be used (see Section 3.7).

Although not all of the eight dimensions combine with one another, we nonetheless examine 187 combinations, of which three apply only to acoustic material. The variety reinforces the need for validation techniques.

## 7.2 Data source

In contrast to many European countries, in Norway dialects are used by people of all ages and social backgrounds both in the private domain and in official contexts (Omdal, 1995). When making recordings the risk is minimal that speakers use a standardized version of their dialect or a variety which is no longer used in every day life. It does not feel unnatural for Norwegian people to read a text aloud in their own dialect.

In the period 1999–2002 Jørn Almberg and Kristian Skarbø (Department of Linguistics, University of Trondheim) compiled a database which consists of recordings of about 50 Norwegian dialects.[1] As a basis the text of the fable 'The North Wind and the Sun' was used. This text was also used in IPA (1949) and IPA (1999) where the text has been transcribed in a large number of different languages. Besides recordings, corresponding transcriptions are also given.

In Gooskens and Heeringa (2004) a perception experiment is described which is based on these recordings (see Section 7.2.1). At the time this experiment was carried out (see Section 7.4.1) recordings of a set of 15 varieties were available. Therefore, the perception experiment was based on 15 varieties. We used the results of this experiment for validation work. In our research the transcriptions of the words in the texts of the same 15 varieties were used as input for a set

[1] The recordings and transcriptions are free available via: `http://www.ling.hf.ntnu.no/nos/`.

of 184 transcription-based comparison methods, which are variants of either the corpus frequency method, the frequency per word method and the Levenshtein distance. Samples of the words in the recordings of these 15 varieties were used as input for a set of 3 recording-based comparison methods, which are variants of the Levenshtein distance.

### 7.2.1 Text recordings

In order to get recordings of translations in different Norwegian dialects of this text, speakers were asked to read the text aloud. The speakers were all given the text in Norwegian beforehand and were allowed time to prepare themselves to be able to read the text aloud in their own dialect. The choice of words and the order of words are sometimes changed to get an authentic rendition. When reading the text aloud speakers were asked to imagine that they were reading the text to someone with the same dialectal background. This was done to ensure a natural reading style and to achieve dialectal correctness.

A set of 15 recordings were used in a perception experiment in order to find perceptual distances among the corresponding 15 varieties. The recordings were made in a soundproof studio in the autumn of 1999 and the spring of 2000. The microphone used for the recordings was a MILAB LSR-1000 and the recordings were made in DAT format using a FOSTEX D-10 Digital Master Recorder. They were edited by means of Cool Edit 96. The perception experiment is explained more extensively in Section 7.4.1.

### 7.2.2 Word transcriptions

On the basis of the recordings transcriptions were made by Jørn Almberg. The transcriptions were made in IPA as well in X-SAMPA (eXtension of Speech Assessment Methods Phonetic Alphabet). In X-SAMPA the IPA symbols are mapped to the ASCII/ANSI characters as found on the keyboard and available in even the most primitive text editors, which makes computational processing of the transcriptions much easier. The big advantage of this data set is that all transcriptions are made by the same person which ensures maximal consistency. The Norwegian translation of the fable 'The North Wind and the Sun' consists of 58 different words. The words are listed in Appendix B Table B.1. The 184 transcription-based computational comparison methods which we validate in this chapter are applied to the transcriptions of translations of these 58 words. For the purpose of validation the same 15 varieties are used as for the perception experiment. Afterwards a larger set of 55 varieties was also used.

Due to the free translation of some phrases a few of the expected words were missing in certain varieties. When two varieties are compared, and when one of the 58 words is missing in a translation of one variety or in both varieties, the word is not taken into account in the calculation of the distance (see Section 5.1.10.1).

Some words occur more than once in the text; e.g., *nordavinden* 'the northwind' normally appears four times in the text. In these cases the mean distance over the variants of one word is used for calculating the distance. (see Section 5.1.10.2). From Section 7.3 it becomes clear that the 58 words are a sufficient basis for reliable dialect comparison.

In Norwegian dialect areas, intonation is one of the most important characteristics. Minimal word pairs can be distinguished by means of tonemes. In the transcriptions three types of tonemes can be found: toneme 1 and toneme 2 (Kristoffersen, 2000) and circumflex (Almberg, 2001). From literature we know that the realization of the same tonemes can vary considerably across the Norwegian dialects. However, no information was given about the precise realization of the tonemes in the transcriptions. We return to this issue below in Section 7.4.1.

### 7.2.3 Word samples

As mentioned in Section 7.2.2 the Norwegian translation of the fable 'The North Wind and the Sun' consists of 58 different words. For all 15 dialects each of the 58 words were cut from the text, so we usually get 58 word samples per dialect. The 3 recording-based computational comparison methods which we validate in this chapter are applied to the word samples which are selected from the recordings.

The same quantifications we note above in Section 7.2.2 about missing elements in recordings apply here as well. Some words occur more than once in the text. In recording-based comparison only the first occurence is selected since the selection of word samples is rather time-consuming.

The voices of different speakers have different pitches. Most obvious is the difference in pitch between male and female voices. Furthermore, the intonation may vary per speaker. When two speakers read the same text aloud, the one may stress different words than the other. To make samples of different speakers as comparable as possible, all word samples were monotonized (see Section 5.2.1). In the set of 15 varieties, 4 recordings were recorded by men, and 11 by women (see Section 7.2.4). The mean pitch of the 4 men was 134 Hz, and of the 11 women 224 Hz. The mean of the means is 179 Hz. So all word samples were monotonized on the mean of 179 Hz. We are aware of the fact that this choice removes all prosodic information about pitch and intonation contours, and that these are known to be significant dialect markers in Norwegian. However, we found no way to exclude speaker-dependent intonation and simultaneously retain dialect-dependent intonation. Furthermore, we are aware of the fact that monotonizing does not remove all gender-dependent information.

### 7.2.4 Varieties

In Figure 7.1 the geographical distribution of the 15 varieties is shown. The dialects are spread over a large part of the Norwegian language area, and cover most major dialect areas as found on the traditional map of Skjekkeland (1997, p. 276). On this map the Norwegian language area is divided in nine dialect areas. In our set of 15 varieties six areas are represented. Figure 7.2 shows which dialect areas the 15 varieties belong to according to the map of Skjekkeland (1997).

For both the perception experiment and the recording-based comparison methods, the distinction between males and females is important. In the set of 15 dialects, the varieties of Bodø, Bø, Herø and Larvik are recorded by male speakers, the other varieties by female speakers.

## 7.3 Consistency

A measure can only be valid when it is reliable. But it may be reliable without being valid. Since reliability is a necessary condition for validity, we check the reliability of the set of methods which calculate distances as the averages of separate word distances. It concerns a total of 147 methods which are variants of the frequency per word method and the Levenshtein distance. The consistency is measured by calculating Cronbach's $\alpha$. In Section 7.3.1 an explanation of this measure of reliability is given. In Section 7.3.2 results are discussed.

### 7.3.1 Cronbach's $\alpha$

Cronbach's $\alpha$ is a popular method to measure consistency or reliability. Cronbach (1951) proposed the coefficient $\alpha$ as a lower bound to the reliability coefficient in classical test theory. Cronbach's $\alpha$ is not a statistical test, it is a coefficient of consistency.

Using a word-based method, the distances between varieties are obtained per word. When calculating the distances between $n_v$ varieties on the basis of $n_w$ words, $n_w$ matrices are obtained, each containing the distances between the $n_v$ varieties on the basis of the pronunciations of one word. Because the distance of a variety with respect to itself is always 0, these distances from the matrices' diagonals are not considered. Since distances between two word pronunciations are symmetric, only the half of the matrix is used. In a matrix totally $(n_v \times (n_v - 1))/2$ distances are taken into account. For each pair of matrices the correlation coefficient can be calculated. The average inter-correlation $\bar{r}$ among the words is calculated as:

$$\bar{r} = \frac{\displaystyle\sum_{i=2}^{n_w}\sum_{j=1}^{i-1} r(w_i, w_j)}{\frac{n_w \times (n_w - 1)}{2}} \tag{7.1}$$

where $r(w_i, w_j)$ is Pearson's correlation coefficient between the matrices of words $w_i$ and $w_j$. Cronbach's $\alpha$ can be written as a function of the number of words and the average inter-correlation among the words:

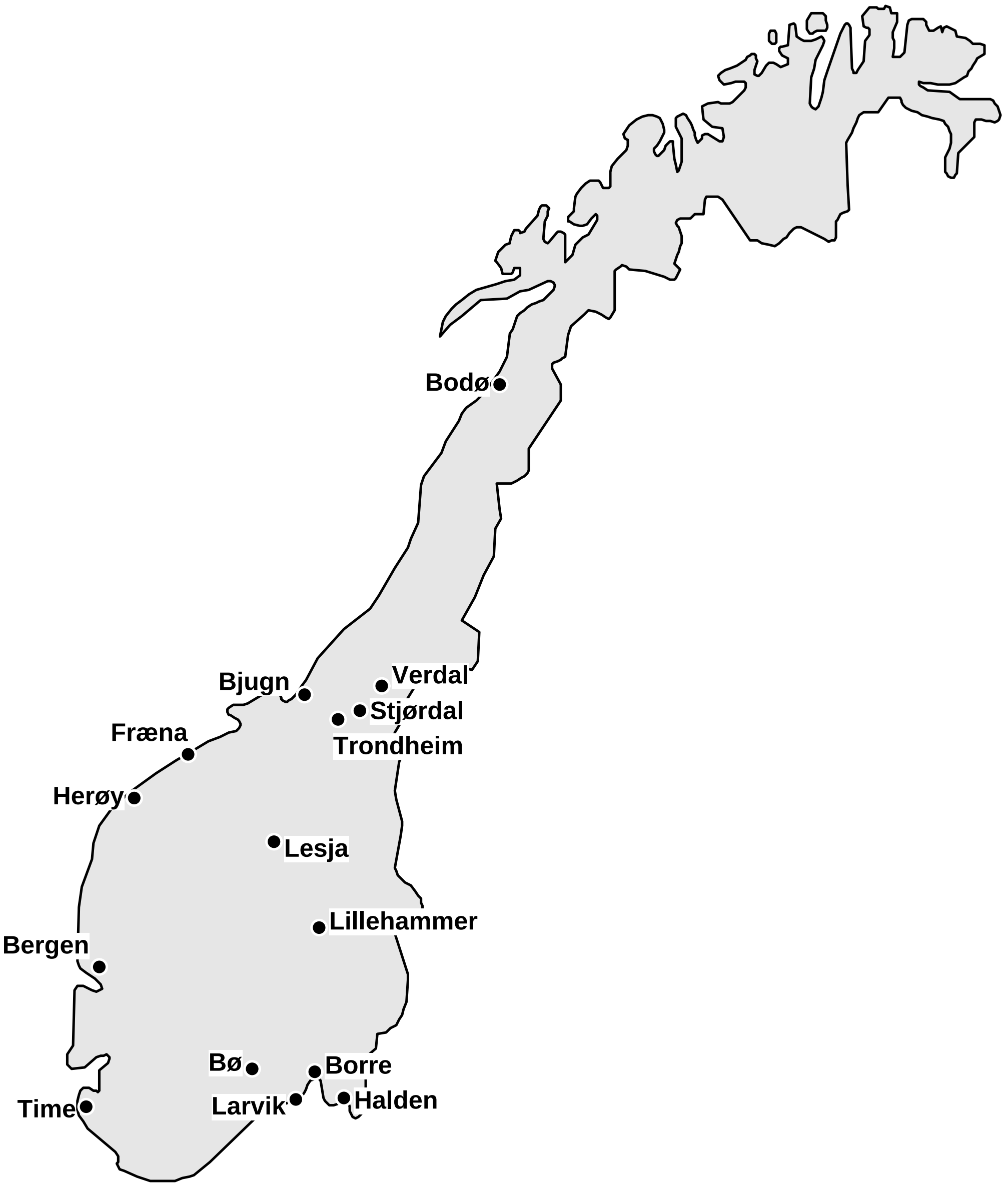


Figure 7.1: The geographic distribution of the 15 Norwegian varieties.

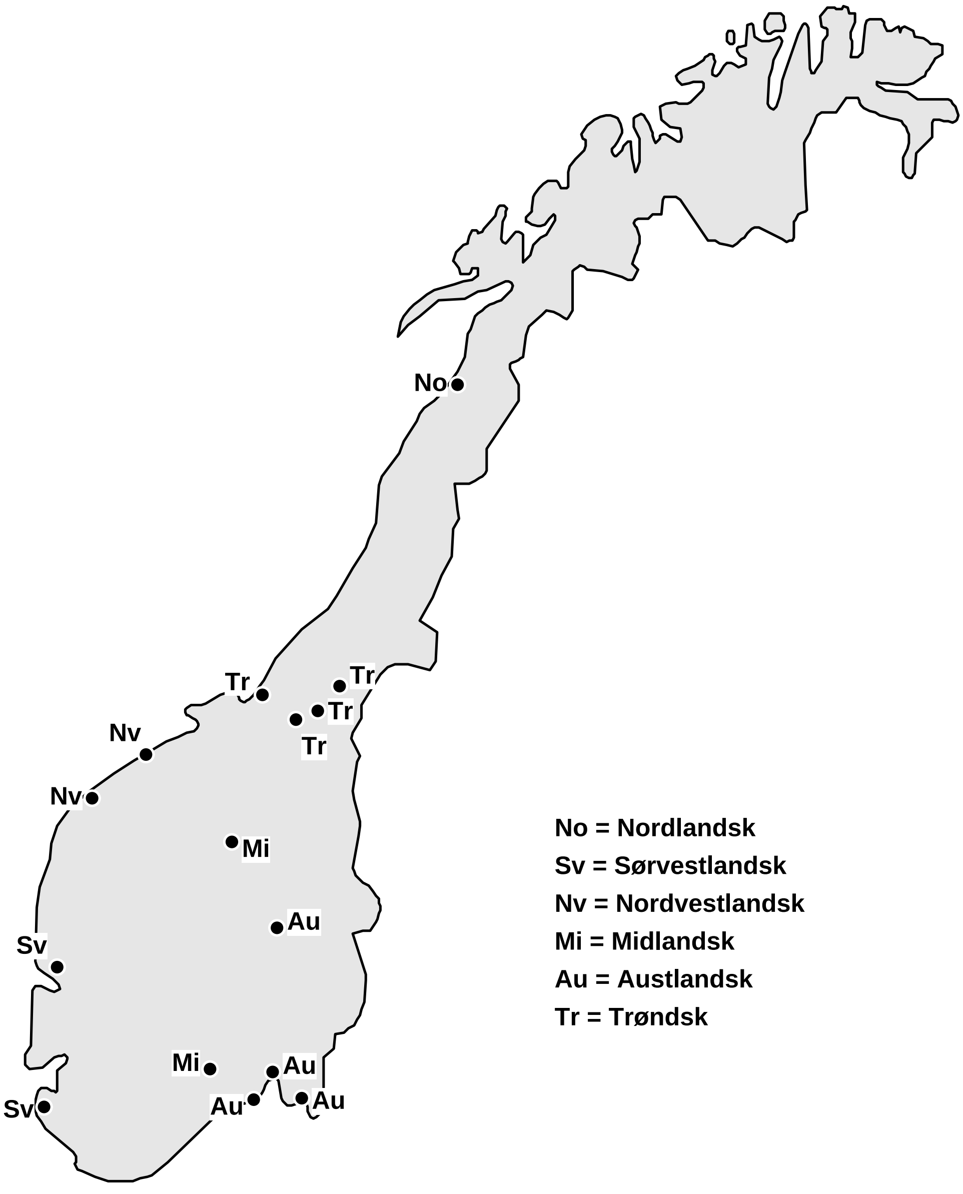


Figure 7.2: According to Skjekkeland (1997) the Norwegian language area can be divided in nine groups. The data points on this map correspond with those in Figure 7.1. In the set of 15 varieties six dialect areas are represented. The same abbreviations are used in the other figures in this chapter.

$$\alpha = \frac{n_w \times \bar{r}}{1 + (n_w - 1) \times \bar{r}} \tag{7.2}$$

As mentioned in Section 7.2.2 the 15 Norwegian varieties are compared on the basis of 58 words. For each matrix corresponding with a word $(58 \times 57)/2 = 1653$ distances are considered. The average inter-correlation is based on $(15 \times 14)/2 = 105$ pairs of matrices.

Usually the Cronbach's $\alpha$ may range between 0 and 1. The higher the $\alpha$, the more reliable the method. A widely-accepted threshold in social science is that $\alpha$ should be 0.70 or higher for a set of items to be considered a scale (Nunnally, 1978).

## 7.3.2 Results

In Table 7.1 the results for the 147 word-based methods are summarized. The main division in the table consists of transcription-based methods on the one hand, and recording-based methods on the other hand. For the transcription-based methods different factors are examined. Results are given for different segment representations, for two and four length gradations processed on the basis of changes in the transcription, for different diphthong representations and for different comparison metrics. For the recording-based methods scores are given for the three acoustic representations of the word recordings. When in the table a score is given for a certain combination of factors, the average is taken over the other factors.

Examining the different transcription-based methods used on the basis of different segment representations, the highest scores were found for the methods using phones, the linear and logarithmic Levenshtein distance using the feature system of H & H, the linear Levenshtein distance using the feature system of A & B and the logarithmic Levenshtein distance using Barkfilters and formant tracks. Further we see that the use of logarithmic segments distances instead of linear ones increases the correlation coefficients of the Levenshtein distances. The high score of the phone-based methods on the one hand, and the improvement which we found when using logarithmic segment distances in the Levenshtein algorithm on the other hand, may indicate that a more reduced number of distance gradations between words will improve the consistency. However, this does not necessarily imply that these most consistent methods will be better methods for validation as well.

From the table it appears that 2 length gradations will in general give higher Cronbach's $\alpha$ values than 4 length gradations. With regard to the diphthong representations, the different representations do not give different scores. Considering the histogram and feature bundle metrics, the Pearson correlation coefficient give the highest scores when using the frequency per word method, and the Euclidean distance gives the best scores when using the Levenshtein distance.

| | Freq.<br>word | Lev.<br>lin. | Lev.<br>log. |
|---|---|---|---|
| *Transcription-based* | | | |
| Segment representation discretely | | | |
| phones | 0.87 | 0.87 | 0.87 |
| features H & H | 0.84 | 0.87 | 0.87 |
| features V & C | 0.82 | 0.85 | 0.86 |
| features A & B | 0.82 | 0.85 | 0.87 |
| Segment representation acoustically | | | |
| Barkfilter | | 0.83 | 0.87 |
| cochleagram | | 0.82 | 0.86 |
| formant tracks | | 0.85 | 0.87 |
| Number of length gradations | | | |
| 2 lengths | 0.84 | 0.86 | 0.87 |
| 4 lengths | 0.83 | 0.85 | 0.86 |
| Diphthongs are represented as | | | |
| 2 segments | 0.83 | 0.85 | 0.87 |
| 1 segment | 0.83 | 0.85 | 0.87 |
| Comparison metric | | | |
| Manhattan | 0.82 | 0.85 | 0.87 |
| Euclidean | 0.83 | 0.87 | 0.88 |
| ‘Pearson’ | 0.84 | 0.85 | 0.85 |
| *Recording-based* | | | |
| Word representation acoustically | | | |
| Barkfilter | | 0.85 | |
| cochleagram | | 0.82 | |
| formant tracks | | 0.77 | |

Table 7.1: Average Cronbach’s $\alpha$ values on the basis of 58 words from 15 Norwegian varieties. The three columns corresponds respectively with the frequency per word method (Freq. word), the linear Levenshtein distance (Lev. lin.) and the logarithmic Levenshtein distance (Lev. log).

Looking at the different word representations which are used in combination with the recording-based Levenshtein distance, the highest score was found when using the Barkfilter representation, the lowest when using the formant track representation. This may be explained by the fact that formant tracks represent only a part of the information in a spectrogram, namely the dominant frequency tracks. This seems to result in less stable results. The formant track-based recording-based Levenshtein distance was at the same time the comparison method with the lowest score compared to all other methods.

In Table 7.1 the lowest Cronbach's $\alpha$ value was equal to 0.77, the highest was equal to 0.87. When examining the Cronbach's $\alpha$ values of all word-based methods separately, it appears that they vary from 0.77 to 0.88. So all methods have a Cronbach's $\alpha$ value which is higher than the threshold of 0.70 (see Section 7.3.1). Our conclusion is that all methods are reliable when using the 58 words of 'the North Wind and the Sun'.

### 7.3.3 Number of items

With Cronbach's $\alpha$ the number of items can be found which are needed to obtain consistent results. More items result in higher $\alpha$ values. In our data set we used 58 items. Using these items $\alpha$ was 0.77 or higher for all computational methods. When the threshold of $\alpha$ is 0.70 (see Section 7.3.1), we may conclude that with all computational methods reliable results can be obtained on the basis of 58 words.

In this section we investigate the effect of the number of items in more detail. For this purpose we use a variant of the transcription-based Levenshtein distance, where segment distances were found on the basis of the Barkfilter representation, four length gradations are used, diphthongs are represented as a sequence of two segments, and logarithmic segment distances are used. The choice of this method is justified in Section 7.5.1, and the method is applied in Section 7.5.2. With this method, distances between the 15 varieties are calculated for each of the words separately. Subsequently, we calculate $\alpha$ values on the basis of subsets of respectively 2 words, 3 words, and so on, through 58 words. The result is a range of 57 $\alpha$ values. The words in a subset are randomly chosen and each word is unique in a subset.

In Figure 7.3 we find a graph in which the x-axis represents the number of words and the y-axis represents Cronbach's $\alpha$. For lower number of words the graph fluctuates strongly. When the number of words increases, the graph becomes more stable and a gradual rise can be seen. From 25 words on $\alpha$ is always higher than 0.70. This means that words yield an acceptable degree of consistency, even using only 25 words and with the computational method we mentioned above. The highest $\alpha$ value is equal to 0.86, obtained on the basis of 58 words. To obtain a higher $\alpha$ value, we should use a larger number of words. The $\alpha$ value can be found with the formula (7.2) in Section 7.3.1. From this

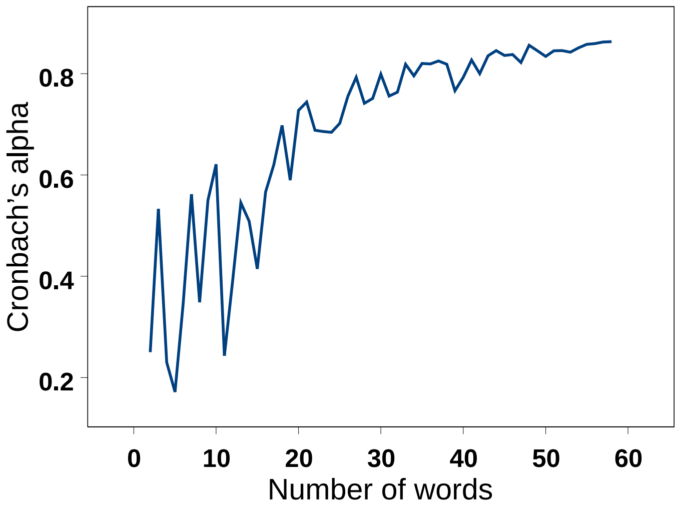


Figure 7.3: Cronbach's $\alpha$ values for 2 through 58 words. For smaller numbers of words the graph strongly fluctuates. When the number of words increases, the graphs becomes more stable and a gradual rise can be seen. From 25 words on $\alpha$ is always higher than 0.70. For 58 words $\alpha$ is equal to 0.86.

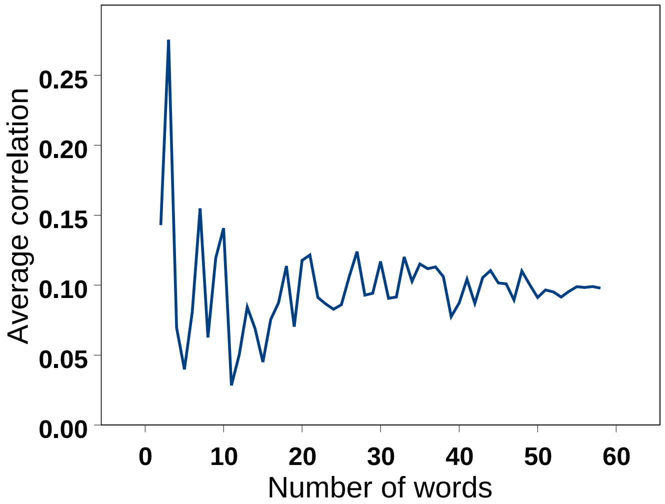


Figure 7.4: Average correlation coefficients for 2 through 58 words. For smaller numbers of words the graph strongly fluctuates. When the number of words increases, the graph becomes more stable. From 55 words on $\bar{r}$ remains stable with a value of about 0.10.

| Cronbach's $\alpha$ | Number of words |
|---|---|
| 0.86 | 55 |
| 0.87 | 60 |
| 0.88 | 66 |
| 0.89 | 73 |
| 0.90 | 81 |
| 0.91 | 91 |
| 0.92 | 104 |
| 0.93 | 120 |
| 0.94 | 141 |
| 0.95 | 171 |
| 0.96 | 216 |
| 0.97 | 291 |
| 0.98 | 441 |
| 0.99 | 891 |

Table 7.2: Number of words needed to obtain different $\alpha$ values. The numbers are rounded. For $\alpha = 1.00$ the number of words is not defined.

formula, we derived another formula with which the number of words $n_w$ can be found that are required to obtain a certain $\alpha$ value:

$$n_w = \frac{\alpha \times (\bar{r} - 1)}{\bar{r} \times (\alpha - 1)} \tag{7.3}$$

When using this formula, $\bar{r}$ should be known. For each of the subsets containing respectively 2 trough 58 randomly chosen words, we calculated $\bar{r}$. In Figure 7.4, we find a graph where the x-axis represents the number of words and the y-axis the corresponding $\bar{r}$'s. Just as we expected, for lower number of words $\bar{r}$ fluctuates strongly. When the number of words increases, $\bar{r}$ becomes more stable. From 55 words $\bar{r}$ remains stable with a value of about 0.10. Using $\bar{r} = 0.10$, we calculated the number of words that are needed to obtain $\alpha$ values from 0.86 to 0.99. The results are given in Table 7.2. From the formula it appears that the number of words is not defined for $\alpha = 1.00$.

We should emphasize before closing this section that the results depend strongly on the average inter-item correlation $\bar{r}$, which may be expected to vary from one family of dialects to another. The results here apply therefore to determine the sample size needed in other language areas only as a very general indication.

## 7.4 Validity

Heeringa et al. (2002) validated computational dialect comparison methods by comparing them to a gold standard. The gold standard provides a classification of language varieties with which (nearly) all experts agree. Varieties which experts disagree about are excluded. In this way the gold standard is incomplete, but it represents consensus. Heeringa et al. (2002) based their gold standard on two different Dutch dialect maps.

A gold standard is represented as a partition. The best validation results were obtained when the results of computational comparison methods were also converted to partitions (by clustering) and when these partitions were compared to the gold standard partition. The comparison of partitions was carried out by calculating the Rand index (Rand, 1971) and the Fowlkes and Mallow index (Fowlkes and Mallows, 1983).

We have since recognized disadvantages in this approach. First, dialect maps (and thus the gold standard) show only groups. The maps do not show precisely the proximity of linguistic relationships among the groups and within the groups. Second, varieties that are excluded (mostly borderline cases), play no role in evaluation even when they are classified into linguistically very different groups. Third, this technique cannot take the *degree* of misclassification into account, e.g., when a misclassified variety belongs to a linguistically very close group. Fourth and finally the measurements of the computational comparison methods are converted to partitions by clustering. The consequence is that information about the linguistic relationships within the groups is lost, and that the clustering itself – along with the distance measure – is then subject of validation.

In Gooskens and Heeringa (2004) distances obtained by a variant of the Levenshtein distance are validated by correlating them with perceptual distances. In the variant of the Levenshtein distance segment distances were determined acoustically. Perceptual distances are found in an experiment in which Norwegian listeners judge distances between 15 Norwegian varieties. The validity of the Levenshtein distance is tested by correlating it with the distances obtained by the perception experiment. The advantage compared to the gold standard-based approach is that validation is not based on simplified representations, namely partitions, but on gradual distances found between each possible pair of varieties. Moreover in this approach there is a clear criterion, namely perception, and no dependence on the use of the investigative technique, clustering.

In this section we will validate all 187 methods by correlating the results with perceptual distances. In Section 7.4.1 we describe the perception experiment with which perceptual distances were found. In Section 7.4.2 we discuss the way in which the perceptual distances are correlated with the distances resulting from each of the 187 computational methods. In Section 7.4.3 we discuss the results.

### 7.4.1 Perception experiment

The perception experiment was carried out by Charlotte Gooskens in the spring of 2000. The experiment is described more detailed in Gooskens and Heeringa (2004). In this Section we give only a brief description.

When the perception experiment was carried out, the recordings of only 15 Norwegian varieties were available. All of them are used. In order to be able to investigate the dialect distances between the 15 Norwegian dialects as perceived by Norwegian listeners, for each of the 15 varieties the corresponding recording of the translation of the fable 'The North Wind and the Sun' was presented to Norwegian listeners in a listening experiment. Because the computational comparison method as validated by Gooskens and Heeringa (2004) did not process intonation, both monotonized and original versions of the recordings were used in the perception experiment. The manipulations were carried out with the computer program PRAAT. In order to monotonize the fragments the pitch contours were changed to flat lines. The recordings of the male speakers were monotonized at 134, which is the average pitch of the four male speakers. The recordings of the female speakers were monotonized at 224 Hz. This was the average pitch of the female speakers.

The listeners were 15 groups of high school pupils, one group from each of the places where the 15 dialects are spoken. All pupils were familiar with their own dialect and had lived most of their lives in the place in question (on average 16.7 years). Each group consisted of 16 to 27 listeners. The mean age of the listeners was 17.8 years, 52 percent were female and 48 percent male.

The perception experiment consists of two sessions. In the first session the monotonized texts of the 15 varieties were presented in a randomized order. After a short break, the original texts of the same 15 varieties were presented again in randomized order. Each session was preceded by a practice recording (of a speaker of Stjørdal, but not one of the 15 recordings used in the experiment itself). Between each two recordings there was a pause of 3 seconds. While listening to the dialects the listeners were asked to judge each of the 15 dialects on a scale from 1 (similar to native dialect) to 10 (not similar to native dialect). This means that each group of listeners judged the linguistic distances between their own dialect and the 15 dialects, including their own dialect. In this way we get a matrix of 15 × 15 distances for each session.

When comparing the matrices, it appeared that the mean judgments are almost the same (7.19 for the monotonous fragments and 7.25 for the original fragments). However, the standard deviation is smaller in the case of the monotonous fragments (1.38) than in the case of the original fragments (1.68).[2] Two explana-

[2] The variances are not significantly different for $\alpha$=0.05 since $\text{P}(\text{F}(\binom{15}{2},\binom{15}{2}) = \frac{1.38^2}{1.68^2}) \Leftrightarrow \text{P}(\text{F}(105,105)=0.6747) \gg 0.05$.

| | Ber | Bju | Bod | Bø | Bor | Fræ | Hal | Her | Lar | Les | Lil | Stj | Tim | Tro | Ver |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bergen | 1.79 | 9.07 | 8.25 | 8.00 | 7.75 | 7.70 | 8.20 | 6.95 | 8.06 | 8.95 | 8.57 | 8.42 | 4.88 | 8.55 | 8.05 |
| Bjugn | 9.16 | 3.44 | 6.44 | 8.26 | 9.29 | 5.80 | 8.30 | 8.05 | 8.44 | 7.32 | 9.10 | 2.21 | 8.00 | 3.30 | 2.85 |
| Bodø | 8.79 | 7.93 | 1.50 | 8.32 | 8.35 | 6.60 | 7.90 | 7.84 | 7.39 | 8.05 | 8.76 | 6.63 | 8.19 | 6.20 | 6.30 |
| Bø | 8.11 | 7.81 | 7.56 | 1.00 | 7.76 | 8.10 | 4.95 | 7.89 | 5.39 | 6.00 | 5.19 | 7.16 | 6.31 | 8.25 | 8.65 |
| Borre | 6.11 | 8.85 | 7.81 | 6.53 | 1.76 | 8.55 | 1.80 | 7.58 | 1.61 | 7.53 | 2.04 | 7.26 | 7.50 | 8.55 | 9.10 |
| Fræna | 9.00 | 7.59 | 7.13 | 8.47 | 8.82 | 3.10 | 8.10 | 7.89 | 8.50 | 7.26 | 9.00 | 6.68 | 7.44 | 6.10 | 7.65 |
| Halden | 7.00 | 8.22 | 8.00 | 6.84 | 4.00 | 8.15 | 2.80 | 7.95 | 2.89 | 6.63 | 3.00 | 7.47 | 7.06 | 8.05 | 8.32 |
| Herøy | 8.63 | 9.37 | 8.44 | 8.53 | 9.18 | 7.05 | 8.65 | 1.26 | 9.33 | 9.32 | 9.48 | 8.58 | 7.50 | 7.50 | 8.22 |
| Larvik | 7.47 | 8.70 | 7.69 | 4.05 | 4.06 | 7.75 | 3.25 | 5.61 | 3.44 | 7.16 | 4.67 | 8.21 | 6.88 | 8.35 | 7.55 |
| Lesja | 8.58 | 7.63 | 7.88 | 7.42 | 8.24 | 7.30 | 7.60 | 7.79 | 7.67 | 1.00 | 7.10 | 6.95 | 7.25 | 7.70 | 8.22 |
| Lillehammer | 6.78 | 8.33 | 8.13 | 6.26 | 4.47 | 8.05 | 3.10 | 7.53 | 4.11 | 7.32 | 2.76 | 7.68 | 6.88 | 8.70 | 8.16 |
| Stjørdal | 8.74 | 3.73 | 6.81 | 7.79 | 8.18 | 6.05 | 7.55 | 7.79 | 8.35 | 7.16 | 8.38 | 2.05 | 7.75 | 3.85 | 3.42 |
| Time | 7.00 | 9.33 | 8.44 | 8.11 | 8.47 | 8.30 | 8.05 | 7.22 | 8.22 | 9.11 | 8.81 | 8.89 | 1.81 | 8.80 | 9.05 |
| Trondheim | 7.84 | 5.89 | 6.75 | 7.53 | 6.47 | 7.35 | 6.05 | 7.16 | 5.94 | 7.94 | 6.33 | 4.47 | 7.63 | 3.35 | 6.84 |
| Verdal | 8.89 | 3.41 | 6.44 | 8.26 | 8.41 | 5.70 | 7.25 | 7.95 | 7.94 | 7.42 | 8.48 | 1.89 | 7.94 | 3.15 | 2.63 |

Table 7.3: Average perceptual distances between all pairs of 15 Norwegian dialects as perceived by 15 groups of listeners judged on a scale from 1 (=similar to own dialect) to 10 (=not similar to own dialect). A column gives the average judgments of different groups of listeners for the same speaker, and a row gives the average judgments for different speakers judged by the same group of listeners.

tions suggest themselves. First the absence of intonation yields unnatural speech. In particular the absence of intonation makes tonemes imperceptible in Norwegian which makes the fragments even more unusual. According to Gooskens and Heeringa (2004) the consequence may be that this makes listeners insecure. This leads to 'safe' judgments, resulting in values which are found closer to the middle of the scale. Second the lower standard deviation for the monotonous distances may have to do with the setup of the experiment. After the first session the listeners know the extremes, i.e., the most similar and most different varieties. This knowledge may be used when judging distances in the second session.

In Gooskens and Heeringa (2004) it is striking that the distances of the comparison method correlate better with the original perceptual distances than with the monotonous perceptual distances, even though the comparison method does not process prosodic information any way. Therefore, we decided only to use the perceptual distances based on the results of the second session, which used the original (non-monotonized) recordings. In this second session the recordings are presented in a natural way. Knowledge about the extremes from the first session is probably used, with the result that the full range of the scale is used. The average judgments as given by the listeners in the second session are given in Table 7.3.

There are two mean distances between each pair of dialects. For example the distance as perceived by the listeners in Bergen with respect to the dialect of Trondheim is different from the distance as perceived by the listeners in Trondheim with respect to the dialect of Bergen. Since both the cluster program and the multidimensional scaling program expect only one value for each pair of different elements, the average of the two mean distances is used when classifying the varieties on the basis of the perceptual distances. In Figure 7.5 a dendrogram is given and in Figure 7.6 a multidimensional scaling plot.

Both the dendrogram and the multidimensional scaling plot accord rather well with the map of Skjekkeland (see Figure 7.2). Sørvestlandsk, Austlandsk and Trøndsk groups can clearly be identified. However, the Midlandsk dialects, Bø and Lesja, do not form a close cluster. Geographically they are rather distant, so they may be rather different although they should be in the same group according to the traditional division. However, in the multidimensional scaling plot the Midlandsk dialect of Lesja is closest to the Midlandsk dialect of Bø. The Nordvestlandsk dialects seem to be very different in both the dendrogram and the multidimensional scaling plot, although they are geographically rather close. Possibly this may be explained by the fact that the map of Skjekkeland is based (partly) on phenomena other than these found in the text 'The North Wind and the Sun'. In our sample the Nordlandsk area is represented by only one variety (Bodø). This variety is grouped with the varieties of the Trøndsk area, which is not unexpected geographically.

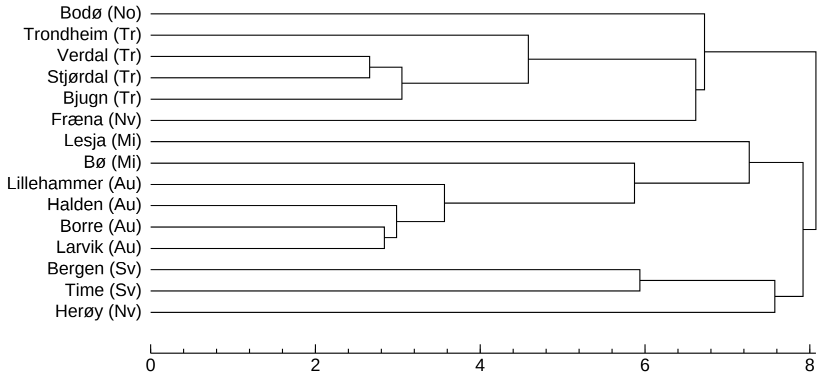


Figure 7.5: Dendrogram derived from the 15 × 15 matrix of perceptual distances showing the clustering of (groups of) Norwegian dialects. UPGMA clustering is used (see Section 6.1.2). On the horizontal scale distances are given in the scale as used by the listeners. The abbreviations between parentheses are explained in Figure 7.2. A Sørvestlandsk, an Austlandsk and a Trøndsk group can clearly be identified. The tree structure explains 91% of the variance.

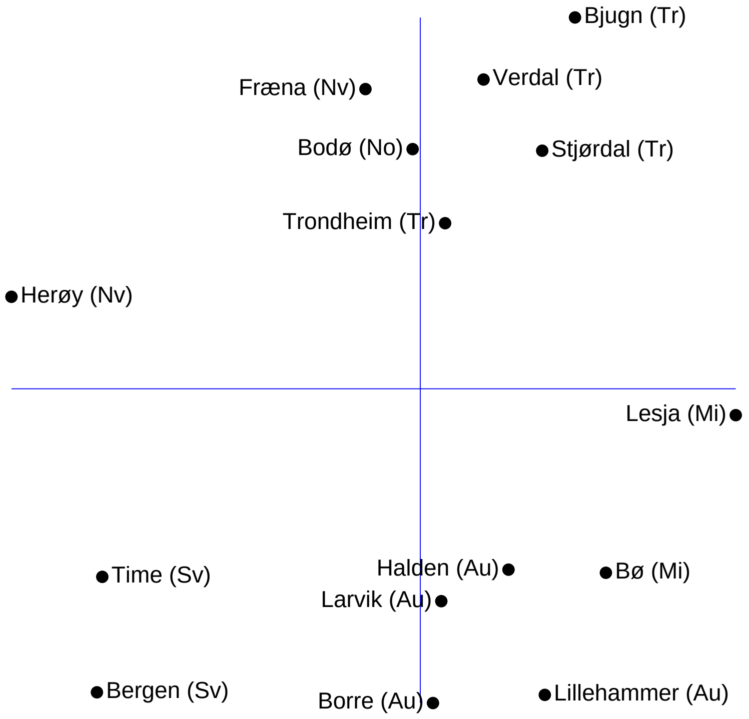


Figure 7.6: Multidimensional scaling of the results derived from the 15 × 15 matrix of perceptual distances. Kruskal's Non-metric MDS is used (see Section 6.2.2). The abbreviations between parentheses are explained in Figure 7.2. The y-axis (first dimension) corresponds with the geographic north-south axis, the x-axis (second dimension) more or less with the west-east axis. Two dimensions explain 67% of the variance.

### 7.4.2 Correlation

When examining Table 7.3, it is striking that the distances of varieties with respect to themselves are mostly higher than 1 (1 was the lowest possible judgment in the experiment). E.g., the listeners of Trondheim give an average judgment of 3.35 when hearing the recording of a speaker of their own city. This may be explained by the fact that there is variation among the dialect speakers in Trondheim. A slight deviation may reflect different idiolects, but a greater deviation may reflect different varieties spoken in different parts in Trondheim. When the listeners come (partly) from different parts of Trondheim than the speaker of the recording does, their average judgment will never be equal to 1. The fact that the listeners of a given location do not recognize every speaker of the same location as familiar, will cause the distance of a location with respect to itself to be greater than 1. This is different from the use of computational comparison methods where speakers are directly compared to each other without the intervention of listeners' judgments.

When comparing varieties of different locations, the variation within each location may again play a role. Assume we want to find the distance between locations $A$ and $B$. In $A$ two related varieties $A_1$ and $A_2$ are spoken. Assume further that in the experiment the speaker of $A$ spoke variety $A_1$, and that the listeners' group of $A$ is familiar with variety $A_2$. In the perception experiment we determine the distance between $A_2$ and $B$. However, when using a computational comparison method the distance between speakers is found, i.e. the distance between $A_1$ and $B$. Therefore, the perceptual and the computational distance need not to be equal: the one may be higher than the other.

It may be clear that the deviation of perceptual distances between varieties at the *same* location always goes in *one* direction compared to the corresponding computational distances: they will be relatively higher. Therefore, when correlating the matrix of perceptual distances with a matrix of computational distances, these higher perceptual distances may cause distortion, which justifies eliminating them. When calculating the correlation coefficient, the values on the diagonal (from upper left to lower right) are not taken into account. However, the distortion of perceptual distances between *different* locations may go into *two* directions compared to the corresponding computational distances: they can be either relatively higher or relatively lower. Therefore, we regarded these deviations as noise which will cause no significant distortion, when correlating the matrix of perceptual distances with a matrix of computational distances. All distances between different locations are considered when calculating the correlation coefficient.

For finding the correlation coefficient, we used the Pearson's correlation coefficient (Sneath and Sokal, 1973, pp. 137–140). When having 15 varieties, a distance matrix will have 15 rows and 15 columns. The correlation coefficient between a matrix $X$ and a matrix $Y$ is calculated as:

$$r(X,Y) = \frac{\sum_{i=1}^{n}\sum_{j\neq i}(X_{ij}-\overline{X})(Y_{ij}-\overline{Y})}{\sqrt{\sum_{i=1}^{n}\sum_{j\neq i}(X_{ij}-\overline{X})^2}\sqrt{\sum_{i=1}^{n}\sum_{j\neq i}(Y_{ij}-\overline{Y})^2}} \tag{7.4}$$

where $n = 15$. Correlation coefficients range from $-1$ (perfect inverse correlation) to $+1$ (perfect correlation). There is no correlation if $r = 0$.

For finding the significance of a correlation coefficient we used the Mantel test, just as in Chapters 3 and 4. The Mantel test is explained in Section 3.8.2. As significance level we choose $\alpha = 0.05$. With the Mantel test it is also possible to determine whether one correlation coefficient is significantly higher than another. In the sections below mostly averaged correlation coefficients are given. Since our implementation of the Mantel test only compares individual correlation coefficients, we were not able to determine whether two averaged correlation coefficients are significantly different.

### 7.4.3 Results

When examining the correlation coefficients of all 187 methods separately, we found that they vary from 0.33 to 0.67. All correlations were significant. But a perfect correlation was not found. We should be aware of the fact that in the perception experiment the listeners were confronted with recordings of spoken texts. These texts include lexical, phonetic, prosodic, morphological and syntactical information. However, when applying computational comparison methods to the transcriptions of word pronunciations, only lexical, phonetic and morphological information is processed. In the feature-based methods the presence of tonemes is also processed, but only in these methods. However, the exact realization of these tonemes is never processed. They are treated as categorical differences (see Section 3.4.1).

In the Tables 7.4, 7.5, 7.6, 7.7 and 7.8 results for the 184 transcription-based methods are given, and in Table 7.9 results for 3 recording-based methods are given. In the tables for the transcription-based methods results are given for the corpus frequency method, the frequency per word method and the Levenshtein distance (using linear or logarithmic segment distances). In these tables the factors mentioned in Section 7.1 are examined. When particular factors are examined in a table, the average is taken over the factors which are not mentioned in that table.

#### 7.4.3.1 Transcription-based comparison methods

In Table 7.4 correlation coefficients are given for different segment representations for each set of transcription-based methods. Examining the phone-based methods, we find that the corpus frequency method performs as well as the frequency per word method. The Levenshtein distance in turn performs better than both the corpus frequency method and the frequency per word method. This confirms our conviction that the Levenshtein distance is methodologically better. The linear and the logarithmic Levenshtein distance have the same correlation. When using phones there exist only two differences: 0 (equal) and 1 (different). So the relative distances are not changed when transformed logarithmically.

Examining the feature-based methods, we see that the frequency per word method performs better than the corpus frequency method, the linear Levenshtein distance performs better than the frequency per word method, and the logarithmic Levenshtein distance performs better than the linear Levenshtein distance for each of the three different feature systems. This order reflects the different steps of improvement in this range of methods. When using features the word-based methods are clearly better than the corpus-based methods. When using phones this improvement was not found. As of this writing we have found no explanation for this.

For both the feature-based Levenshtein distances and the acoustic-based Levenshtein distances the use of the logarithmic distances instead of linear distances improves the correlation. It confirms our idea that the use of the logarithm mimics perception better.

It is striking that the phone-based methods and the acoustic-based logarithmic Levenshtein distances perform best. Even the two frequency-based methods using the phone representation perform well. Why do especially the phone-based methods perform better than the more refined feature-based methods? Perhaps this may be explained if, for dialect speakers, all distances between different segments are the same. This means that the distance between [i] and [ɪ] is equal to the distance between [i] and [ɒ]. In fact the first distance is relatively too large (it should be smaller than the distance between [i] and [ɒ]) and the second relatively too small (it should be larger than the distance between [i] and [ɪ]). This may result in a logarithmic effect in which small differences weigh disproportionally, which in turn yields a higher correlation with the perceptual distances.

Although the phone-based methods perform as well (or even any better) than the acoustic-based logarithmic Levenshtein distances, we prefer the methods last mentioned. The results shown here are based on a small set of 15 varieties.[3] Having more varieties results in a network with a higher density, in which small details may become more important. These details are not processed in the phone-based methods, only in the acoustic and feature-based methods. Validation on the basis of a larger set of varieties is advisable in future work.

[3] Having 15 varieties the correlation is still based on $15 \times 14 = 210$ distances!

| | Freq. corp. | Freq. word | Lev. lin. | Lev. log. |
|---|---|---|---|---|
| Segment representation discretely | | | | |
| phones | 0.66 | 0.66 | 0.67 | 0.67 |
| features H & H | 0.47 | 0.61 | 0.64 | 0.65 |
| features V & C | 0.45 | 0.59 | 0.61 | 0.63 |
| features A & B | 0.45 | 0.58 | 0.62 | 0.64 |
| Segment representation acoustically | | | | |
| Barkfilter | | | 0.65 | 0.66 |
| cochleagram | | | 0.64 | 0.66 |
| formant tracks | | | 0.64 | 0.66 |

Table 7.4: The effect of methods and representations is shown in the average correlation coefficients of transcription-based distance measures with respect to perceptual distances on the basis of 15 Norwegian varieties. The four columns present respectively the corpus frequency method (Freq. corp.), the frequency per word method (Freq. word), the linear Levenshtein distance (Lev. lin.) and the logarithmic Levenshtein distance (Lev. log). For each method the average scores for different segment or word representations are given.

#### 7.4.3.2 Representation of segments

In Section 7.4.3.1 we found that the phone-based methods and the acoustic-based logarithmic Levenshtein distances perform best. In this section we compare the different segment representations in more detail.

**Different feature representations** When comparing the different feature representations in Table 7.4, we find that the feature system of H & H gives better results than the other systems. This is especially striking since the V & C and A & B systems were developed especially for measuring transcription differences, while the basis of the H & H system, the features of Chomsky and Halle's *The Sound Pattern of English*, was developed for encoding phonological rules. It was our expectation that the perceptually-motivated system of V & C would give the best results. Even though, the system was originally developed for Dutch. Later on the system was extended so that it contains all vowels and pulmonic consonants of the IPA system (see Section 3.1.3). Although the extensions were made along the lines of the original system, they are not directly based on the perception of listeners. This may possibly explain the lower correlations in Table 7.4.

**Feature vs. phone representation** Although the use of phones gives better results than the use of features for all methods, we found the greatest difference for the corpus frequency method in Table 7.4. In a histogram of phones, frequencies are given for the phones individually. In a histogram of features, feature frequencies are given. The frequency per feature gives the number of sounds for which that feature was positive. Different sounds may contribute to the frequency of the same feature. The result is that we lose information. We illustrate this by a hypothetical example. Assume the corpus of a dialect contains a [b] and a [f]. The feature *voice* gets a frequency of 1 (due to the [b]) and the feature *continuant* get a frequency of 1 (due to the [f]). Another dialect contains a [p] and a [v]. The features *voice* and *continuant* get a frequency of 1 (only due to the [v]). The consequence is that the different dialects get the same frequencies and will erroneous appear to be equal when comparing the feature frequencies. This will not happen when using phones. For the first dialect the [b] and the [f] get a frequency of 1, and for the second dialect the [p] and the [v] get a frequency of 1. The dialects are clearly distinguished, although differences may be exaggerated when comparing the phone-based histograms.

**Different acoustic representations** The acoustic representations are only used in combination with Levenshtein distance. For the Barkfilter we found a higher correlation than for the two other acoustic representations when using the linear Levenshtein distance. Using the Barkfilter representation voiceless and voiced sounds are more sharply distinguished and the vowels and r-like sounds are closer than when using the cochleagram representation. Comparing the Barkfilter with the formant track representation, we get a vowel quadrilateral rather than a triangle. Using the Barkfilter representation plosives and fricatives are clearly distinguished, which is not the case when using the formant track representation (see Sections 4.3.1.2, 4.3.2.2 and 4.3.3.2). When using the logarithmic Levenshtein distance, all acoustic representations have the same correlation. Using logarithmic distances, smaller distances become relatively more important than larger differences. This makes the acoustic-based segment distances more similar.

**Acoustic vs. feature representation** Since the acoustic representations are only used in combination with Levenshtein distance, the comparison between acoustic and feature representations can only be made on the basis of the results of Levenshtein distance. Among the different feature representations, we found the highest correlation for the feature system of H & H. Equally good or better results are obtained when using acoustic representations. In the feature system of H & H consonants are mainly distinguished by the manner of articulation. This applies for the acoustic representations as well. However, in the systems of V & C and A & B the place of articulation is clearly represented. Our results suggest that the manner of articulation plays a more important role in perception. From

Table 7.4 it appears that the use of acoustic representations gives better results than the use of feature representations in general.

#### 7.4.3.3 Number of length gradations

Table 7.5 presents – for each comparison method and for each segment representation – a comparison of 2 length gradations versus 4 length gradations as processed by changing the transcription. We should be aware of the fact that for phone-based and acoustically-based methods half-long and long are not processed when using 2 length gradations. When using a feature representation, half-long and long are processed by changing a feature value.

In the table in three cases we found that 2 length gradations result in a higher correlation coefficient than 4 length gradations, and in nine cases 4 length gradations is superior.[4] So we conclude that the use of 4 length gradations in general gives better results than the use of 2 length gradations, but the differences are seldom large.

We found no systematic distinction between the phone-based and acoustically-based methods on the one hand, and the feature-based methods on the other hand. For example, the phone-based corpus frequency method performs better when using 4 gradations, but the frequency per word method performs better when using 2 gradations. Looking at the feature-based methods, we found clear improvements for the H & H-based and V & C-based corpus frequency methods when using four gradations. H & H themselves used two gradations while processing half-long and long by changing a feature value.

#### 7.4.3.4 Representation of diphthongs

Table 7.6 compares two diphthong representations for each comparison method and for each segment representation. In the first a diphthong is processed as the sequence of two monophthongs and in the second as one segment. It seems to make no difference whether diphthongs are represented as two segments or as one segment. In three cases we found that the representation as two segments results in a higher correlation than the representation as one segment, and in three cases the representation as one segment results in a higher correlation than the representation as two segments.[5] On the basis of these outcomes a clear conclusion cannot be drawn. This may be explained by the fact that only a small set of diphthongs were defined for the NOS data.

Examining the feature-based corpus frequency methods, we observe that when using the feature system of H & H, the higher correlation coefficient is obtained

[4] Since the phone-based logarithmic Levenshtein distance yields the same results as the linear counterpart, we only counted results of the latter. See Section 7.4.3.1.

[5] Just as in Section 7.4.3.3 we only counted results of the linear Levenshtein distance when considering the phone-based methods.

| | Freq. corp. | Freq. word | Lev. lin. | Lev. log. |
|---|---|---|---|---|
| Phones | | | | |
| 2 lengths | 0.66 | 0.67 | 0.66 | 0.66 |
| 4 lengths | 0.66 | 0.66 | 0.67 | 0.67 |
| Features H & H | | | | |
| 2 lengths | 0.43 | 0.61 | 0.64 | 0.65 |
| 4 lengths | 0.50 | 0.61 | 0.64 | 0.65 |
| Features V & C | | | | |
| 2 lengths | 0.43 | 0.59 | 0.61 | 0.62 |
| 4 lengths | 0.47 | 0.60 | 0.62 | 0.63 |
| Features A & B | | | | |
| 2 lengths | 0.45 | 0.58 | 0.62 | 0.64 |
| 4 lengths | 0.44 | 0.59 | 0.62 | 0.64 |
| Barkfilter | | | | |
| 2 lengths | | | 0.64 | 0.66 |
| 4 lengths | | | 0.65 | 0.67 |
| cochleagram | | | | |
| 2 lengths | | | 0.64 | 0.66 |
| 4 lengths | | | 0.64 | 0.66 |
| formant tracks | | | | |
| 2 lengths | | | 0.65 | 0.66 |
| 4 lengths | | | 0.64 | 0.66 |

Table 7.5: The effect of *segment length discrimination* is shown in the average correlations of transcription-based methods with perceptual distances on the basis of 15 Norwegian varieties. For each comparison method and for each segment representation the average scores for 2 length gradations versus 4 length gradations can be compared.

| | Freq. corp. | Freq. word | Lev. lin. | Lev. log. |
|---|---|---|---|---|
| Phones | | | | |
| 2 segments | 0.66 | 0.66 | 0.67 | 0.67 |
| 1 segment | 0.66 | 0.66 | 0.66 | 0.66 |
| Features H & H | | | | |
| 2 segments | 0.47 | 0.61 | 0.64 | 0.65 |
| 1 segment | 0.46 | 0.61 | 0.64 | 0.65 |
| Features V & C | | | | |
| 2 segments | 0.45 | 0.59 | 0.61 | 0.63 |
| 1 segment | 0.46 | 0.59 | 0.61 | 0.63 |
| Features A & B | | | | |
| 2 segments | 0.44 | 0.58 | 0.62 | 0.64 |
| 1 segment | 0.45 | 0.59 | 0.62 | 0.64 |
| Barkfilter | | | | |
| 2 segments | | | 0.65 | 0.66 |
| 1 segment | | | 0.65 | 0.66 |
| cochleagram | | | | |
| 2 segments | | | 0.64 | 0.66 |
| 1 segment | | | 0.64 | 0.66 |
| formant tracks | | | | |
| 2 segments | | | 0.65 | 0.66 |
| 1 segments | | | 0.64 | 0.66 |

Table 7.6: The effect of *diphthong representation* is shown in the average correlations of transcription-based methods with perceptual distances on the basis of 15 Norwegian varieties. For each comparison method and for each segment representation the average scores for two diphthong representations are compared. In the first a diphthong is processed as the sequence of two monophthongs and in the second as one segment. There is very little difference in treating diphthongs as one versus two segments.

when a diphthong is represented as two segments. On the other hand for the feature systems of V & C and A & B the higher correlation coefficients are obtained when a diphthong is represented as one segment. This difference between H & H on the one hand, and V & C and A & B on the other hand, concerns the way in which height is defined for closing diphthongs (in the NOS data there are no centering diphthongs). In H & H the height is equal to the height of the first segment, in the other systems the mean of the heights of the first and the second segment is used.

#### 7.4.3.5 Comparison of feature histograms or feature bundles

In Table 7.7 different metrics for finding distances between histograms and feature bundles are compared for the feature-based comparison methods. For the corpus frequency method, Manhattan gives the best results when using the feature systems of H & H and V & C, and Euclidean gives the best results when using the feature system of A & B. For all systems the use of Pearson's correlation coefficient gives the lowest correlation with the perceptual distances. This is especially striking for the H & H system, since H & H themselves used this metric in all their publications. Looking at the results of the word-based methods, the Euclidean distance unanimously appears to be the best metric.

Why does Manhattan give the better results most of the time when using the corpus frequency method, and does Euclidean appear to be the best metric when using word-based methods? When comparing the Manhattan metric to the Euclidean metric (see the formulas given in respectively (3.1) and (3.2) in Section 3.6.2.5), we see that feature differences are squared when using the Euclidean metric. The result is that larger differences are weighted relatively more heavily than smaller differences. When using the corpus frequency method, dialect distances are measured with the metrics. Using the frequency per word method and Levenshtein distance, respectively word distances and segment distances are calculated with these metrics. This indicates that on the highest level (comparison of dialects) feature differences should be weighted equally, but on the deeper levels (comparison of words or segments) larger differences should be weighted relatively more heavily than smaller ones.

In Table 7.8 in fact the same scores are given as in Table 7.7. However, now for each comparison method and for each histogram or feature bundle comparison metric the average scores for the different feature-based segment representations are given. The table shows that, regardless which comparison method and histogram or feature bundle metric is used, the feature system of H & H gives mostly better results than the other systems. The findings for all three comparison metrics are nearly the same. Using the corpus frequency method the feature system of V & C performs as well or better than the feature system of A & B. For the frequency per word method it turns out that the feature system of V & C gives better results than the system of A & B when using Manhattan or Euclidean,

| | Freq. corp. | Freq. word | Lev. lin. | Lev. log. |
|---|---|---|---|---|
| Features H & H | | | | |
| Manhattan | 0.49 | 0.61 | 0.63 | 0.65 |
| Euclidean | 0.48 | 0.64 | 0.66 | 0.66 |
| 'Pearson' | 0.43 | 0.58 | 0.64 | 0.63 |
| Features V & C | | | | |
| Manhattan | 0.46 | 0.59 | 0.61 | 0.63 |
| Euclidean | 0.47 | 0.61 | 0.64 | 0.65 |
| 'Pearson' | 0.43 | 0.58 | 0.59 | 0.61 |
| Features A & B | | | | |
| Manhattan | 0.46 | 0.56 | 0.61 | 0.64 |
| Euclidean | 0.45 | 0.59 | 0.65 | 0.66 |
| 'Pearson' | 0.42 | 0.60 | 0.61 | 0.63 |

Table 7.7: The effect of *feature bundle comparison metrics* is shown in the average correlations of transcription-based methods with perceptual distances on the basis of 15 Norwegian varieties. For each comparison method and for each feature-based segment representation the average scores for different feature bundle comparison metrics are given.

but when using Pearson's correlation coefficient it is the opposite. When using the Levenshtein distances the feature system of A & B performs better than that of V & C, regardless the choice of the histogram of feature bundle metric. We conclude that there is no interaction between feature representation and feature metric.

#### 7.4.3.6 Recording-based comparison methods

In Table 7.9 the scores of the recording-based Levenshtein distances are given. As mentioned in the Sections 7.2.3 and 7.2.4 in the set of 15 varieties there were 4 male speakers and 11 female speakers. Correlation coefficients are given for both genders separately, and for all speakers. When using all speakers the use of the formant track representation gives the highest correlation. This correlation is higher than those for the transcription-based corpus frequency methods using features. However, when using Barkfilters or cochleagrams as acoustic word representations, the correlations are lower than those for all transcription-based methods. We explain the low correlations of the recording-based methods by two facts. First the way in which sample sizes are normalized for speech rate is very rough (see Section 5.2.3). This may hamper the finding of correct alignments and corresponding distances by the Levenshtein algorithm. Second voice quality may still play a role, although the samples are monotonized. Differences between

| | Freq. corp. | Freq. word | Lev. lin. | Lev. log. |
|---|---|---|---|---|
| Manhattan | | | | |
| Features H & H | 0.49 | 0.61 | 0.63 | 0.65 |
| Features V & C | 0.46 | 0.59 | 0.61 | 0.63 |
| Features A & B | 0.46 | 0.56 | 0.61 | 0.64 |
| Euclidean | | | | |
| Features H & H | 0.48 | 0.64 | 0.66 | 0.66 |
| Features V & C | 0.47 | 0.61 | 0.64 | 0.65 |
| Features A & B | 0.45 | 0.59 | 0.65 | 0.66 |
| ‘Pearson’ | | | | |
| Features H & H | 0.43 | 0.58 | 0.64 | 0.63 |
| Features V & C | 0.43 | 0.58 | 0.59 | 0.61 |
| Features A & B | 0.42 | 0.60 | 0.61 | 0.63 |

Table 7.8: The interaction between feature sets and feature bundle comparison metrics is shown in the average correlations of transcription-based methods with perceptual distances on the basis of 15 Norwegian varieties. For each comparison method and for each histogram or feature bundle comparison metric the average scores for the different feature-based segment representations are given. It is clear that Manhattan and Euclidean distance are similar, and that the Pearson measure is inferior.

male and female voices may influence the result (Heeringa and Gooskens, 2003). This appears to be confirmed by the fact that the correlation coefficients on the basis of the female speakers are higher than on the basis of all speakers. The correlation coefficients on the basis of the male speakers are much lower. This may indicate that diversity in voice quality is relatively larger for male speakers than for female speakers. The highest correlation coefficient for male speakers – just as for male and female speakers, taken together – was obtained using formant tracks. Apperently, this representation is less sensitive to voice quality. However, the transcription-based word-based comparison methods perform still much better.

## 7.5 Choice and results

The findings of the Sections 7.3 and 7.4 enable us to select the optimal method for finding distances between varieties. The choice of the method is made in Section 7.5.1. In Section 7.5.2 we apply this method to Norwegian data and show results.

| | Male | Female | All |
|---|---|---|---|
| Barkfilter | 0.08 | 0.44 | 0.33 |
| cochleagram | 0.12 | 0.55 | 0.41 |
| formant tracks | 0.36 | 0.55 | 0.50 |

Table 7.9: Correlations of recording-based Levenshtein distances with perceptual distances on the basis of 15 Norwegian varieties using three different acoustic word representations. Scores are given for the male and female speakers separately, and for all speakers.

### 7.5.1 Choice of method

In Section 7.3 we examined the reliability of the word-based methods. We found that all methods are reliable when using the 58 words of 'the North Wind and the Sun'. In Section 7.4 we validated transcription-based and recording-based methods with respect to the results of a perception experiment. Examining the transcription-based methods we found that the phone-based methods and the acoustic-based logarithmic Levenshtein distances give results that correlate strongest with perceptual distances. Among the feature representations, the H & H system yields the best results. Among the acoustic representations we found the Barkfilter representation better than the other two representations, but only when using the linear Levenshtein distance. We found that the use of 4 length gradations will in general give better results than the use of 2 length gradations. For the diphthongs we compared the representation as two segments with the representation as one segment but could not draw a definite conclusion. When using features, the Manhattan metric gives the better results most of the time when using the corpus frequency method. When using word-based methods, the use of the Euclidean metric is preferable. Examining the recording-based methods we found that the three Levenshtein variants gave less satisfying results.

Therefore, we choose one of the transcription-based methods. From this range of methods we have to choose from the phone-based methods and the acoustic-based logarithmic Levenshtein distances. We found the highest average score for the phone-based Levenshtein distances. Nevertheless, we maintained our preference for the acoustic-based logarithmic Levenshtein distances in Section 7.4.3.1. In a small set of 15 varieties, the rougher phone-based methods may perform well, but for a denser sampling, minor differences may play a stronger role. Using the acoustic-based logarithmic Levenshtein distance, these differences are taken into account to a greater deal. Examining Table 7.4, we find that different acoustic representations give the same average scores when using the logarithmic Levenshtein distance. When looking at the results of the linear Levenshtein distance, we find the highest score when the Barkfilter is used. Since we are forced to make a choice, we choose the Barkfilter representation. With regard to the number of length gradations, of course we prefer 4 length gradations. As mentioned above no obvious conclusion could be drawn about the representation of diphthongs. When examining the scores of the acoustic-based Levenshtein distances in Table 7.4.3.4, we only found different results for the linear Levenshtein distance using the formant track representation. For this method, the two-segmental representation gave better results. Therefore, we choose the two-segmental representation of diphthongs. With that we have made our choice, namely the method with the parameters we chose above. We do not need to choose a feature bundle metric since we use an acoustic representation.

It may be expected that the method which we chose on the basis of different parameters, belongs to the better ones in the set of all of the 187 methods. To examine this, all 187 methods were sorted according to their correlation with the perceptual distances. When examining the sorted list of methods, it appears that our method even has the highest correlation coefficient. The Cronbach's $\alpha$ for this method is 0.86. The correlation coefficient between distances obtained with this method and the perceptual distances is equal to 0.67. Therefore, we applied this method to the NOS data in Section 7.5.2 and Chapter 8, and to the RND data in Chapter 9.

### 7.5.2 Analysis of Norwegian

In this section first we apply the method chosen in Section 7.5.1, to our set of 15 varieties. In Table 7.10 the distances are given as percentages. The way in which percentages are found is described in the Sections 5.1.8 and 5.1.10. Given the high correlation between the distances obtained with this method and perceptual distances we may expect that the classification results will be similar as those in Section 7.4.1. Since the distance matrix is symmetric, only one half is used, while the zero values on the diagonal from upperleft to lowerright are not used. In Figure 7.7 a dendrogram is given. In the dendrogram, the scale distance is given as a percentage. In Figure 7.8 a multidimensional scaling plot is shown.

Comparing the dendrogram in Figure 7.7 with the dendrogram obtained on the basis of the perceptual distances (Figure 7.5), both show an Austlandsk group which contains the varieties of Larvik, Halden, Lillehammer and Borre, and a Trøndsk group which contains the varieties of Verdal, Bjugn and Stjørdal. Although the two dendrograms do not cluster the Midlandsk varieties as one group, in the perceptual dendrogram they appear to be more related than in the computational dendrogram. In the perceptual dendrogram the Midlandsk dialect of Lesja is clustered with the Austlandsk varieties, although not very close. In the computational dendrogram this dialect belongs to the Trøndsk varieties. Geographically the variety is located about midway between the two areas. In both the perceptual and computational dendrogram Bø is clustered with the Austlandsk varieties, but in the perceptual dendrogram the relation appears to be stronger. The Sørvestlandsk varieties of Bergen and Time form one (rather loose) cluster in the perceptual dendrogram. In the computational dendrogram they do

| | Ber | Bju | Bod | Bø | Bor | Fræ | Hal | Her | Lar | Les | Lil | Stj | Tim | Tro | Ver |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Bergen | 00.0 | 34.9 | 31.3 | 35.6 | 27.5 | 35.1 | 26.6 | 41.7 | 28.7 | 39.8 | 24.7 | 39.2 | 27.7 | 31.2 | 37.8 |
| Bjugn | 34.9 | 00.0 | 23.2 | 32.1 | 29.4 | 26.1 | 28.4 | 32.6 | 28.1 | 25.9 | 28.5 | 20.0 | 37.0 | 23.8 | 16.9 |
| Bodø | 31.3 | 23.2 | 00.0 | 33.1 | 28.6 | 30.8 | 27.8 | 34.9 | 23.1 | 30.2 | 26.7 | 27.2 | 34.2 | 27.5 | 27.7 |
| Bø | 35.6 | 32.1 | 33.1 | 00.0 | 28.5 | 37.9 | 27.0 | 31.3 | 27.9 | 30.9 | 28.8 | 39.3 | 33.0 | 30.6 | 34.0 |
| Borre | 27.5 | 29.4 | 28.6 | 28.5 | 00.0 | 38.8 | 17.5 | 39.7 | 21.3 | 36.3 | 15.0 | 39.0 | 31.1 | 25.6 | 32.8 |
| Fræna | 35.1 | 26.1 | 30.8 | 37.9 | 38.8 | 00.0 | 36.1 | 31.7 | 35.2 | 29.6 | 37.2 | 28.5 | 37.9 | 33.1 | 29.5 |
| Halden | 26.6 | 28.4 | 27.8 | 27.0 | 17.5 | 36.1 | 00.0 | 39.5 | 14.4 | 33.2 | 11.8 | 37.6 | 31.4 | 22.1 | 30.2 |
| Herøy | 41.7 | 32.6 | 34.9 | 31.3 | 39.7 | 31.7 | 39.5 | 00.0 | 37.7 | 35.4 | 38.1 | 39.7 | 38.3 | 36.7 | 37.1 |
| Larvik | 28.7 | 28.1 | 23.1 | 27.9 | 21.3 | 35.2 | 14.4 | 37.7 | 00.0 | 32.9 | 15.2 | 35.0 | 31.6 | 23.1 | 30.1 |
| Lesja | 39.8 | 25.9 | 30.2 | 30.9 | 36.3 | 29.6 | 33.2 | 35.4 | 32.9 | 00.0 | 32.1 | 24.9 | 35.9 | 34.7 | 29.6 |
| Lillehammer | 24.7 | 28.5 | 26.7 | 28.8 | 15.0 | 37.2 | 11.8 | 38.1 | 15.2 | 32.1 | 00.0 | 35.3 | 29.3 | 23.1 | 31.3 |
| Stjørdal | 39.2 | 20.0 | 27.2 | 39.3 | 39.0 | 28.5 | 37.6 | 39.7 | 35.0 | 24.9 | 35.3 | 00.0 | 42.0 | 32.4 | 25.9 |
| Time | 27.7 | 37.0 | 34.2 | 33.0 | 31.1 | 37.9 | 31.4 | 38.3 | 31.6 | 35.9 | 29.3 | 42.0 | 00.0 | 34.6 | 38.7 |
| Trondheim | 31.2 | 23.8 | 27.5 | 30.6 | 25.6 | 33.1 | 22.1 | 36.7 | 23.1 | 34.7 | 23.1 | 32.4 | 34.6 | 00.0 | 22.6 |
| Verdal | 37.8 | 16.9 | 27.7 | 34.0 | 32.8 | 29.5 | 30.2 | 37.1 | 30.1 | 29.6 | 31.3 | 25.9 | 38.7 | 22.6 | 00.0 |

Table 7.10: Average Levenshtein distances between all pairs of 15 Norwegian dialects given as percentages.

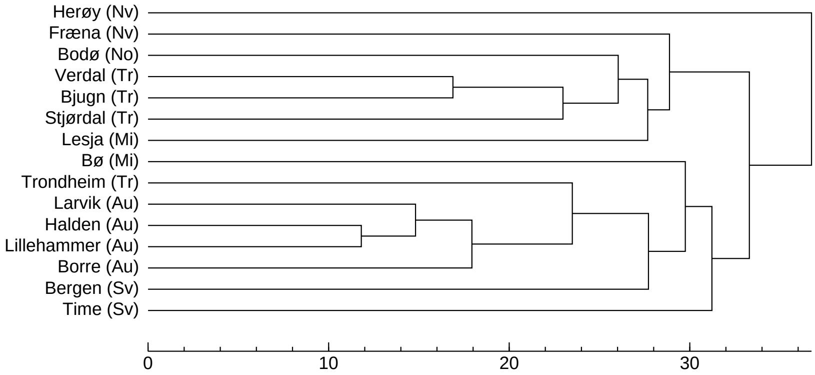


Figure 7.7: Dendrogram derived from the 15 × 15 matrix of Levenshtein distances showing the clustering of (groups of) Norwegian dialects. UPGMA clustering is used (see Section 6.1.2). The scale distance is given as a percentage. The abbreviations between parentheses are explained in Figure 7.2. An Austlandsk and a Trøndsk group can clearly be identified. The tree structure explains 68% of the variance.

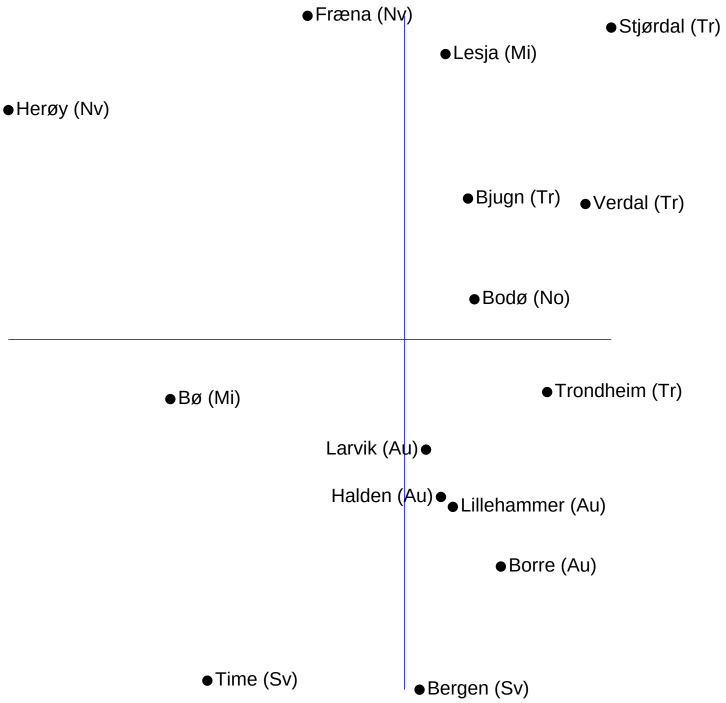


Figure 7.8: Multidimensional scaling of the results derived from the 15 × 15 matrix of Levenshtein distances. Kruskal's Non-metric MDS is used (see Section 6.2.2). The abbreviations between parentheses are explained in Figure 7.2. The y-axis (first dimension) corresponds with the geographic north-south axis, the x-axis (second dimension) more or less with the west-east axis. Two dimensions explain 83% of the variance.

not form one cluster, but appear to be related to some extent (see Section 6.1.4). In both dendrograms the two Nordvestlandsk varieties do not form one cluster. In both Fræna is clustered with the Trøndsk varieties. However, Herøy is clustered with the Sørvestlandsk varieties in the perceptual dendrogram, while in the computational dendrogram it belongs to none of the groups, but appears to be distinct from all the other varieties. In both dendrograms Bodø is clustered with the Trøndsk varieties. However, in the computational dendrogram Bodø looks as if it were closer to the Trøndsk varieties than in the perceptual dendrogram. However, the cluster with Verdal, Bjugn and Stjørdal is geographically not impossible. A striking difference can be found with regard to the dialect of Trondheim, which is clustered with the Trøndsk varieties in the perceptual dendrogram, but in the computational dendrogram it is clustered with Austlandsk varieties. Possibly the listeners recognized the recording of Trondheim as the dialect of Trondheim and let influence their judgments by geography. However, the dialect of larger cities may be in contrast with their surrounding and more related to geographically more distant varieties. We conclude that the two dendrograms are rather similar, due to the fact that especially the closer clusters in the one dendrogram are also found in the other one.

Comparing the multidimensional scaling plot in Figure 7.8 with the multidimensional scaling plot obtained on the basis of the perceptual distances (Figure 7.6), an Austlandsk and a Trøndsk group can be found in both. In the computational plot the Austlandsk varieties are closer than in the perceptual plot. However, in the perceptual plot the Trøndsk varieties are closer than in the computational plot. In the perceptual plot the Trøndsk dialect of Trondheim is most distant from the other Trøndsk varieties. In the computational plot the Trondheim dialect is even more distant to the varieties of the same group. In the perceptual plot the geographically distant Midlandsk varieties of Bø and Lesja are not very close, but in the computational plot they are much more distant. The Nordvestlandsk varieties of Fræna and Herøy are about equally distant in the two plots. In the perceptual plot the Sørvestlandsk varieties of Bergen and Time are closer than in the computational plot. In both plots the Nordlandsk variety of Bodø is found near (perceptually) or among (computationally) the Trøndsk varieties. In the perceptual plot there is a rather sharp division between northern and southern varieties. In the computational plot the northern and southern varieties form a continuum. This may indicate that listeners perceive differences in a more categorical way than the Levenshtein distance suggests. This may also explain the other differences. Since the differences between the plots are relatively small, we conclude that the Levenshtein distances reflect the perceptual distance a great deal.

# Chapter 8

# Measuring Norwegian dialect distances

In Chapter 7 a range of computational comparison methods was validated. The method with the highest score is a variant of the Levenshtein distance, where (i) segment distances are found on the basis of the Barkfilter representation, (ii) four length gradations are used, (iii) diphthongs are represented as a sequence of two segments, and (iv) logarithmic segment distances are used (Section 7.5.1). This method was applied to a small set of 15 Norwegian dialects (Section 7.5.2). In this chapter we apply the same method to a larger set of 55 Norwegian varieties. Results will be compared to the dialect map of Skjekkeland (1997).

In Section 8.1 the set of 55 varieties will be discussed. On the basis of the Levenshtein distances we will perform cluster analysis and multidimensional scaling. In Section 8.2 results of cluster analysis are presented, and in Section 8.3 the results of multidimensional scaling. The discussion of the results should be considered as an initial impetus. Further analysis of the results may be useful future work. In Section 8.4 we draw some conclusions.

## 8.1 Data source

In Section 7.2 we described a database which contains recordings of different Norwegian varieties. The database was compiled by Jørn Almberg and Kristian Skarbø. For each variety a recording and a transcription is given of the fable 'The North Wind and the Sun'. The text consists of 58 words which are given in Appendix B Table B.1.

When the perception experiment was carried out (see Section 7.4.1), recordings of only 15 varieties were available. Later on this database was extended. In this chapter results are presented which are obtained on the basis of a set of 55 varieties. Figure 8.1 shows the geographical distribution of the dialects. The set of 55 varieties covers all nine dialect areas as found on the map of Skjekkeland

(1997). Figure 8.2 shows the distribution of the varieties over the dialect areas as given by Skjekkeland. For some locations more than one recording and transcription was available. Therefore, these locations are numbered in the figures in this chapter.[1]

- Alstahaug
  The two versions are based on different recordings of different informants, the first from Sandnessjøen (Alstahaug 1) and the second from Tjøtta (Alstahaug 2). The first version is most representative for the area of Alstahaug.

- Bergen
  The two versions are based on different recordings of the same informant. The older version (Bergen 1) is no longer available on the web, but was used in validation work (see Section 7.2). The newer version (Bergen 2) is the better one according to the speaker.

- Bodø
  The two versions are based on different recordings of the same informants. The older version (Bodø 1) is no longer available on the web, but was used in validation work. The newer version (Bodø 2) is the better one according to the speaker.

- Rana
  The two versions are based on different recordings by different informants, both from Rana (Rana 1 and Rana 2). The second version is more representative for the area around Rana.

- Stavanger
  The two versions are based on different recordings by different informants, the first from Hafrsfjord (Stavanger 1) and the second from Hundvåg (Stavanger 2). Both are equally representative for the surrounding of Stavanger, but when we are forced to make a choice, we select the second version.

- Stjørdal
  The three versions are based on different recordings of different informants, the first and the second from Stjørdal (Stjørdal 1, Stjørdal 2), and the third from Stjørdalshalsen (Stjørdal 3). The first version is most representative. In validation work the second version is used, which was available earlier.

[1] We are grateful to Jørn Almberg (personal communication) for advice at several points below, e.g., the question as to which of two versions is the more typical for a given site.

- Time
  The two versions are based on different recordings of different informants, the first from Bryne (Time 1) and the second from Undheim (Time 2). The first version is most representative for the area of Time. This version was also used in validation work.

## 8.2 Classification

### 8.2.1 Cluster analysis

Using the Levenshtein variant we mentioned at the beginning of this chapter, we calculated the distances between the 55 varieties. On the basis of these distances we applied cluster analysis (see Section 6.1). In Figure 8.3 a dendrogram is given, showing the classification of 55 Norwegian varieties. In the dendrogram, the scale distance shows percentages.

Examining the nine most significant groups we find from upper to lower the dialects of Herøy and Fræna, a central group, the dialect of Bø, an eastern group, a southeastern group, a northern group, a western group and a southwestern group. The same groups are geographically visualized in Figure 8.4.

When regarding only the 5 most significant groups, Herøy and Fræna appear to be one cluster. Both varieties belong to the Nordvestlandsk varieties. However they are not clustered with the other Nordvestlandsk varieties, which are found in the western group. When considering the 9 most significant groups, each of these two varieties appears to be a separate dialect, not clustered with any of the other groups. This indicates that the two varieties are very marked dialects among the other Nordvestlandsk varieties.

The *central* group contains the Trøndsk varieties of Sunndal and Oppdal and the geographically rather close Midlandsk variety of Lesja. It is striking that Sunndal and Oppdal are not clustered with the other Trøndsk varieties, which are for the greater part found in the northern group. We expected that Lesja would be clustered with the other Midlandsk variety of Bø. However, just as for the set of 15 varieties, this is not the case. Geographically the two varieties are distant. The variety of Bø appears to be a separate variety which does not belong to any of the other varieties. It is striking that Bø is suggested to be closest to the eastern varieties, and not with the geographically closer varieties of the southeastern group.

The Austlandsk varieties are divided into an *eastern* group and a *southeastern* group. In the southeastern group the geographically adjacent Sørlandsk varieties are found as well. More striking is the presence of the Sørvestlandsk variety of Bergen and the Trøndsk variety of Trondheim in this group. We cannot explain this. However, it is not uncommon that varieties of larger cities are dialect islands

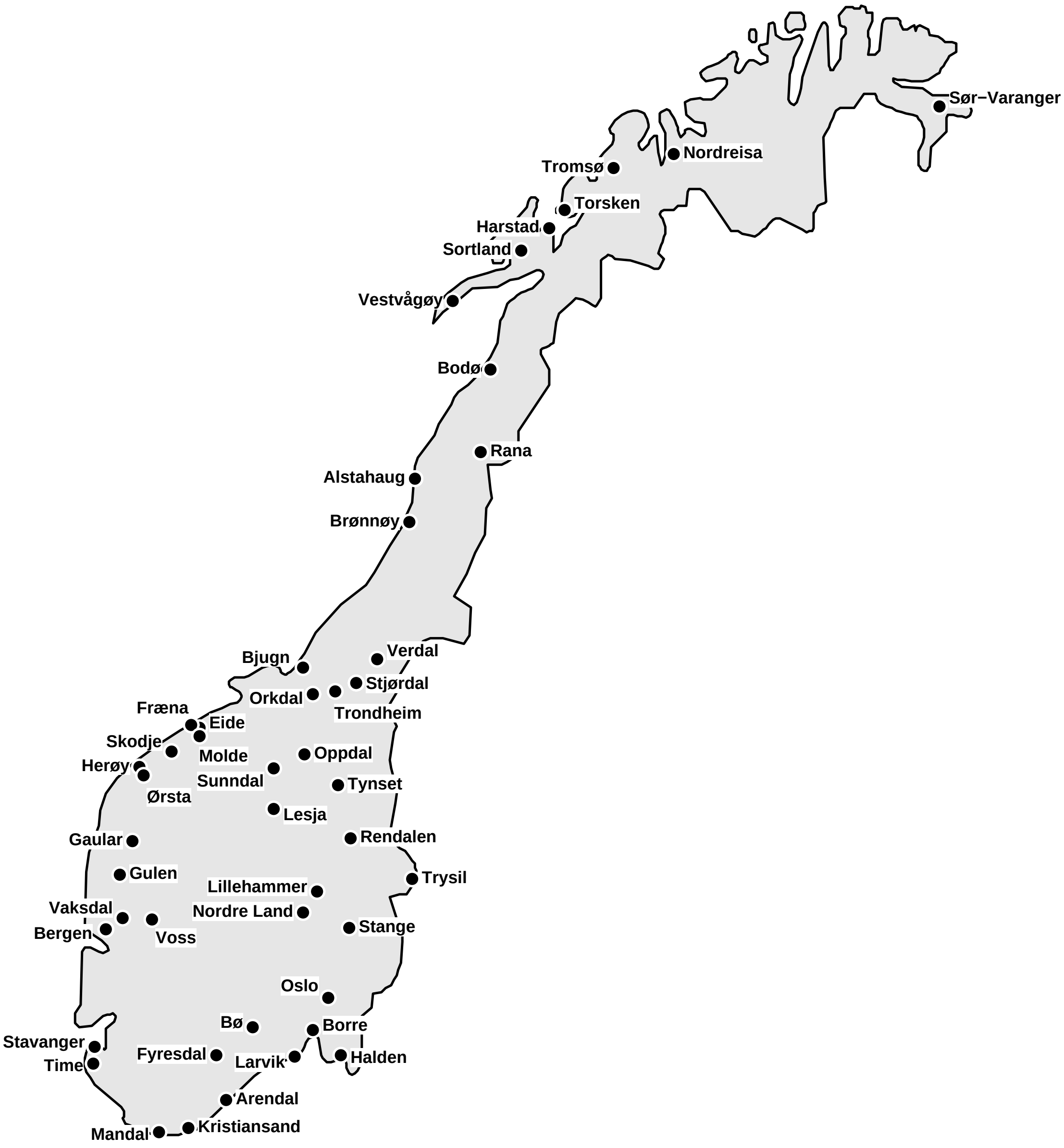


Figure 8.1: The geographic distribution of the 55 Norwegian varieties.

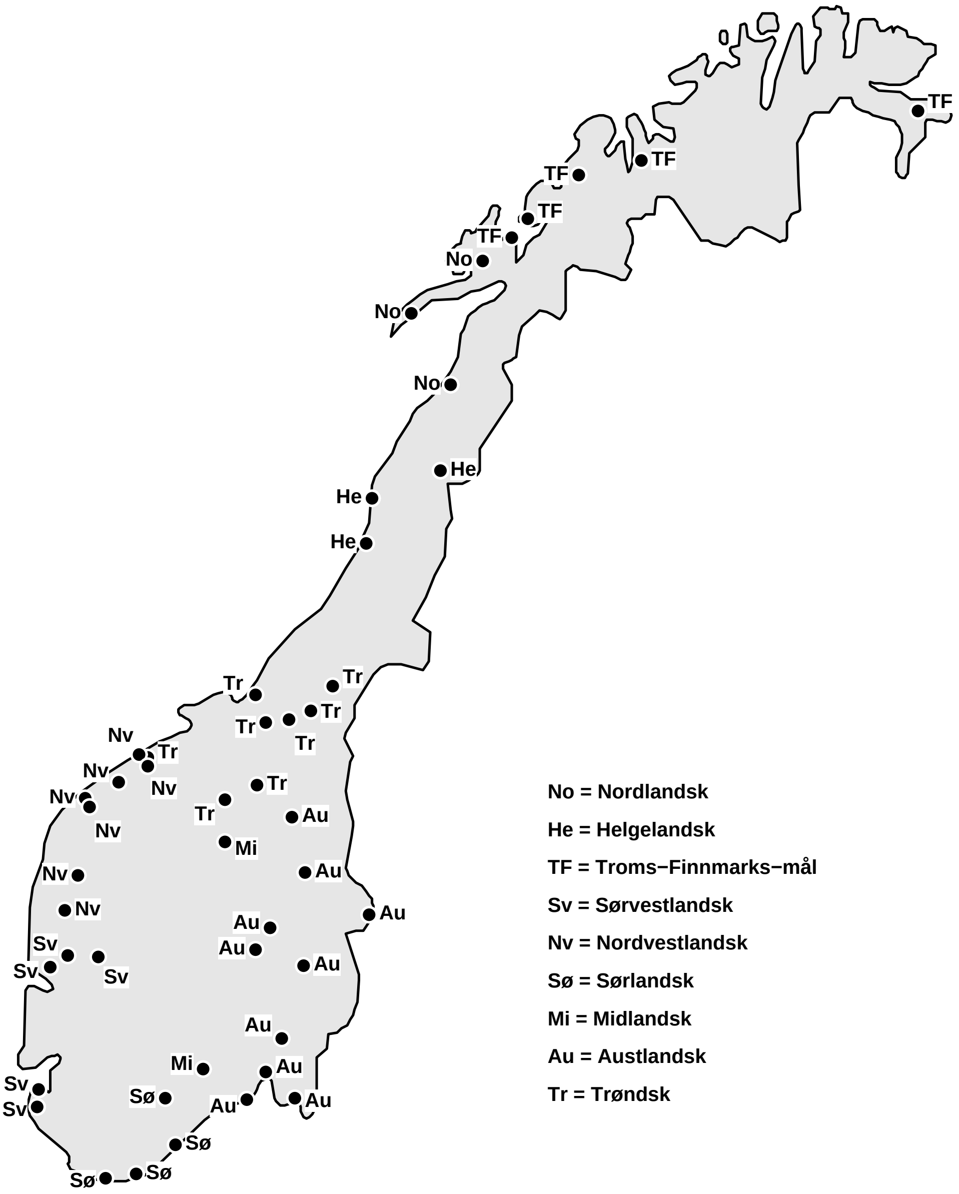


Figure 8.2: According to Skjekkeland (1997) the Norwegian language area can be divided in nine groups. The data points on this map correspond with those in Figure 8.1. In the set of 55 varieties all dialect areas are represented. The same abbreviations are used in the other figures in this chapter.

which are related to geographically more remote dialects.[2] Rather unexpected is that the varieties of Tynset and Lillehammer are in the southeastern group, and not in the eastern group. Lillehammer is closest to Halden. Both Halden and especially Lillehammer are nearly standard (i.e., close to *bokmål*), and therefore not very typical dialect versions from their respective geographic regions. The reading of Tynset is also quite standard, which may be the reason why it is judged to be closer to Oslo. In the southeastern group two versions of Bergen, which are recorded by the same informant, do not form one cluster, but are rather close.

The largest group in the dendrogram is the *northern* group. It contains Nordlandsk, Helgelandsk, Troms-Finnmarks-mål and Trøndsk. The group may be divided in a Trøndsk group on the one hand, and a group containing the other varieties on the other hand. In the latter group, no systematic division between Nordlandsk, Helgelandsk and Troms-Finnmarks-mål varieties can be found. Perhaps the division in these three areas has become blurred over time. The two varieties of Rana are rather close, although they do not form one cluster. The varieties of Alstahaug are obviously more distant, indicating dialect diversity in a small area. Stjørdal 1 and 2 are rather close. Compared to these two varieties Stjørdal 3 is relatively distant, indicating again strong variation in a small area. The two versions of Bodøy are recorded by the same person. They neatly form one cluster.

In the *western* group Nordvestlandsk varieties are mainly found. The adjoining Sørvestlandsk varieties of Vaksdal and Voss are in this group as well. More surprisingly is that the Sørlandsk dialect of Fyresdal is also in this group. It would be more fitting if this dialect were clustered with other Sørlandsk varieties. We cannot explain this. In the dendrogram the Sørlandsk varieties cannot be found as a group. The *southwestern* group mainly contains Sørvestlandsk varieties. The geographically adjacent Sørlandsk variety of Mandal is also in this group. The two varieties of Time neatly cluster together, just as the two varieties of Stavanger.

### 8.2.2 Area map

In the map in Figure 8.4 we treated some varieties as dialect islands, i.e. their color was not expanded to their surrounding by triangulation. The varieties of Herøy, Fræna and Bø do not belong to groups, instead they are treated as dialect islands. The varieties of Trondheim, Tynset, Lillehammer and Bergen were clustered with the southeastern varieties. As stated above, this is unexpected since they are geographically rather distant from the other varieties in the southeastern group and found among varieties of other groups. Therefore, we dealt with them as language islands. Finally the classification of Fyresdal with

[2] Compare, e.g., the town Frisian dialect islands in the Frisian dialect area on the map of Daan and Blok (1969).

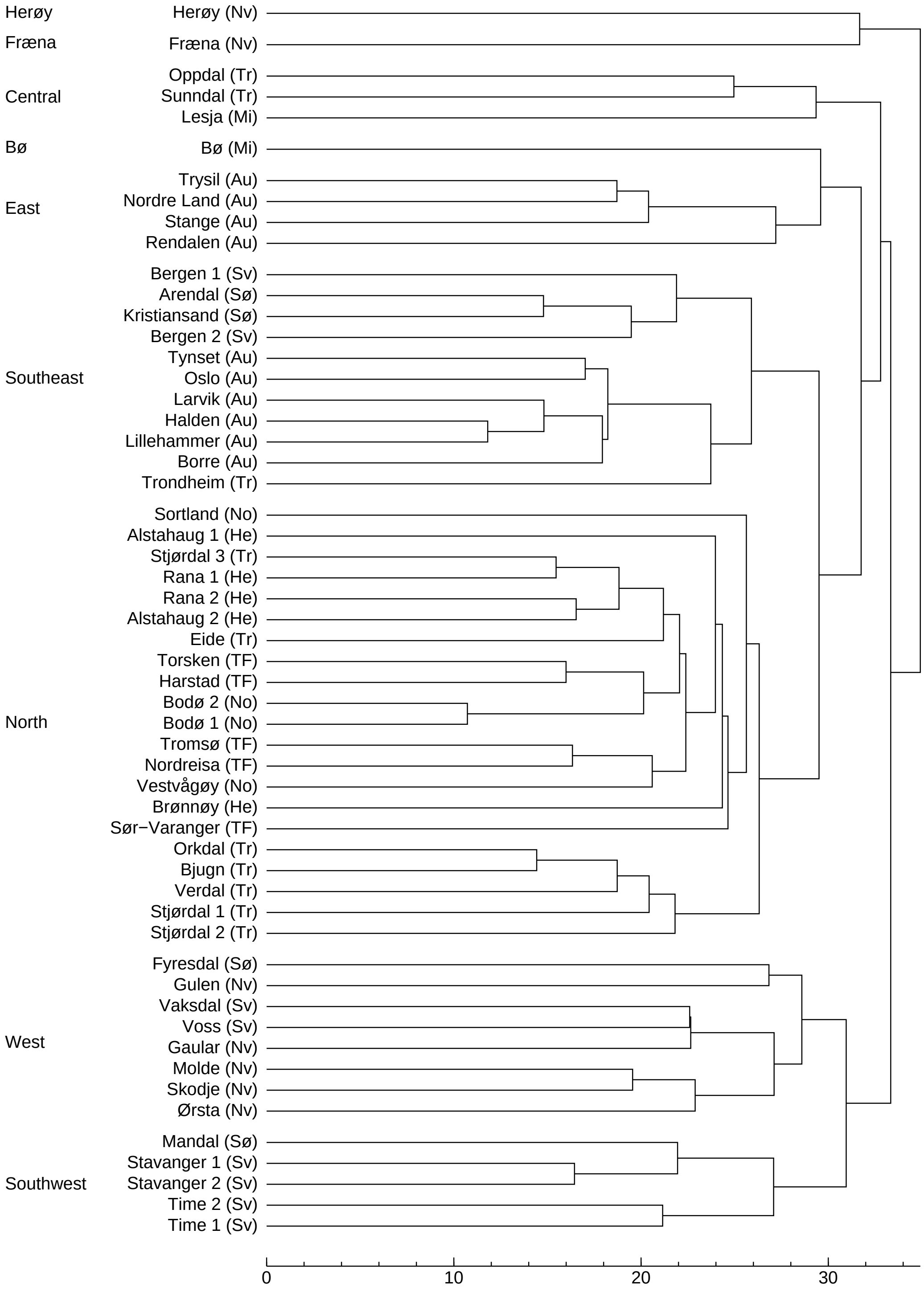


Figure 8.3: Dendrogram derived from the 55 × 55 matrix of Levenshtein distances showing the clustering of (groups of) Norwegian dialects. UPGMA clustering is used (see Section 6.1.2). The scale distance shows percentages. The abbreviations between parentheses are explained in the caption to Figure 8.2. The nine most significant groups are labeled and geographically visualized in Figure 8.4. The tree structure explains 48% of the variance.

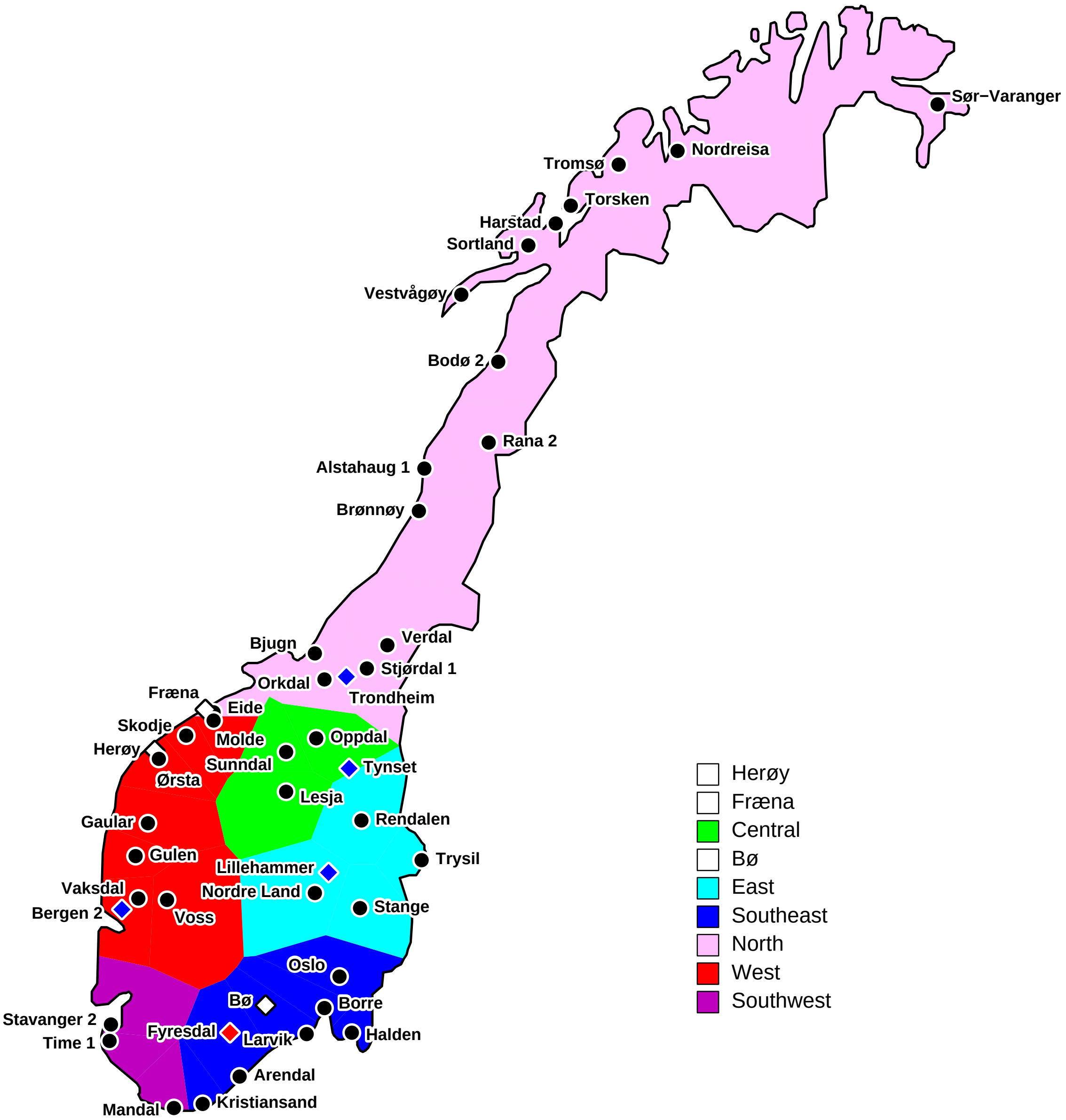


Figure 8.4: The nine most significant groups as derived from the dendrogram in Figure 8.3. UPGMA clustering is used (see Section 6.1.2). Varieties treated as dialect islands are marked with a diamond. The dialects of Herøy, Bø and Fræna are marked with a white diamond. Each of them is a separate group.

the western group is unexpected as stated above. The variety is treated as a language island. In this way we get a clear division into six areas. For those locations for which more than one transcription was available, we selected only the most reliable transcription, or the transcription which is most representative for the surroundings of the location (see Section 8.1 for more details).

When comparing our map in Figure 8.4 with the map of Skjekkeland (1997) in Figure 8.2 it is most striking that the Midlandsk group on the map of Skjekkeland is not found on our map. As explained above, the fact that Lesja and Bø do not form one group may be explained by the fact that they are geographically rather distant. The northern part of the Austlandsk group corresponds with our East group. Our Southeast group covers the southern part of the Austlandsk varieties and the eastern part of the Sørlandsk varieties. Our Southwest group corresponds with the southern part of the Sørvestlandsk varieties, but also the Sørlandsk variety of Mandal is in this group. Our West group includes the Nordvestlandsk varieties. However on our map the southern border is shifted to the South. Our central group covers the southern part of the Trøndsk group. The northern Trøndsk varieties, the Helgelandsk varieties, the Nordlandsk varieties and the Troms-Finnmark-mål are found as a North group on our map, covering a large geographic area.

## 8.3 Continuum

### 8.3.1 Multidimensional scaling

We also applied multidimensional scaling to the Levenshtein distances between the 55 Norwegian varieties (see Section 6.2). When applying this classification technique we found that one dimension explains 55% of the variance, two dimensions 79%, three dimensions 89%, four dimensions 91%, five dimensions 93%, six dimensions 94% and seven dimensions 95%. Using more than three dimensions only a small improvement of the explained variance is obtained. Therefore, we regard the three-dimensional solution which is shown in Figure 8.5.

The y-axis represents inversely the first dimension. At the top the central Trøndsk varieties are found, and at the bottom the southern Sørlandsk varieties. This accords with geography. However it is striking that the northern Helgelandsk, Nordlandsk and Troms-Finnmark-mål varieties and some Austlandsk varieties are found about intermediate between the Trøndsk and Sørlandsk varieties, which does not agree with geography. This suggests that the northern varieties are more related to the southern varieties than might be expected on the basis of simple geographical distance. The x-axis represents the second dimension. For the southern varieties a division in western and eastern varieties can be found. On the left the western Nordvestlandsk and Sørvestlandsk varieties are found, and on the right the eastern Austlandsk and most Sørlandsk varieties are found. For the central and northern varieties we found no clear division in West and East. The grey tones represent the third dimension. The Austlandsk varieties of Fyresdal, Bø, Nordre Land, Stange, Trysil and Rendalen are represented by black dots (low values), the other Austlandsk varieties by darker grey dots, most remaining varieties by lighter grey dots, and the Nordlandsk variety of Sortland by a white dot (high value).

To get insight into the relation between the variation per dimension on the one hand, and variation in word pronunciations on the other hand, first we calculated distances between varieties per dimension. When two varieties have respectively the values $x$ and $x'$ in a dimension, the distance is equal to $|x - x'|$. In this way a distance matrix of $(55 \times 54)/2$ distances is obtained per dimension. Subsequently Levenshtein distances are calculated on the basis of the pronunciations of a single word. In this way we get 58 matrices for 58 words, each containing $(55 \times 54)/2$ distances. The distances which we calculated per dimension are correlated with the distances of each of the 58 matrices. In this way we found the strongest correlating word per dimension.

It appears that the distances in the first dimension correlate most strongly with distances obtained on the basis of pronunciations of the word *mann* 'man' ($r = 0.55$). In the northern and central varieties this word is mostly pronounced as [$^{1}$mɑɲː] while it is usually pronounced as [$^{1}$mɑnː] in the southern varieties. They differ by the last segment: [ɲ] versus [n]. Distances in the second dimension correlate most strongly with distances obtained on the basis of pronunciations of the word *enige* or *samde* 'agreed' ($r = 0.52$). In the western varieties forms like [$^{2}$ʔeːnig], [$^{2}$eːnig], and [$^{2}$æinigə] are used. In the eastern varieties mostly forms like [$^{2}$ʔeːni], [$^{2}$eːni], and [$^{2}$eːniə] are found. So in the western varieties a [g] is pronounced, but in the eastern varieties the [g] is elided. Furthermore, only in Herøy did we find the form [$^{2}$sɑmdə]. Distances in the third dimension correlate most strongly with distances obtained on the basis of pronunciations of the word *kven* or *hven* 'who' ($r = 0.44$). In Rendalen and Stange the word is pronounced as [$^{2}$ɔkːən], in Nordre Land and Bø as [$^{2}$hɔkːən], and in Trysil as [ˌhøkːən]. In the other varieties forms like [$^{1}$k$^{h}$ɛmː], [$^{1}$k$^{h}$æn] and [$^{1}$ʋemː] are used. Lexical differences are represented by this word. Per dimension we also examined other strongly correlating words. However we found no system in the phenomena which causes differences between the pronunciations.

### 8.3.2 Continuum map

In Section 6.2.4 we explained that on the basis of three dimensions of the three-dimensional solution each variety can be represented by a color. The three dimensions are mapped to the intensities of red, green and blue. We used this approach to create a map in which each variety get its own unique color. We assigned the colors to the three dimensions so that the different areas can be recognized rather clearly. The first dimension represents inversely the intensity of green, the second

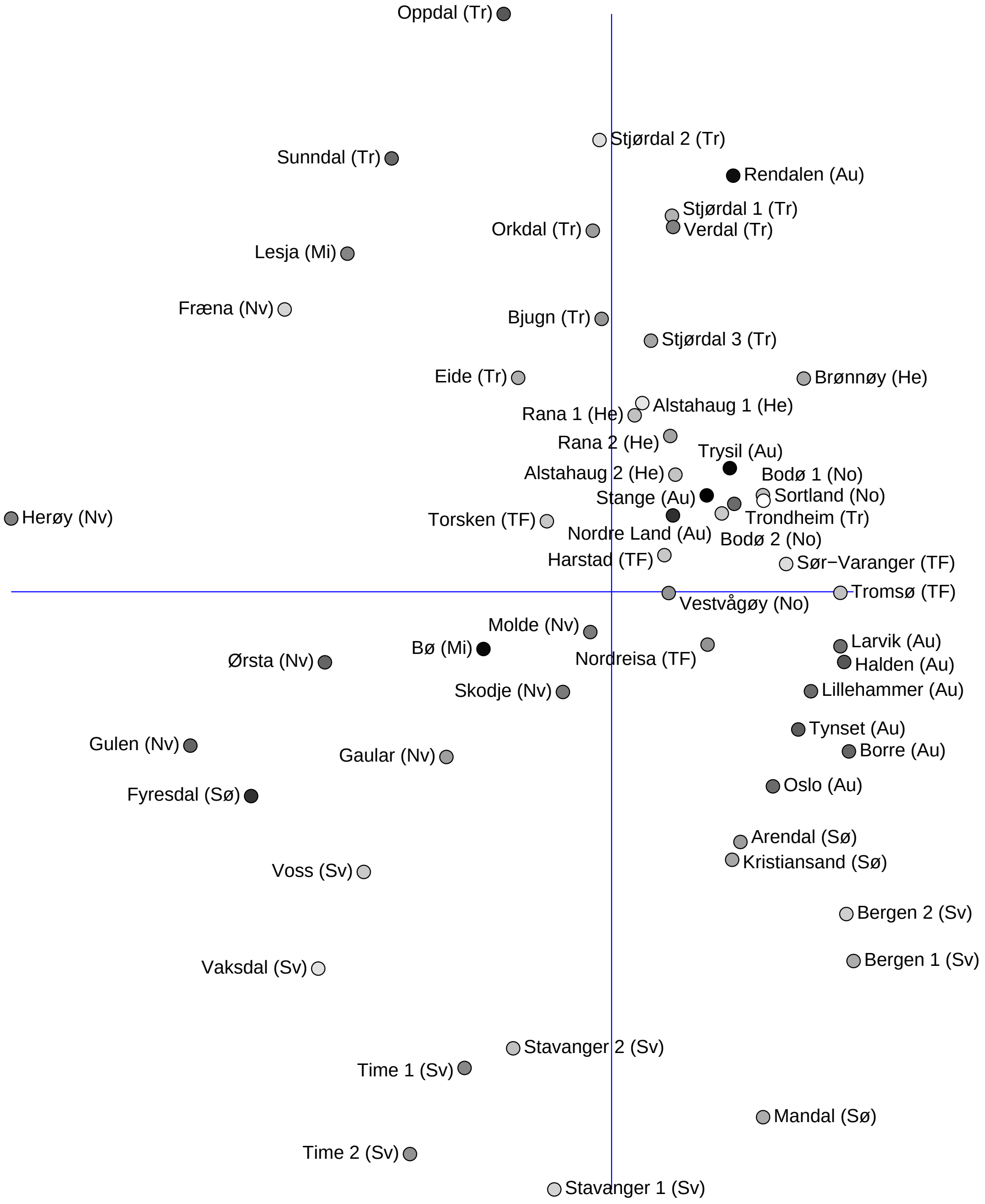


Figure 8.5: Multidimensional scaling of the results derived from the 15 × 55 matrix of Levenshtein distances. Kruskal's Non-metric MDS is used (see Section 6.2.2). The abbreviations between parentheses are explained in Figure 8.2. The y-axis represents inversely the first dimension, the x-axis represents the second dimension and grey tones the third dimension. Three dimensions explain 89% of the variance.

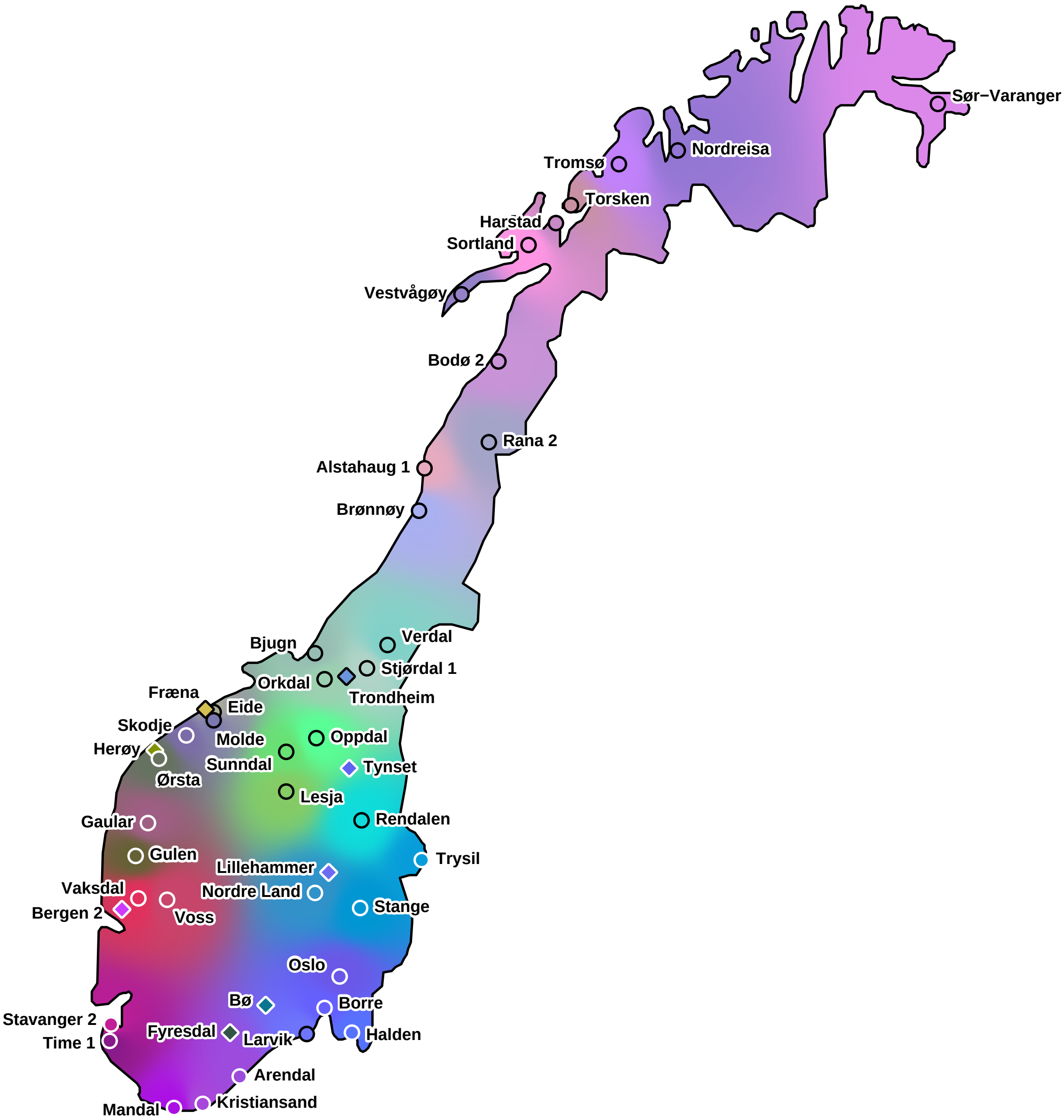


Figure 8.6: Dialect variation represented by color variation. The first MDS dimension is mapped inversely to green, the second is mapped to blue and the third to red. Kruskal's Non-metric MDS is used (see Section 6.2.2). The color of intermediate points is determined by interpolation using Inverse Distance Weighting. Dialect islands are marked with a diamond. They are not involved in the interpolation process.

dimension represents the intensity of blue, and the third dimension the intensity of red.

In Figure 8.6 a color map based on three MDS dimensions is shown. In the map the space between the points is colored on the basis of MDS values which are found by interpolation using Inverse Distance Weighting (see Section 6.2.4). In this way the dialect landscape is represented as a continuum. Dialect islands are excluded from the interpolation process. The same varieties are treated as dialect islands as in the map in Figure 8.4. Where two or more transcriptions were available for the same location, just as in Figure 8.4 we selected only the most reliable one or the one which is most representative for its surroundings (see Section 8.1 for more details).

Comparing this map with the map of Skjekkeland (see Figure 8.2) we see that the eastern blue area corresponds with the Austlandsk varieties. Examining the blue shades in more detail, we find three areas. In and around Rendalen we find greenish blue. A more pure blue area is represented by Trysil, Nordre Land and Stange. Darker blue is found in and around the varieties of Oslo, Borre, Halden and Larvik. The purple area in the furthest south corresponds with the Sørlandsk varieties. In the West we find an area varying from red in the North to red-purple more south. The area corresponds with the Sørvestlandsk varieties. In the Northwest different colors are found, illustrating that the Nordvestlandsk varieties do not form a homogeneous group. The southern Trøndsk varieties of Oppdal and Sunndal and the Midlandsk variety of Lesja represent a central green area. This area is not found on the map of Skjekkeland (1997). In and around most northern Trøndsk varieties we find a blue-green color. The Helgelandsk, Nordlandsk and Troms-Finnmarks-mål varieties are represented by different shades of purple. However the three different groups are not clearly distinguished. The purple shades suggest a strong relation with the Sørlandsk varieties in the furthest south. In our map we found no Midlandsk area.

## 8.4 Conclusions

Examining the dendrogram in Figure 8.3, the map in Figure 8.4, the multidimensional scaling plot in Figure 8.5 and the map in Figure 8.6 we found some minor and some major differences compared to the map of Skjekkeland (1997). We explain this by two factors. First the text ‘the North Wind and the Sun’ is a rather short text. We are not sure of the extent to which the translations of this text are representative pictures of the varieties. In Bolognesi and Heeringa (2002) a word list is used where the words are randomly chosen from a corpus. In that way the data will be more representative. The classification as given in this section may not interpreted as *the* classification of Norwegian dialects, but as *one* classification which only reflects the variation in the translations of the fable ‘the North Wind and the Sun’. Second the map of Skjekkeland (1997) is based on a restricted number of phenomena. Possibly the map may reflect the historical situation to some extent. To get more clarity about this, it would be interesting if a new map were created on the basis of the arrow method, just as was done by Daan and Blok (1969) for the Netherlandic part of the Dutch language area.

When comparing the map in Figure 8.4 with the map of Skjekkeland, we found some similarities, but also a lot of differences. In our opinion the map in Figure 8.5 is much more like the map of Skjekkeland. Figure 8.4 is based on the dendrogram in Figure 8.3. This dendrogram explains 48% of the variance. Figure 8.6 is based on the multidimensional scaling plot in Figure 8.5. This plot explains 89% of the variance. Therefore, we judge the map in Figure 8.6 to be more reliable than the map in Figure 8.4.

# Chapter 9

# Measuring Dutch dialect distances

On the basis of a small data set of Norwegian varieties we validated a range of different computational methods in Chapter 7. We found that the method with the highest score is a variant of the Levenshtein distance, where (i) segment distances are found on the basis of the Barkfilter representation, (ii) four length gradations are used, (iii) diphthongs are represented as a sequence of two segments, (iv) and logarithmic segment distances are used (Section 7.5.1). This method was applied to a larger set of Norwegian varieties in Chapter 8. In this chapter we apply the same method to Dutch dialects. We use data from the *Reeks Nederlandse Dialectatlassen* (RND), a series of Dutch dialect atlasses which were edited by Blancquaert and Peé in the period 1925–1982. The goal of this chapter is to show that the application of the Levenshtein distance to the Dutch material of the RND gives interesting and useful results, despite the shortcomings of the RND transcriptions.

In Section 9.1, the RND will be discussed in more detail. In Section 9.2 the selected variant of the Levenshtein distance is applied to this data source, and the resulting distances are discussed. On the basis of these distances, the dialects are classified. Results of cluster analysis are presented in Section 9.3, where a main classification is given. In Section 9.4 each of the groups that are found in the main classification is discussed in more detail. In Section 9.5, results of multidimensional scaling are given. Using this technique the Dutch language area may be viewed as a continuum. The Dutch dialects are also compared to Standard Dutch. A ranking of difference with respect to Standard Dutch is given in Section 9.6. In Section 9.7 we draw some conclusions.

## 9.1 Data source

The *Reeks Nederlandse Dialectatlassen* (RND) is a series of atlasses covering the Dutch dialect area. The Dutch dialect area comprises the Netherlands, the northern part of Belgium, a smaller northwestern part of France and the German county Bentheim. The atlas series consist of 16 parts. Although the Dutch language area consists of 16 provinces as well, the 16 volumes do not exactly correspond with the 16 provinces. In 1925, the first volume appeared, compiled by Blancquaert. The volume contains transcriptions of dialects in Klein-Brabant. The first recordings for this volume were already made in 1921 or 1922 (Goossens, 1997). After this, Blancquaert initiated a project in which recordings were made of varieties in the whole Dutch language area. To speed up the progress, Blancquaert engaged several collaborators. Unfortunately, Blancquaert died before all the volumes were finished. Peé was his successor and finished the project. The last recordings were made by Entjes in 1975. These recordings are found in part 14 (Zuid-Drenthe and Noord-Overijssel), that appeared in 1982 as a last installment.

### 9.1.1 Words

In the RND, the same 141 sentences are translated and transcribed in phonetic script for each dialect. Blancquaert mentions that the questionaire was conceived as a range of sentences with words that illustrate particular sounds. The design saw to it that, e.g., possible changes of old-Germanic vowels, diphthongs and consonants are represented in the questionaire. Morphologic and syntactic phenomena are also represented in the sentences (Blancquaert, 1948, p. 13). Since digitizing the phonetic texts is time-consuming on the one hand, and since the Levenshtein distance is a word-based method on the other hand, we selected only 125 words from the text. These words were digitized for each dialect and used as input for the Levenshtein distance. The words represent (nearly) all vowels (monophthongs and diphthongs) and consonants. Also the consonant combination [sx] is represented, which is pronounced as [sk] in some dialects and as [ʃ] in some other dialects. The words are listed in Appendix B Table B.3.

Since the RND transcriptions consists of sentences, the same word may vary lexically in different dialects. We digitized and processed only forms which were semantically equal, as far as we could judge. Among different lexical nouns it may appear that one form has a determiner, and another has not. We always left out the determiner. In a sentence, assimilation phenomena can be found. So these phenomena are also found in the word transcriptions which we cut from the sentences. When two succeeding words in a sentence were not separated by a space, we tried to find the border between the two words by comparing the transcription with the transcriptions of other nearby varieties in which the words were separated by a space, or by comparing with comparable transcriptions in the same text. If we found that the last segment of a word was shared with the first segment of the next word, we included the segment in both words when cutting them from the sentence. E.g. we split [ɔsəbludrɪŋkən] in [ɔsəblud] ‘ox-blood’ and [drɪŋkən] ’drink’.

Sometimes a few of the expected words are missing for certain varieties, e.g. as the consequence of a free translation of some sentences. When two varieties are compared, and for one of the 125 words a translation is missing in one variety or in both varieties, the word is not taken into account in the calculation of the distance (see Section 5.1.10.1). For some words, more than one pronunciation was given, since e.g. an older and a newer form may be in circulation simultaneously. In these cases, the mean distance over the variants of one word is used for calculating the distance (see Section 5.1.10.2).

An extended discussion about the selection of words from the RND can be found in Heeringa (2001). The detailed presentation and discussion of the data is not repeated here, since we focus here on the analysis. The digitized data is publically available at `http://www.let.rug.nl/~heeringa/dialectology/atlas/` with the kind permission of the publisher, De Sikkel.[1]

### 9.1.2 Varieties

The RND contains transcriptions of 1956 Dutch varieties. It would be very time-consuming to digitize all transcriptions. Therefore, we made a selection of 360 dialects. When selecting the dialects the goal was to get a net of evenly scattered dialect locations. First, we selected all locations in the RND which have more than 5,000 and fewer than 10,000 inhabitants. In smaller locations the dialect may be less stable due to moving or deaths, while in larger towns there may exist more than one dialect. Where the density remained too low, smaller locations were also used. Where an irregular pattern arose, the larger locations were sometimes replaced by smaller ones. A denser sampling resulted in the areas of Friesland and Groningen, and in the area in and around Bentheim.

In the map of Hof (1933, p. 14a) the Frisian area is divided into Bildts, Woudfries, Zuidhoeks and Stellingwerfs.[2] The selection was adjusted so that each of these groups was represented. A special group of Frisian varieties that do not form one geographical area are the ‘town Frisian’ varieties (*Stad(s)fries*). Just as the dialect of *het Bildt*, town Frisian dialects may be regarded as an intermediate form of Dutch and Frisian. Town Frisian varieties are spoken in Midsland, Dokkum, Harlingen, Franeker, Leeuwarden, Bolsward, Sneek, Heerenveen, Staveren and on the island of Ameland. In the map of Daan and Blok (1969) the town Frisian locations appear as language islands in the ‘pure’ Frisian language continuum. All of these locations are included in our data set. The map of Daan suggested that

[1] Later on this publisher was taken over by De Boeck, Antwerpen.

[2] A clearer print of this map is found on the cover of the thesis of Breuker (1993).

the variety of Kollum belongs to Kollumerlands. In our results it will appear that this variety belongs rather to the town Frisian varieties. Therefore, we also regard Kollum as a town Frisian language island. The geographical island of Ameland is represented by Hollum and Nes. Since the collection of 'pure' Frisian locations have the same density as the sample in the remaining part of our study, and the town Frisian locations are added to them, we get a relatively higher density in Friesland.

For the Frisian locations of Appelscha, Donkerbroek and Tjalleberd two texts are given in the RND. Appelscha is located in the *Stellingwerf* area. In addition to a Low Saxon *Stellingwerf* variety, a Frisian variety is spoken, introduced by Frisian laborers who moved to Appelscha at the time of peat-diggings. We process the Frisian variety as a language island. Below, 'Appelscha 1' refers to the *Stellingwerf* variety and 'Appelscha 2' to the Frisian variety. A Frisian and a *Stellingwerf* variety are also spoken in Donkerbroek. In the map of Daan and Blok (1969) it can be seen that the river Kuinder (or Tjonger) is the boundary between Frisian (west) and Low Saxon (east). Since Donkerbroek is located west of this river, we regarded the Frisian variety as part of the Frisian language continuum and the *Stellingwerf* variety as a Low Saxon language island in the Frisian language continuum. Below 'Donkerbroek 1' refers to the Frisian variety and 'Donkerbroek 2' to the *Stellingwerf* variety. In Tjalleberd, most people spoke Frisian when the RND recordings in Friesland were made. However, a minor part of the population spoke Tjalleberds (or 'Gietersk'), a variety introduced by peat laborers from northwestern Overijssel (Giethoorn and surroundings). We process the Tjalleberd variety as language island. The Frisian variety is referred to as 'Tjalleberd 1', and the *Stellingwerf* variety as 'Tjalleberd 2'.

In Reker (1993, p. x) the province of Groningen is divided in West-Groningen, North-Groningen, *Oldambt*, *Westerwolde*, *Veenkoloniën*, and the city of Groningen. The northern part of the province of Drenthe, south of the province of Groningen, is not displayed on this map. However, the varieties of this area are strongly related to the Groningen varieties. Because of personal interest, relatively more varieties are chosen in Groningen and North-Drenthe. The varieties are chosen so that each of the different areas is represented.

In part 14 (Zuid-Drenthe and Noord-Overijssel), varieties of the German county Bentheim are included. In a study about Dutch-German contact in and around Bentheim, the German transcriptions on the one side, and the Dutch transcriptions on the other side of the Dutch-German border are used (Heeringa et al., 2000). Since the same varieties are used in the present study, a higher density is found in and around Bentheim.

The RND includes also some varieties in the Belgium province of Luik. Just as the varieties in Bentheim, these varieties do not belong to the Dutch language area. The dialects are found south of the Dutch province of Limburg in the northeastern part of the Belgium province of Luik. We selected Aubel and Baelen which belong to the French language area, and Eupen and Raeren which belong to the German language area.

The geographical distribution of the 360 Dutch varieties is shown in Figure 9.1. Since in Appelscha, Donkerbroek and Tjalleberd two varieties are spoken, the map shows only 357 localities. We divided the Dutch language area in a northwestern, northeastern, southwestern and southeastern part. Each of these parts are visualized in more detail in the Figures 9.10, 9.13, 9.15, 9.17, 9.19, 9.21, 9.22 and 9.24. More about the selection of varieties from the RND can be found in Heeringa (2001).

To be able to compare the varieties with respect to Standard Dutch, we also added a transcription of Standard Dutch. To assure consistency with the existing RND transcriptions, the Standard Dutch transcription is based on the *Tekstboekje* of Blancquaert (1939). However, we transcribed words such as *komen*, *rozen* and *open* as [koˑmə], [roːzə] and [oˑpə]. In the *Tekstboekje* of Blancquaert these words would end on an [n], just as suggested by the spelling. For more details see Heeringa (2001).

### 9.1.3 Groups

The most recent dialect map of the Dutch language was published in 1969 and compiled by Jo Daan (Daan and Blok, 1969). The map was already mentioned in Section 2.2.1 where we discussed the arrow method. With the arrow method, dialect borders are found on the basis of the perception of the dialect speakers. In this map the Dutch language area is divided into 28 different groups. The groups are mentioned in Table 9.1. The map in Figure 9.2 shows the classification of our set of varieties according to the map of Daan. For 49 borderline cases we found it unclear to which of the groups they belong. We left them out, so the map is eventually based on 311 varieties. In this set, 26 of the 28 groups are represented. Not represented are Daan's groups 8 and 16. Three small groups were represented by only one variety, namely group 2 (Egmond aan Zee), group 4 (Koog aan de Zaan) and group 12 (Geraardsbergen). In Section 9.3 we will compare the results of the Levenshtein distance to the classification that is given in the map of Daan.

### 9.1.4 Consistency

As mentioned above the RND consists of 16 different parts. The recordings were made during a time interval of more than 50 years. Therefore, differences in pronuncation may be in some cases the result of differences in time. The volumes of two adjacent areas never differ by more than 30 years, so the effect of temporal difference is mainly found when comparing varieties which are geographically more distant. Sometimes a rather large interval per volume was found as well. We

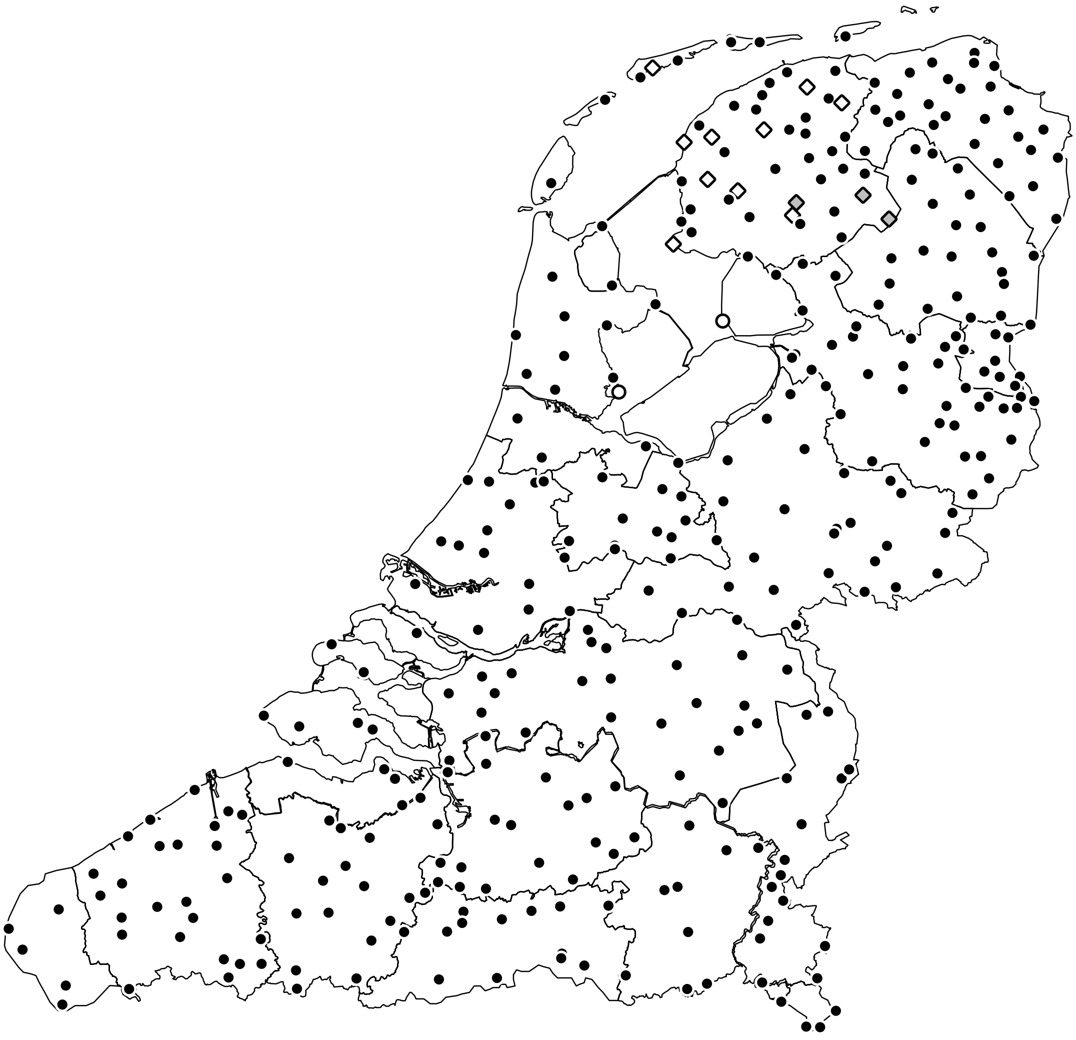


Figure 9.1: Distribution of the 357 localities, corresponding with 360 different varieties. White diamonds represent language islands and grey diamonds represent localities with two dialects where one of the two dialects is a dialect island. Circles represent small geographic islands. In the Figures 9.10 through 9.24 the different parts of the Dutch language area of shown in more detail.

| | |
|---|---|
| 1 | Dialect of Zuid-Holland |
| 2 | Dialect of Kennemerland |
| 3 | Dialect of Waterland |
| 4 | Dialect of Zaan region |
| 5 | Dialect of northern Noord-Holland |
| 6 | Dialect of the province of Utrecht and the Alblasserwaard region |
| 7 | Dialect of Zeeland |
| 8 | Dialect of region between Holland and Brabant dialects |
| 9 | Dialect of West Flanders and Zeeuws-Vlaanderen |
| 10 | Dialect of region between West and East Flanders dialects |
| 11 | Dialect of East Flanders |
| 12 | Dialect of region between East Flanders and Brabant dialects |
| 13 | Dialect of the river region |
| 14 | Dialect of Noord-Brabant and northern Limburg |
| 15 | Dialect of Brabant |
| 16 | Dialect of region between Brabant and Limburg dialects |
| 17 | Dialect of Limburg |
| 18 | Dialect of the Veluwe region |
| 19 | Dialect of Gelderland and western Overijssel |
| 20 | Dialect of western Twente and eastern Graafschap |
| 21 | Dialect of Twente |
| 22 | Dialect of the Stellingwerf region |
| 23 | Dialect of southern Drenthe |
| 24 | Dialect of central Drenthe |
| 25 | Dialect of Kollumerland |
| 26 | Dialect of Groningen and northern Drenthe |
| 27 | Frisian language |
| 28 | Dialects of het Bildt, Frisian cities, Midsland, and Ameland Island |

Table 9.1: In the map of Daan and Blok (1969) 28 groups are distinguished. In the map in Figure 9.2 the locations of the groups are displayed.

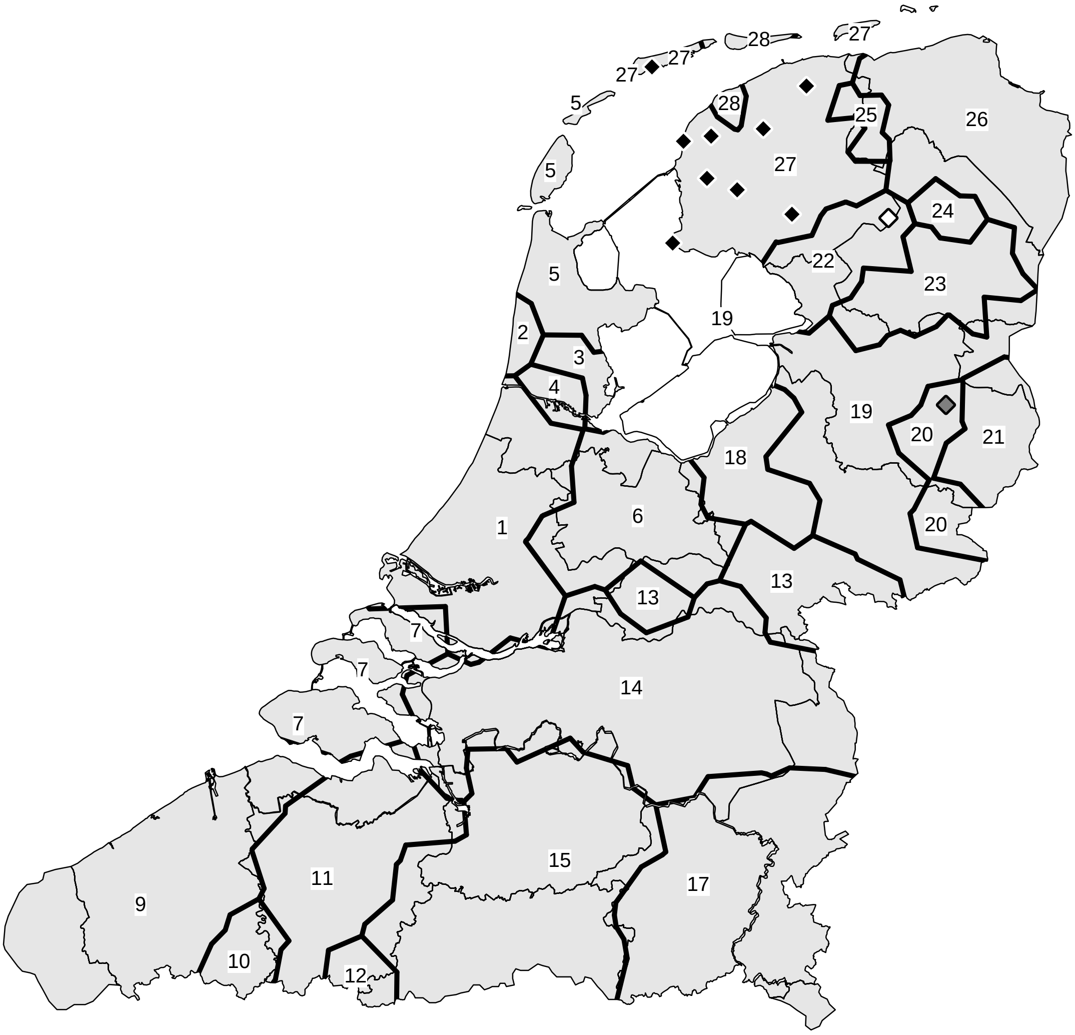


Figure 9.2: Locations of the 28 groups as distinguished in the map of Daan and Blok (1969). Provincial borders are represented by thinner lines and dialect borders by thicker ones. The numbers are explained in Table 9.1. Diamonds represent dialect islands. The black diamonds represent Frisian cities, which belong to group 28. The white diamond represents Appelscha, where both the dialect of group 22 and group 27 is spoken. The grey diamond represents Vriezenveen which contrasts strongly with its surroundings.

found the largest interval in volume 12 (Gelderland and Zuid-Overijssel) which was compiled during a period of 20 years. The 16 volumes of the RND were compiled by 16 different authors (and some assistants). Some authors worked on more than one volume, and some volumes were compiled by more than one author.

In Table 9.2 an overview is given of volumes and authors. The recording periods are borrowed from Reker (1997). The codes refer to the map in Figure 9.3. In this map, the areas per volume and per author are given. When we discuss results in Section 9.3, we can check whether the border we find represent borders between volumes or authors by consulting this map. For volume 2 (the southern part of East Flanders) the recordings are for the greater part made by E. Blancquaert, and for a lesser part by H. Vangassen in cooperation with Blancquaert. In the map, we mention both authors since the volume does not specify which transcriptions were made by which author. All locations in volume 5 (Zeeland Islands) were visited by both E. Blancquaert and P. J. Meertens. Therefore, we mention both authors again. In the map white diamonds represent dialects which were recorded by Blancquaert while surrounding dialects were recorded by one or more other authors. Grey diamonds in the area of volume 11 (Zuid-Holland and Utrecht) represent dialects which were recorded by Blancquaert and L. Oyen. Black diamonds in the area of volume 15 (Friesland) mark dialects recorded by Blancquaert, K. Boelens and G. van der Woude. By means of these joint recordings Blancquaert introduced the new fieldworkers to his system, ensuring a basic level of consistency.

Slightly different questionnaires are used in the RND. Generally speaking a Flemish version (volumes 1 through 8) and a Dutch version (volumes 9 trough 16) were used. For volume 6 many words were replaced by French equivalents. For volume 15 in most cases a Frisian questionnaire was used, and for a fewer cases a Dutch questionnaire. Our list of 125 words contains only words for which equivalents are found in all questionnaires. A questionnaire may direct dialect speakers to a certain degree, especially on the lexical level. Therefore the use of different transcriptions can make the RND material less consistent. On the other hand, questionaires were adapted so that they accord better with the dialect area for which they were used. This may result in more true transcriptions. A directing effect will be found only in transition zones. More about differences between questionaires is discussed in Heeringa (2001).

Although the goal was that transcribers should work using Blancquaert's guidelines to ensure consistency, the different transcribers use slightly different notations (see also Goossens (1977, pp. 71–72)). Examining the transcriptions of different varieties, it is not always clear whether differences are transcriber differences or real pronunciation differences, which makes it hard to trace all transcriber differences. However, we found a limited set of differences that were obviously transcriber differences. In this section they will be discussed briefly. For each of them, we describe how we normalized the data for them. In Heeringa

| Volume | Author(s) | Period | Code |
|---|---|---|---|
| 1 | E. Blancquaert | 1922–1925 | bla 1 |
| 2 | E. Blancquaert and H. Vangassen | 1927–1930 | bla/van 2 |
| 3 | E. Blancquaert | 1933–1935 | bla 3 |
| 4 | H. Vangassen | 1933–1935 | van 4 |
| 5 | E. Blancquaert and P. J. Meertens | 1932–1939 | bla/mee 5 |
| 6 | Willem Pée | 1934–1940 | pee 6 |
| 7 | Willem Pée | 1946–1953 | pee 7 |
| 8 | J. C. Claessens | 1937–1948 | cla 8 |
| 8 | W. Goffin | 1937–1948 | gof 8 |
| 8 | A. Stevens | 1937–1948 | ste 8 |
| 9 | A. Weijnen | 1939–1949 | wei 9 |
| 10 | A. R. Hol | 1949–1959 | hol 10 |
| 10 | J. Passage | 1949–1959 | pa 10 |
| 11 | L. Oyen | 1950–1962 | oye 11 |
| 12 | H. Entjes | 1950–1970 | ent 12 |
| 12 | A. R. Hol | 1950–1970 | hol 12 |
| 13 | Jo Daan | 1950–1962 | daa 13 |
| 14 | H. Entjes | 1974–1975 | ent 14 |
| 15 | K. Boelens | 1950–1951 | boe 15 |
| 15 | G. van der Woude | 1950–1951 | wou 15 |
| 16 | A. Sassen | 1956–1961 | sas 16 |

Table 9.2: List of volumes and authors of the RND together with the periods during which the recordings were made. The codes are used in the map in Figure 9.3.

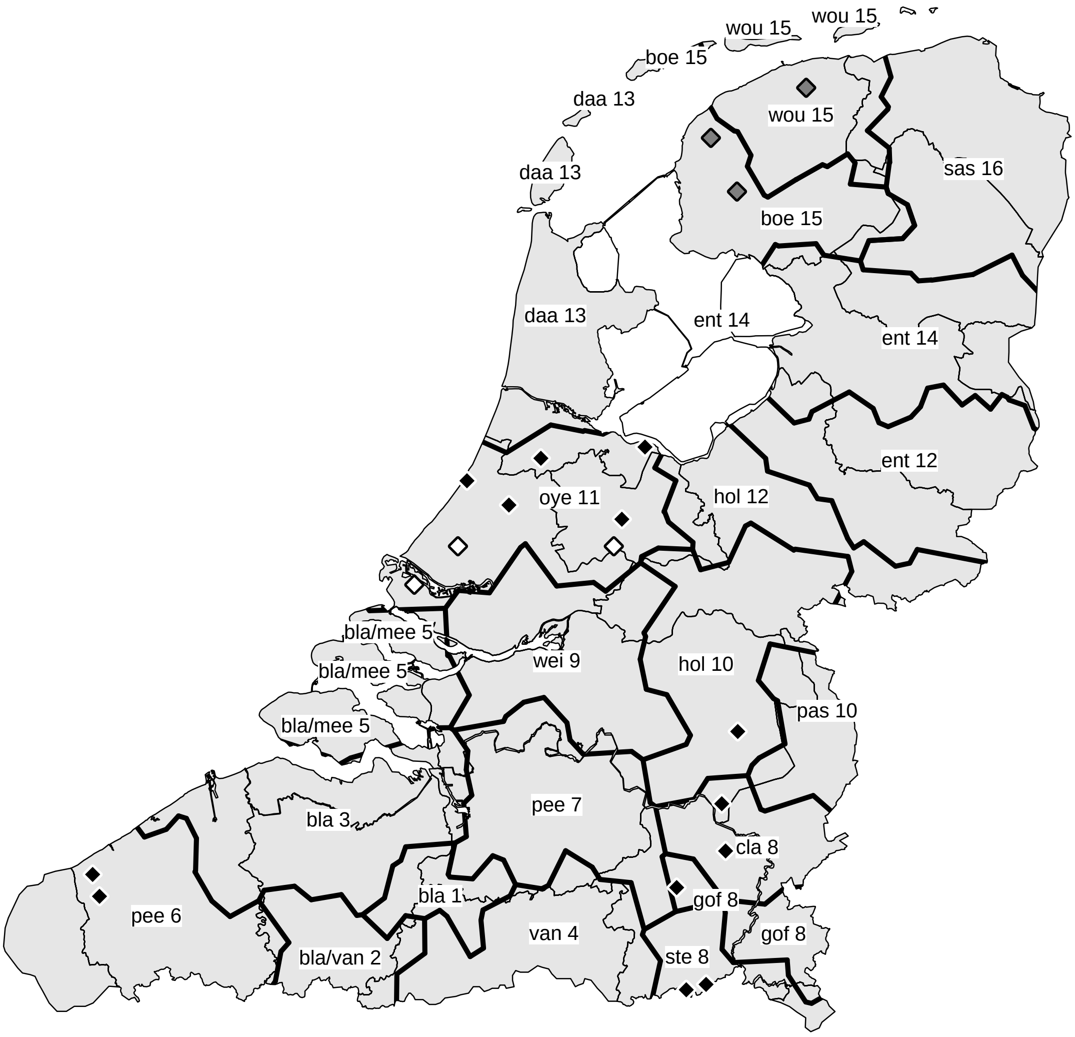


Figure 9.3: Distribution of volumes and authors over the 360 RND varieties. Provincial borders are represented by thinner lines and volume/author areas by thicker ones. Black diamonds represent recordings of E. Blancquaert, white diamonds represent recordings of Blancquaert and L. Oyen, and grey diamonds represent recordings of Blancquaert, K. Boelens and G. van der Woude. The codes are explained in Table 9.2.

(2001) some of the same consistency problems are discussed more extensively, but the way in which they are solved may differ slightly from the way we described in this section.

#### 9.1.4.1 Vowel + trill

In the RND, sometimes the *ee*, *oo* and *eu* before *r* are transcribed as respectively [eː], [oː] and [øː] (see e.g., Blancquaert (1948)) and sometimes as [ɪˑ], [ʊˑ] and [ʏˑ] (see e.g., the introduction of volume 13). Sometimes one author even used both notations intermixed (see e.g. the introduction of volume 16). To standardize different notations with the same meaning, we could replace each [e], [o] or [ø] before [r] or [ʀ] by respectively [ɪ], [ʊ] or [ʏ]. Since an [r] may also be weakened to a [ə], the [e], [o] and [ø] before [ə] should also be replaced by respectively [ɪ], [ʊ] or [ʏ]. However, it is not always clear whether an [ə] after an [e], [o] or [ø] is a weakened [r]. If not, the [e], [o] and [ø] should not be changed, to avoid that e.g., the relation between two different (dialect) pronunciations of *zee* 'sea', namely [zeː] and [zeːə̆] (the latter would be changed to [zɪːə̆]) is lost. However it is infeasible to determine the exact meaning of the large number of schwa's in the large number of varieties. The other possibility is to replace each [ɪ], [ʊ] or [ʏ] before [r] or [ʀ] by respectively [e], [o] or [ø]. However, when applying these substitutions, problems arise since the *r* is deleted in some pronunciations. E.g. the relation between [pɪˑrt], which is a dialect pronunciation of *paard* 'horse' (and which would be changed in [peˑrt]) and [pɪtjə], which is a dialect pronunciation of *paardje* 'small horse' will be is lost. To overcome all the problems mentioned, we replaced simply each [ɪ], [ʊ] and [ʏ] everywhere by respectively [e], [o] and [ø], not only when they appear before [r], [ʀ] or [ə], but in all other contexts as well. On the one hand, in the IPA quadrilateral the substitutes are very close to the substituted vowels. On the other hand, some contrasts are lost. However, we prefer the loss of these contrasts to retaining contrasts that only reflect notation differences and no real differences in pronunciation.

#### 9.1.4.2 Nasal + nasal

In volume 12 of the RND, we found that *bloemen* 'flowers' (sentence 2) was noted as [bloːmˑ] by Entjes (dialect of Laren), and as [blumn̩] by Hol (Spankeren). *Stenen* 'stones' (sentence 25) was noted as [steːnˑ] by Entjes (Groenlo), and as [steˑnn̩] by Hol (Spankeren). *Brengen* 'bring' (sentence 39) is noted as [brɛŋˑ] by Entjes (Laren), and as [brɛŋn̩] by Hol (Spankeren). The examples show that transcribers do not note the Dutch ending *en* as pronounced in Low Saxon dialects in the same way. We found similar variation between and even within transcriptions of Flemish dialects. Although it is conceivable that some of the different transcriptions represent genuine pronunciation differences (and our procedures are equipped to deal with this), we preferred again to err on the side of caution.

In the introduction of volume 12, Entjes mentioned that he transcribed the word *kunnen* ‘can’ as [kʏnˑ], while Hol noted the same pronunciation as [kʏnn̩]. Entjes writes that he only heard one longer [n], and not two [n]’s as suggested by Hol. To make the data as consistent as possible we have to replace either the two-nasal notations by one-nasal notations, or the one-nasal notations by two-nasal notations. We prefer to use the two-nasal notations which are also suggested by Twilhaar (1990). Considering the one-nasal notations, we found for e.g. the Dutch word *spannen* ‘to put’ (in the context of: put a horse to a cart) the following transcriptions: 1) [spɑn] (Nieuw Schoonebeek), 2) [spɑnˑ] (Oldemarkt), 3) [spɑnː] (Blankenberge), 4) [spɑn̩ː] (Alveringem), and 5) [spɑn̩] (Borger). Replacing the half-long, long and syllabic nasals as in the cases 2), 3), 4) and 5) can be done by an automatic procedure. Since the short nasal in case 1) represents probable de-gemination, we would like to let this short [n] also be replaced by a two-nasal notation especially to retain the relation with the half-long nasal. However, this cannot be done by an automatic procedure. Only nasals that correspond with the Dutch syllable *en* should be replaced. But not each short [n] corresponds with the Dutch syllable *en* as in case 1), so each short [n] in the data should be checked by hand. Therefore, we made a conversion in the other direction. We retained the notations where only one nasal is noted. This nasal may be noted as half long, long or syllabic. We replaced the two-nasal notations [mn], [nn] and [ŋn] by respectively [m̩], [n̩] and [ŋ̍] when they are found at the end of a word. If they are not found at the end of a word but rather are followed by a vowel, the substitutions are only made when the second nasal is noted as half-long, long or syllabic. If they are followed by another consonant, the substitutions are always made since in these contexts the second nasal can hardly be pronounced as non-syllabic. When replacing the two-nasal notation by a one-nasal notation, diacritics of the second nasal are left out. If the second nasal was respectively an [m], [n] or [n], the same procedure was followed.

#### 9.1.4.3 Plosive + nasal

In volume 12 of the RND, we found that *dopen* ‘baptize’ (sentence 35) was transcribed as [døːpm̩] by Entjes (Groenlo) and as [døːpn̩] by Hol. In volume 16 *Hebben* ‘have’ (sentence 106) was noted as [hɛbm̩] by A. Sassen (Bellingwolde). In volume 12 the same word was transcribed as [hɛbn̩] by Hol (Spankeren). In volume 12 Entjes noted *bakken* ‘bake’ (sentence 113) as [bɑkŋ̍] (Wilp) while Hol noted this word as [bɑkn̩] (Hoenderlo). In part 15 G. van der Woude noted *geslagen* ‘hit’ (sentence 131) as [slɑˑgŋ̍] (Kollum). The same author noted this word also as [slɑ̞ˑgn̩̆] (Dokkum).

In our opinion, an [n] after [p], [b], [k] and [g] is an unnatural pronunciation. On the other hand, an [m] after [p] or [b] and a [ŋ] after [k] or [g] may be pronounced easily. Therefore, we replaced [pn] by [pm̩], [bn] by [bm̩], [kn] by [kŋ̍] and [gn] by [gŋ̍] if the combinations were found at the end of a word. If they were not found at the end of a word and followed by a vowel, the substitutions are only made when the nasal is noted as half-long, long of syllabic. If the nasal was followed by another consonant, the substitutions are always made since in these contexts the second nasal can hardly be pronounced as non-syllabic. Existing diacritics of either the plosive or the nasal are not changed.

#### 9.1.4.4 Voiceless palatal plosive

In the feature table of Hoppenbrouwers and Hoppenbrouwers (2001, p. 40) the [c], [tj] and [t$^j$] get the same definition. The [c] is only found in volume 16. We follow Hoppenbrouwers and Hoppenbrouwers by changing all [tj]'s and [t$^j$]'s in [c]'s. For the [tj] the substitution is made regardless of whether the [t] or [j] or both are noted as extra-short.

#### 9.1.4.5 Voiceless velar fricative

In the phonetic overview in volume 16 A. Sassen explicitely mentioned the [g̭] (RND notation) as a voiceless fricative. As a sample word Sassen gives the Dutch word *wasgoed* 'wash'. We processed this simply as the IPA [x].

## 9.2 Distances

Using the Levenshtein distance, we find the distance between two pronunciations of the same word. The distance between two varieties is equal to the average of a series of Levenshtein distances computed from a series of word pairs. For 360 varieties, the average Levenshtein distance is calculated for each possible pair of varieties. The result is a 360 × 360 matrix. In Figure 9.4, the distances are geographically visualized. Strongly related varieties are connected by darker lines, while more distant varieties are connected by lighter ones. Where no lines are seen, the varieties are actually connected by white lines, indicating large distances.

In the picture, the strong relationships between the 'pure' Frisian varieties (Northwest) are clearly shown. When examining the picture, we should be aware of the fact that lighter lines, which represent the weaker relations between Frisian and town Frisian, are covered by the darker lines. Also the Groningen dialects (east of Frisian) form a group. Especially the most northern Groningen varieties are as close to each other as the Frisian varieties are. South of Groningen in Drenthe another small but close group is found. South of this group a large group is found in Overijssel. Especially the northern varieties are close. South of this group a sharp boundary is found, known as the boundary between Low Saxon (northeastern dialects) and Low Franconian (western, southwestern and southern dialects). In the rest of the map groups can also be found, although they are less distinct. However, when looking at the map from some distance, they

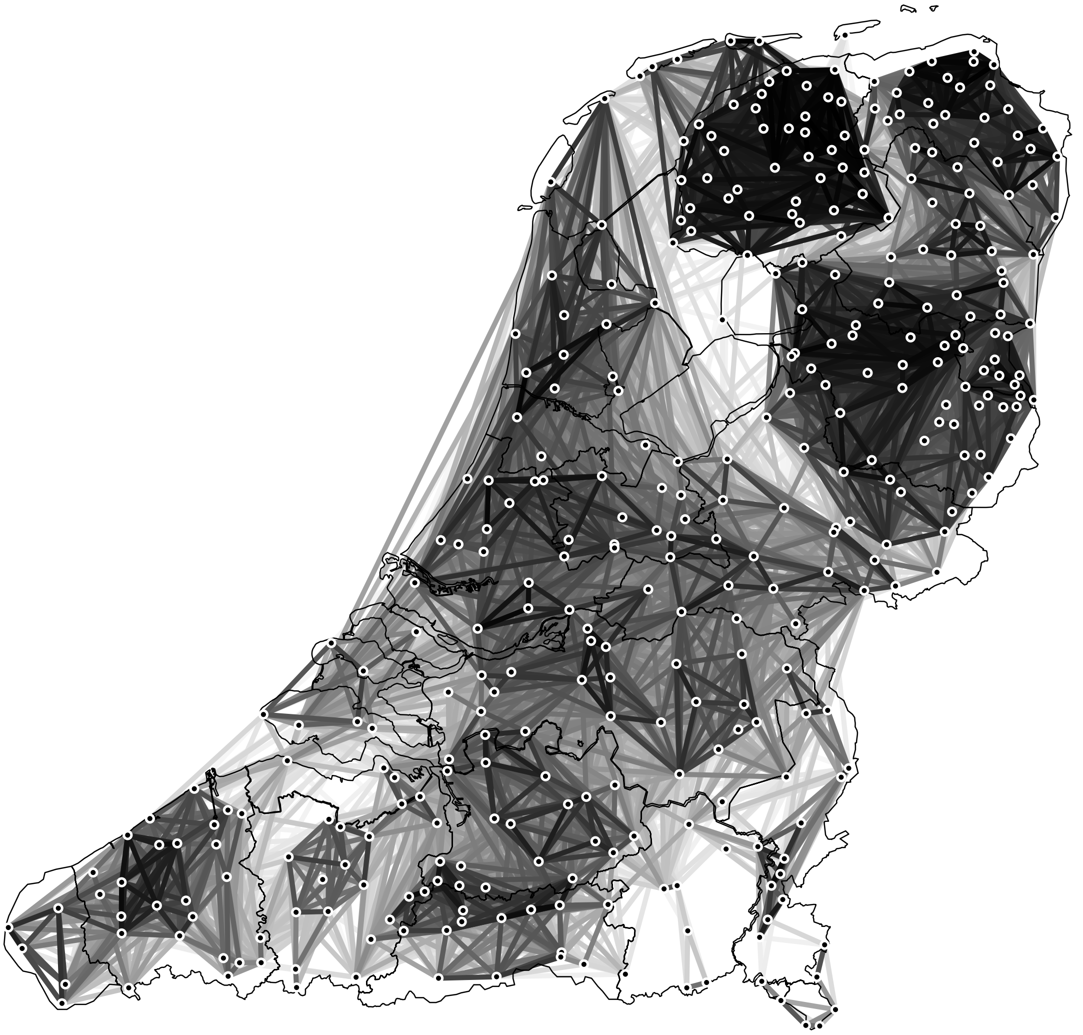


Figure 9.4: Average Levenshtein distances between 360 RND varieties. Darker lines connect close varieties, lighter lines more remote ones.

can be found. In the southwest, we find a French and West Flemish group. In the center of this area a group of strongly related dialects can be found. East of this area we find an East Flemish group, although the varieties are not so close. The same applies for the Zeeland varieties north of the East Flemish varieties. They are not close but emerge as a group since distances to other dialects are large, resulting in a lighter stroke around this set of varieties. East of the East Flemish group, we find an Antwerpen group (north) and a Brabant group (south). The two groups are connected in the East. East of these two groups, a lot of white area is seen, indicating the large distances that exist among the Limburg dialects. In the Dutch part of the Limburg area a small core area is found, where varieties are rather close. In the remaining part of the Dutch language area, it is hard to recognize groups on the basis of this map alone. Therefore, cluster analysis is described in Section 9.3. The result is an obvious division into groups. This division is compared to the division as shown in Daan's map.

Some borders suggested by the picture may not be real dialect borders. When looking at Figure 9.3, the border between Frisian and Groningen varieties, Frisian and Overijssel varieties and between Groningen and Overijssel varieties coincide with transcriber borders. This may be accidental, but we will keep track of it in the sections below.

## 9.3 Classification

### 9.3.1 Cluster analysis

On the basis of the distances between the 360 RND varieties we perform cluster analysis (see Section 6.1). The result is a large dendrogram in which all varieties are hierarchically ordered. In Figure 9.5 the dendrogram is displayed, showing only the 13 most significant groups. The scale distance shows percentages. The way in which percentages are found is described in the Sections 5.1.8 and 5.1.10. In the map in Figure 9.6, the 13 groups in Figure 9.5 are visualized by different colors. The colors are chosen by hand and inspired by the dialect map of Daan. When neighboring points belong to different groups, the exact border between the points is found on the basis of triangulation (see Section 6.1.5). To keep the picture simple, the dots and labels are given only for a restricted set of (in general) better-known locations. In the map, diamonds with and without labels can also be found. A diamond represents a language island, i.e. a variety which is only spoken in the location itself, and not in the area around the location.

We choose 13 groups since most of them correspond neatly with the groups which we found in the map in Figure 9.4. Some groups were not found in Figure 9.4. The Frisian mixed varieties were not found since they are geographically spread among the Frisian varieties. The Southwest Limburg group and the Northeast Luik group were not found since they consist of only a few dialects and are

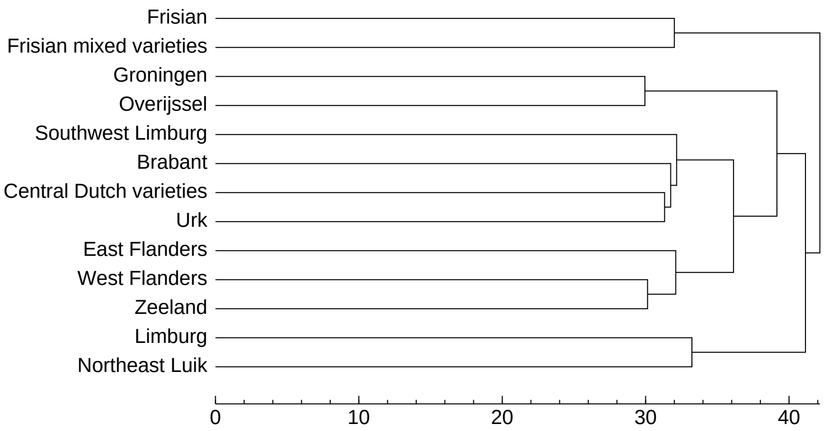


Figure 9.5: Dendrogram derived from the 360 × 360 matrix of Levenshtein distances showing the clustering of groups of Dutch dialects. UPGMA clustering is used (see Section 6.1.2). The scale distance shows percentages. Each of the 13 most significant groups is summed in one label and geographically visualized in Figure 9.6. The tree structure explains 70% of the variance.

considered a heterogenous area. Urk was not found as a 'group' since it consists of only one dialect. However, all of these groups are more significant than some other groups that were clearly found in Figure 9.4, e.g. the Zeeland group and the West Flanders group. When choosing more than 13 groups, we get groups that are not clearly recognized as groups in Figure 9.4 or groups with only a few, or even one dialect. Therefore, the number of groups in the main division was restricted to 13. In Section 9.4 each of the groups is discussed in more detail.

When considering the four main groups in the dendrogram in Figure 9.5, we get respectively Frisian (Frisian and Frisian mixed varieties), Low Saxon (Groningen and Overijssel), Low Franconian (Southwest Limburg ... Zeeland) and Limburg varieties (Limburg and Northeast Luik). A difference between our division and the division of Hoppenbrouwers and Hoppenbrouwers (2001, p. 58) is that we find Frisian to be most distant from the central Dutch varieties, while in the division of Hoppenbrouwers and Hoppenbrouwers the Frisian varieties are more closely related to the dialects in Noord-Holland, Zuid-Holland and Brabant. This difference confirms our leading hypothesis that regarding words as linguistic units and considering the structure of a word is important. These two aspects are not processed in the methodology of Hoppenbrouwers and Hoppenbrouwers. Similar to the main division of (Hoppenbrouwers and Hoppenbrouwers, 2001, p. 58) we find that the Limburg varieties do not belong to the Low Franconian varieties, but form a separate group.

### 9.3.2 Area map

When comparing the map in Figure 9.6 with Daan's map in Figure 9.2, we found similarities and differences. Our Frisian group corresponds perfectly to group 27 in the map of Daan. Our group of Frisian mixed varieties includes group 28 in Daan's map, but also the northern part of group 22. Daan's group 27 contains town Frisian varieties (most diamonds on our map), Ameland island (the island north of Leeuwarden) and the dialect of *het Bildt* (the area northwest of Leeuwarden). Group 22 contains the *Stellingwerf* varieties (on our map northeast of Steenwijk and southwest of Assen). It is striking that Daan's group 27 and a part of group 22 are one group on our map. Possibly the speakers of the *Stellingwerf* area in Daan's study did not consider the town Frisian language islands, but focused mainly on the sharp contrast between the *Stellingwerf* continuum and the Frisian continuum in their judgments.

In three locations in the Frisian continuum, and in one variety in the *Stellingwerf* continuum, two varieties are spoken. In Tjalleberd (the higher diamond south of Grouw), a Tjalleberd variety and a Frisian variety are spoken. The lighter blue color in the diamond represents the Tjalleberd dialect island, and the darker blue around the diamond the Frisian variety. In Donkerbroek (the lighter blue diamond east of Grouw and west of Assen) both a Frisian and a *Stellingwerf* variety is spoken. The lighter blue color in the diamond represents the *Stellingwerf* dialect island and the darker blue color around the diamond the Frisian variety. In Appelscha (the darker blue diamond southwest of Assen) the same two varieties as in Donkerbroek are spoken. The darker blue color in the diamond represents the Frisian language island, the lighter blue color around the diamond the *Stellingwerf* variety. The three locations are discussed in more detail in the sections below.

The Groningen group corresponds with the groups 25, 26 and 24 of Daan's map. However, the south border on our map is found more southerly, probably as the result of transcriber differences (see Figure 9.3). The Overijssel group corresponds with the groups 23, 19, 20 and 21 of the map of Daan. The south border is the border between the Low Saxon area (north) and the Low Franconian area (south). The border does not coincide with the border between group 19 (north) and groups 18 and 13 (south) in Daan's map. The difference may be explained by transcriber differences that influenced our results. In Daan's map both the Groningen group and the Overijssel group are divided into smaller groups. In our results, a closer division is not found when regarding only the 13 most significant groups. Daan and Blok (1969, p. 28) writes that when dialect differences are small, misgivings may be justified as to whether non-linguistic criteria had greater influence than linguistic-criteria. For example differences in social and economic structure may influence the awareness of the speakers. Furthermore, we suppose that borders in heterogeneous areas may represent larger differences than in homogeneous areas. Our method does not reckon with non-linguistic

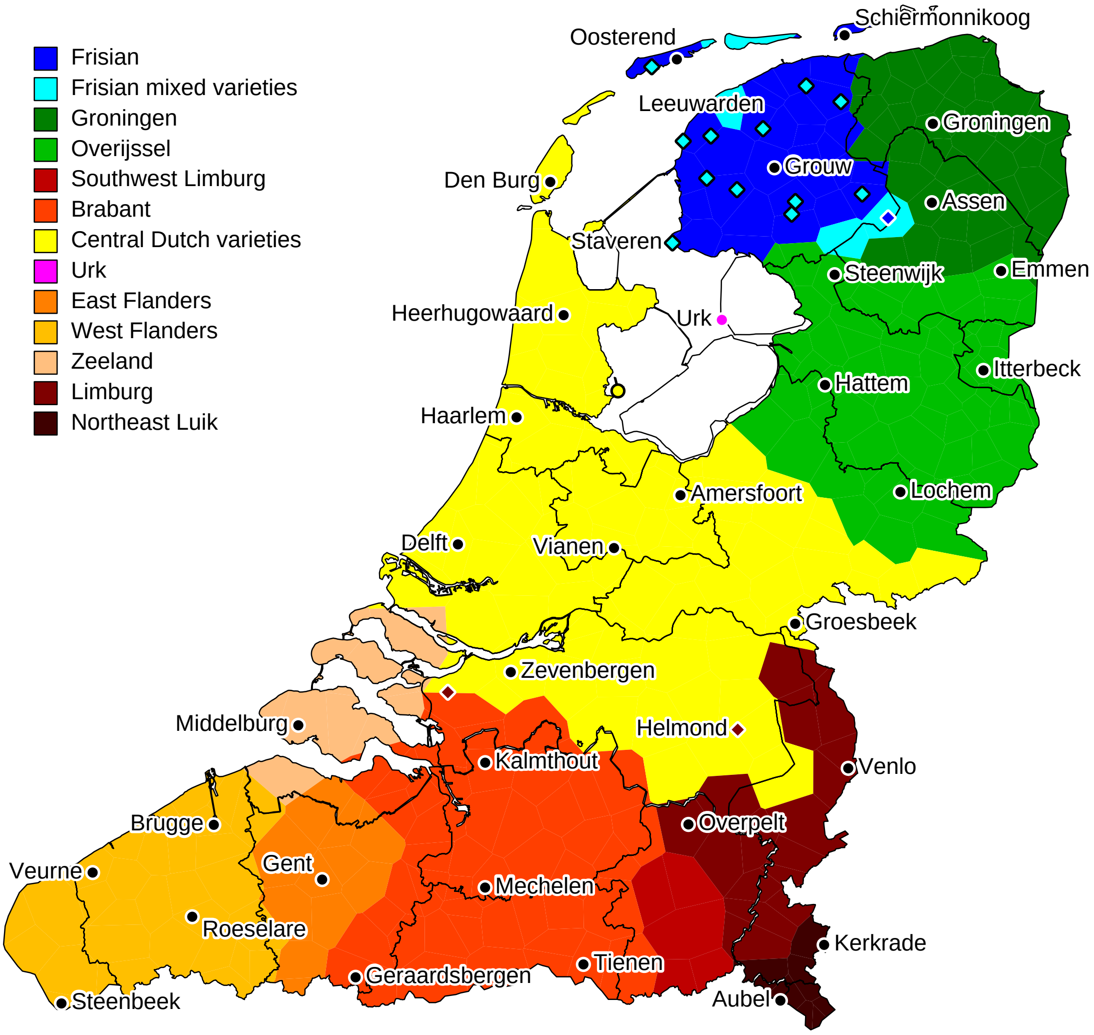


Figure 9.6: The 13 most significant groups as given in the dendrogram in Figure 9.5. UPGMA clustering is used (see Section 6.1.2). Diamonds represent language islands. Colors are chosen by hand and inspired by the dialect map of Daan.

factors on the one hand, and the degree of homogeneity on the other hand. This may explain why groups on our map are divided into several groups in Daan's map.

Our group of Central Dutch varieties corresponds with the groups 18, 6, 5, 2, 3, 4, 1, 13 and 14 in the map of Daan. The greater part of the south border corresponds rather well with the border in the map of Daan between the groups 1 and 14 (north) and the groups 7, 15 and 17 (south). Differences can be explained by transcriber differences. However, the northern part of the province of Limburg (the part north of Venlo and south of Groesbeek) is a part of group 14 in Daan's map, while it belongs to the Limburg group on our map. In the map of Te Winkel (1901) this area is separated from both group 14 and from the our Limburg group, while it is extended more to the north. We will discuss this in more detail in Section 9.4.12. Just as our Groningen and Overijssel group, the group of Central Dutch varieties is divided into a large number of groups in Daan's map. We mentioned some possible causes in the previous paragraph. However, for our Limburg group we found the opposite. Group 17 in the map of Daan divides into a southwest Limburg group, a Limburg group and a Northeast Luik group on our map. The fact that the borders did not exist in the awareness of the speakers may be explained by the fact that the Limburg area is a very heterogeneous area (Hoppenbrouwers and Hoppenbrouwers, 2001, p. 187). In an area with great dialect variation, it is more difficult to recognize groups (compare the Limburg area in the map in Figure 9.4). It is striking that the varieties of Steenbergen (the darker red diamond southwesterly of Zevenbergen) and Helmond were found among the Limburg dialects. Their relation with the Limburg varieties is discussed in Section 9.4.12.

Examining the remaining groups, we see that our Brabant group corresponds with groups 15 and 12 in Daan's map, our East Flanders group corresponds with group 11 of the map of Daan, our West Flanders group corresponds with groups 10 and 9 in Daan's map, and our Zeeland group corresponds with group 7 in the map of Daan. The border between the group of central Dutch varieties and the Brabant group is similar to the border between group 14 (north) and group 15 (south) in Daan's map, but is also influenced by transcriber differences. A striking difference between our map and the map of Daan is that the northeast part of group 11 forms a part of our Brabant group, and not a part of the East Flanders groups as may be expected on the basis of Daan's map. This is discussed in Section 9.4.6. Furthermore, the groups 12 and 11 form one group on our map as well as the groups 10 and 9. When examining these differences, we should be aware of the fact that the Belgian part of Daan's map is not based on the arrow method, but on information of language geographers who often belonged to dialect-speaking groups themselves. However, we did not know the exact criteria that were used by the language geographers, and the weightings of the criteria. Although the groups 12 and 10 can be found as separate groups on a deeper level (see Sections 9.4.9 and 9.4.10), their significance is not strong enough for them to be recognized as groups in the map of Figure 9.4 or found among the 13 most significant groups.

Among the 13 groups, the dialect of Urk is also found as a separate group. In Daan's map, this dialect belongs to group 19. Possibly the dialect of Urk is most like the dialect of group 19, so speakers judge the dialect of Urk as nearly the same as the dialect of group 19, although strong differences may exist. In our results, Urk is closest to the group of central Dutch varieties, but nonetheless appears as one of the 13 most significant groups. For more details see Section 9.4.8.

### 9.3.3 Composite cluster map

On the basis of the dendrogram in Figure 9.5 (including the subtrees of the main groups) we create a composite cluster map. The map is shown in Figure 9.7. In this map the borders between the most significant groups are darker blue, blue and green. These borders can also be found in Figure 9.4 more or less, and distinguish the areas as shown in Figure 9.6. Less significant borders are lighter green, greenish yellow, yellow and lighter yellow. The least significant borders are white. The benefit of this picture compared to Figure 9.6 is that it shows both the main groups and further classifications per group. To keep a clear picture, dialect islands which belong to the Frisian mixed varieties, and the dialect islands Steenbergen and Helmond are excluded. The dialect of Urk belongs to one of the 13 main groups. However, no borders are drawn around this variety since this former island is isolated by an area in which only Standard Dutch was spoken.

Examining the 13 main groups, we find that the Frisian group appears as an homogeneous area. Only the dialect of Hindeloopen (southwest) appears to be rather deviant from the other varieties of this group. Frisian mixed varieties are spoken on the Ameland island (north of the mainland), in *het Bildt* (northwest on the mainland) and in the Stellingwerf area (along the southeast province border). The Groningen group is divided in a northern and a southern part. The border partly coincides with the province border between Groningen (north) and Drenthe (south). The Overijssel group is divided in a western and an eastern part. The western part appears to be very homogeneous. The eastern part is less homogeneous. On the border between the western and the eastern part, the dialect of Vriezenveen appears as a dialect island. This variety is discussed further in Section 9.4.4. In the map it can be seen that the large group of Central Dutch varieties is divided in a western and a southeastern part. The Limburg area is divided a Limburg group, a southwest Limburg group and a southeast Luik group. The southeast Luik group is divided in a western and a eastern group. The division in these two groups reflects the Benratherlinie (see Section 9.4.13). In the Brabant group we find in the furthest Southeast a small but rather deviant group. In this group we find a west/east division. In the two western varieties (Diest in the north and Tienen in the south) the uvular [ʀ] is used, just as in the varieties of the Limburg group and the Northeast Luik group. This may

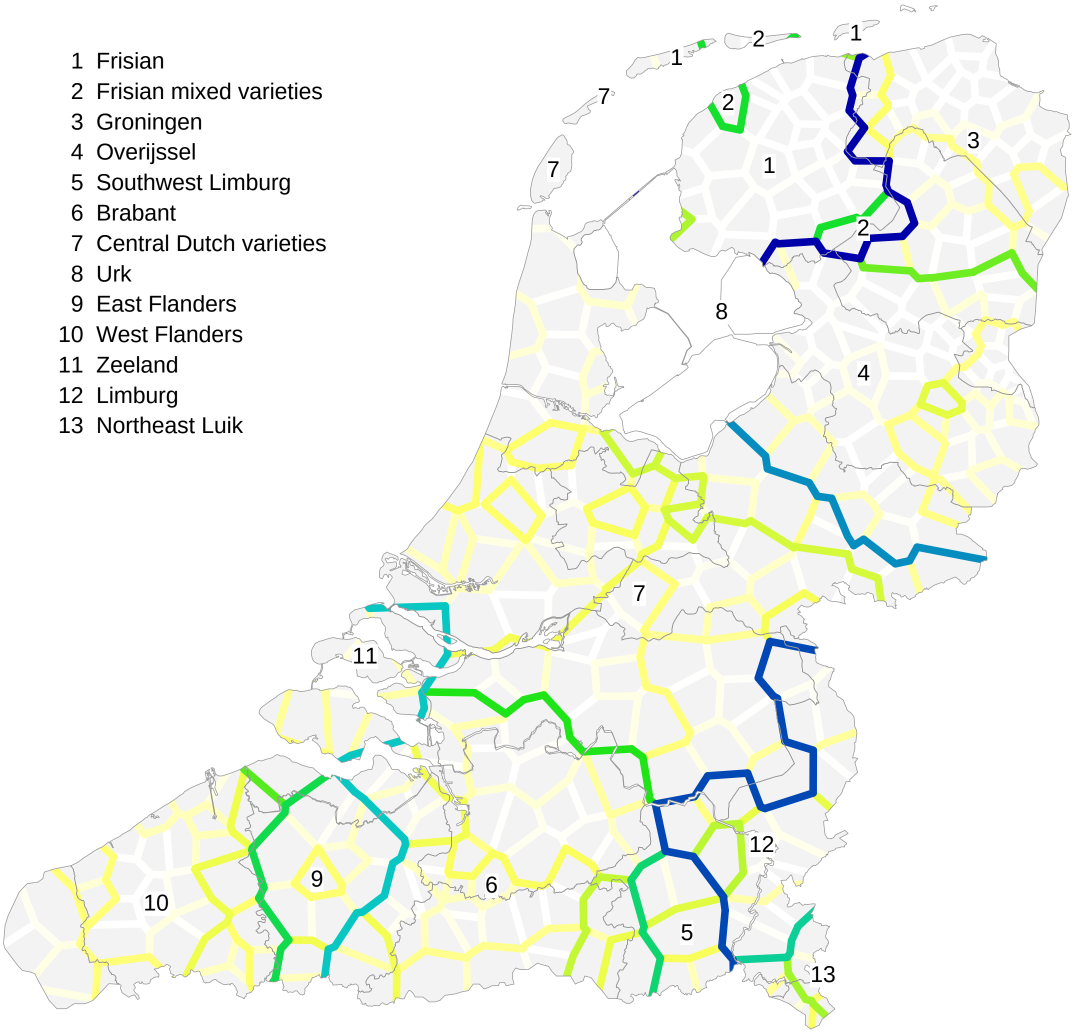


Figure 9.7: Composite cluster map on the basis of the dendrogram in Figure 9.5. UPGMA clustering is used (see Section 6.1.2). The most significant borders are represented by darker blue, blue and green lines, less signifant borders by lighter green, greenish yellow, yellow and lighter yellow lines. The least significant borders are white. Dialect islands (the diamonds in Figure 9.6) are excluded.

distinguish them from the other Brabant varieties. In Daan's map this group belongs to the Limburg varieties, and not to the Brabant varieties. However in the western variety (Velm) the alveolar [r] is used. Examining the Brabant group further we find a rather separate group in the Northwest. This group is discussed in Section 9.4.6. Furthermore we find a division between Anwerpen varieties (north) and Brabant varieties (south). In the center of the East Flanders group the dialect of Gent is found as a dialect island. Phonologically the dialect of Gent differs strongly from the surrounding varieties. E.g. all vowels in the variety of Gent are longer than in the varieties around this city. In the West Flanders group the central part appears to be rather homogeneous. The varieties along the eastern province border form a separate group which possibly may be seen as a transition zone between East Flanders and West Flanders. In the furthest Southwest we find the French Flanders varieties as an separate group. They are separated from the other varieties of the West Flanders group by a transition zone. Finally we find some borders in the Zeeland group. However these borders are not so sharp.

In Section 9.4 the classification of each of the 13 main groups is discussed in more detail.

## 9.4 Classification per subgroup

In this section each of the 13 groups as found in the dendrogram in Figure 9.5 and displayed in the map in Figure 9.6 is discussed in more detail. In Sections 9.4.1 through 9.4.13 a dendrogram will be given per group. From a dendrogram a closer division per group can be derived. The smaller clusters correspond with smaller areas within the larger area. These areas are displayed in Figure 9.8. The grey diamonds represent varieties that do not actually belong to the group in which they are geographically found. The white dots represent locations where two varieties are spoken. The one belongs to the local group, the other is a dialect island that does not belong to the group. When considering this map one should be aware that the significance of the groups is different. However, the main goal of this map is to find the varieties in a cluster of a dendrogram more quickly in the map. In Sections 9.4.1 through 9.4.13 different parts of the map are shown in more detail.

### 9.4.1 Frisian

In the map of Daan, group 28 represents the 'pure' Frisian varieties. This group is not divided further. However, a closer division is given in Hof's map (1933, p. 14a). This map is based on isoglosses. Furthermore, a closer division of

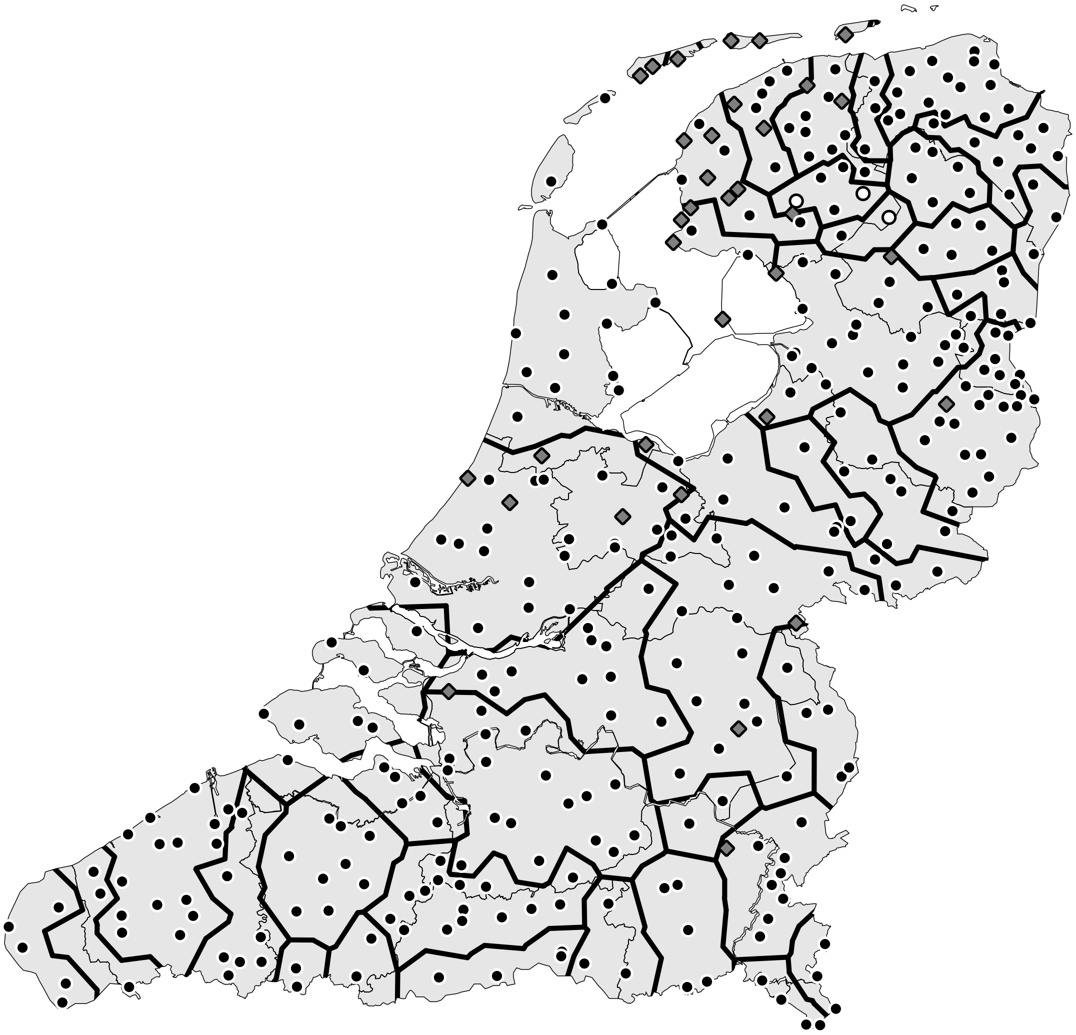


Figure 9.8: Closer division of the Dutch language area on the basis clusters as found within each of the 13 main groups. Provincial borders are represented by thinner lines and dialect cluster borders by thicker ones. The grey diamonds represent varieties that do not actually belong to the group in which they are geographically found. The white dots represent locations where two varieties are spoken. The one belongs to the group, the other is a dialect island that does not belong to the group.

mainland-rural Frisian is given by Van der Veen (1986, 1994). Van der Veen obtained his division on the basis of computational processing of isoglosses. These isoglosses are based on high-frequency words gathered from different sources (e.g. the RND).

Our division is given in Figure 9.9. The locations of the varieties can be found in the map in Figure 9.10. Most distinct within the Frisian group are Schiermonnikoog, Oosterend, West-Terschelling and Hindeloopen. The first three varieties are found on islands, and Hindeloopen is known as an isolated place inhabited by fishermen with an archaic dialect (Hoppenbrouwers and Hoppenbrouwers, 2001, p. 99). Apart from these four varieties, we find a division in a northern and a southern group. When examining the map of transcriber borders in Figure 9.3 this division perfectly reflects the division between the two transcribers who made the recordings in the Frisian area. However, within each of these groups more interesting results can be found.

In the 'northern' group the cluster Hallum ... Holwerd belongs to the Kleifries varieties, and the cluster Rottevalle ... Bakkeveen to the Woudfries varieties. Looking at the 'southern' group, we find the variety of the small city of IJlst as most distinct, and next, a cluster containing the varieties of Koudum and Lemmer. This cluster represents the Zuidhoeks (Frisian: Súdhoeksk) varieties. Going one level deeper, we find two groups. The cluster Workum ... Appelscha 2 may be considered as belonging to Woudfries (Frisian: Wâldfrysk), although Workum and Tjalleberd 1 are unexpected members of this group. For Workum we cannot explain this. For Tjalleberd this may have to do with the fact that in this place not only a Frisian, but a Low Saxon variety is spoken as well. The cluster, containing the varieties Oudeschoot ... Spannum represents a part of the Kleifries (Frisian: Klaaifrysk) area.

When combining the 'northern' Woudfries varieties with the 'southern' Woudfries varieties, and the 'northern' Kleifries varieties with the 'southern' Kleifries varieties, we obtain a division in three groups (apart from Hindeloopen, the island varieties and IJlst): Woudfries, Kleifries and Zuidhoeks varieties. These groups correspond rather well with the groups as given in Hof's map. However, the exact course of the border between Woudfries and Kleifries in our results is different from the course of the same border in Hof's map, and is nearly equal to the course of this border as suggested in the map of Van der Veen (1994, p. 7). We have the most confidence in the map of Van der Veen because of his well-considered choice and weighting of the isoglosses.

### 9.4.2 Frisian mixed varieties

In Daan's map group 27 includes the dialects of *het Bildt*, the Frisian cities, Midsland, and Ameland Island. In addition, group 22 is a transitional zone, consisting of the *Stellingwerf* varieties. In this section, the Frisian mixed varieties encompass group 27 and the northern part of group 22. We have not found a

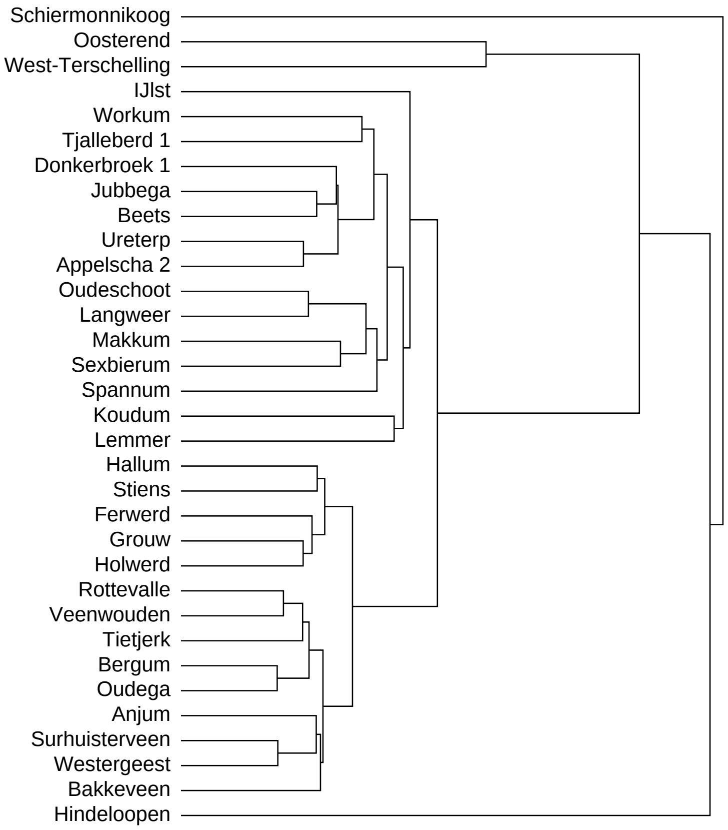


Figure 9.9: Subtree of the dendrogam in Figure 9.5, representing the Frisian group.

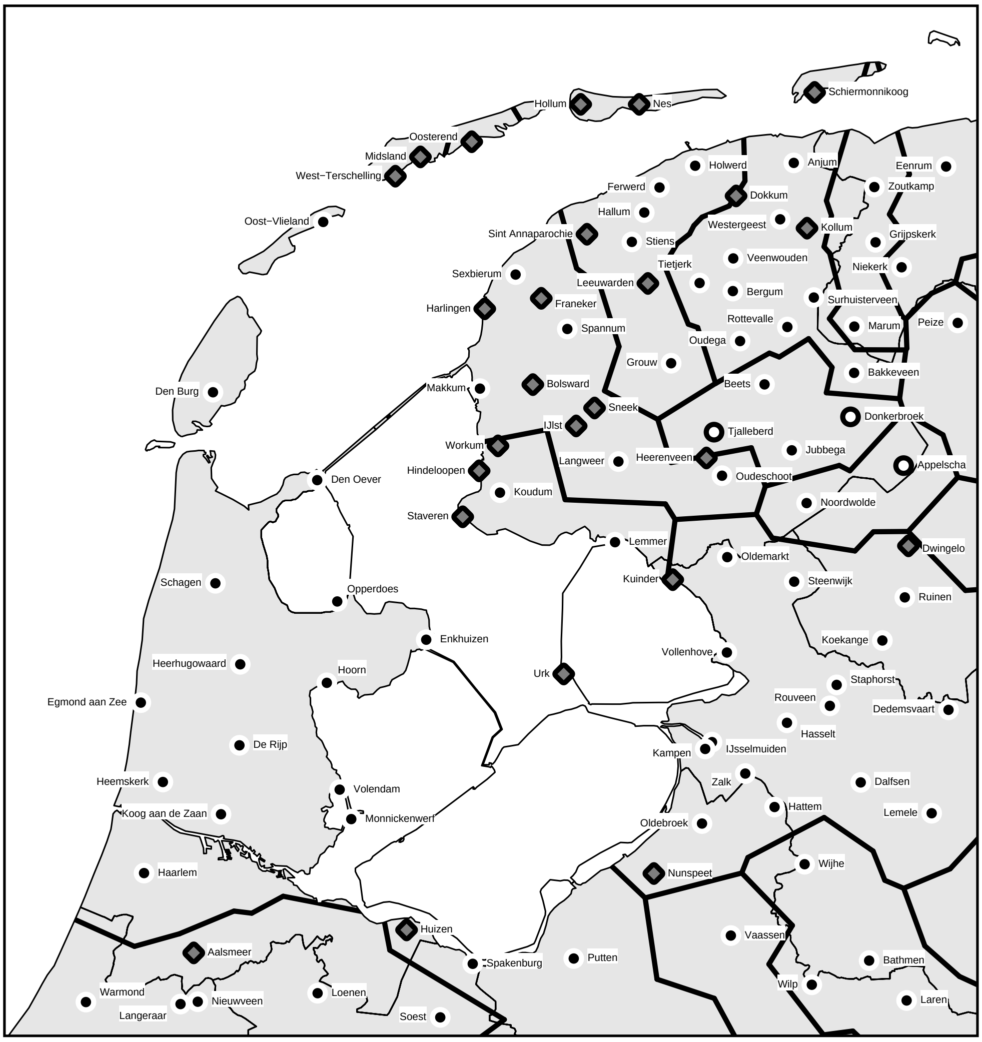


Figure 9.10: The northwestern part of the Dutch language area. Province borders are represented by thinner lines and dialect cluster borders by thicker ones. The grey diamonds represent varieties that do not actually belong to the group in which they are found geographically. The white dots represent locations where two varieties are spoken. The one belongs to the local group, the other is a dialect island which does not belong to this group.

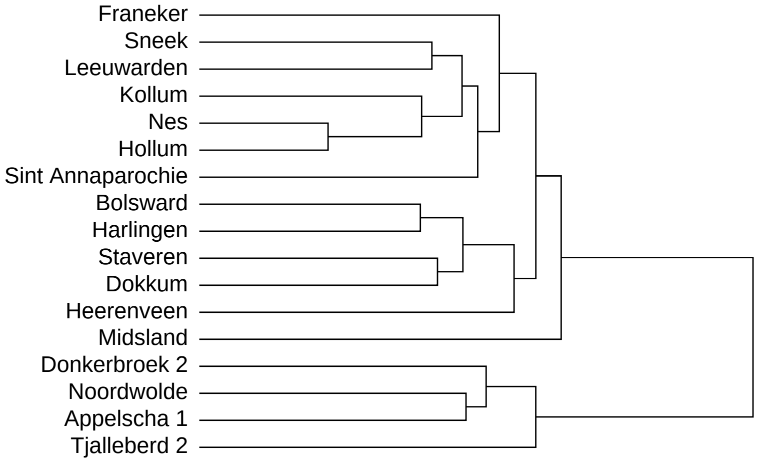


Figure 9.11: Subtree of the dendrogram in Figure 9.5, representing the Frisian mixed varieties group. The group contains town Frisian varieties and the *Stellingwerf* varieties. The dialect of *het Bildt* (represented by Sint Annaparochie) is found among the town Frisian varieties.

detailed discussion on the division of the groups 27 and 22 in the literature yet. In Daan's map, group 25 is also a transitional area, known as the *Westerkwartier.* This area is discussed in Section 9.4.3.

In Figure 9.11, a dendrogram is given that shows the division of the Frisian mixed varieties. The locations of the varieties can be found on the map in Figure 9.10. The dendrogram gives a clear division between *het Bildt* and town Frisian varieties on the one hand (cluster Franeker . . . Midsland), and the *Stellingwerf* varieties on the other hand (cluster Donkerbroek 2 . . . Tjalleberd 2).

Considering the cluster Franeker . . . Midsland, and ignoring Midsland we again find a northern and a southern group. Unfortunately, the border between these two groups represents the transcriber border (see Figure 9.3). Nonetheless, some conclusions can be drawn. First, the variety of Kollum clearly belongs to the town Frisian varieties, although it is found at the border of the Low Saxon area Kollumerlands (group 25) in Daan's map. The same finding was also found by Hoppenbrouwers and Hoppenbrouwers (2001, p. 96). Furthermore, the dialect of Sint Annaparochie (representing the dialects of *het Bildt*) does not appear as an outlier. Therefore, it is correct that in the map of Daan the dialect of *het Bildt* is considered as a town Frisian dialect. While the town Frisian varieties originated in making Frisian more Dutch, the dialect of *het Bildt* originated in making Dutch more Frisian. Our division suggested that these developments resulted in similar varieties.

Looking at the cluster Donkerbroek 2 . . . Tjalleberd 2, we found that the Low Saxon variety of Tjalleberd is clustered with the *Stellingwerf* varieties, although it appears as the most deviant variety within this group. On the map of Daan, the *Stellingwerf* area is colored green, suggesting that this area belongs to the Low Saxon varieties. However, in our results the *Stellingwerf* varieties of Donkerbroek, Noordwolde and Appelscha are not found among the Low Saxon varieties. They form a cluster (together with Tjalleberd) which is clustered with the cluster of the dialects of the Frisian cities (and other related dialects). Hoppenbrouwers and Hoppenbrouwers (2001) who obtained a similar classification, cited Sassen (1953, p. 305) who described the *Stellingwerf* dialect as a Drenthe dialect which became more like Frisian. On the other hand, we found some other *Stellingwerf* varieties clustered among the Low Saxon varieties (see Sections 9.4.3 and 9.4.4). The fact that the *Stellingwerf* varieties in our data set are found among both Frisian and Low Saxon varieties may partly be explained by transcriber differences, and meanwhile the question cannot be answered whether *Stellingwerf* varieties are more related to Frisian than to Low Saxon, or the other way round.

### 9.4.3 Groningen

We labeled the group in this section as Groningen varieties since the greater part is found in this province of Groningen. However, the northern part of Drenthe also belongs to this group. The group encloses the groups 25, 26, 24 and a small northern part of group 23 of Daan's map. In the map of Reker (1993, p. x) the province of Groningen is divided in West-Groningen, North-Groningen, *Oldambt*, *Westerwolde*, *Veenkoloniën*, and the city of Groningen. The map is based on isoglosses. A map of Heeroma (1963) suggests that the varieties in the northern part of Drenthe are more related to the Groningen varieties than to the varieties in the southern part of Drenthe. In Daan's map Groningen and northern Drenthe are clearly one group.

The division of the Groningen and northern Drenthe varieties is given in Figure 9.12. The location of the varieties can be found on the map in Figure 9.13. At the highest level in the dendrogram we find two groups. The cluster Marum ... Zoutkamp represents the *Westerkwartier* varieties, found as group 25 (Kollumerlands) on the map of Daan, and as the West-Groningen group in the map of Reker (1993). Going one level deeper, we get a division into Groningen dialects on the one hand (cluster Niekerk ... Groningen), and dialects mainly found in the northern part of Drenthe on the other hand (cluster Onstwedde ... Dwingelo). According to the map of Daan, the dialect of Zoutkamp should not belong to the *Westerkwartier* cluster, but to the cluster of Groningen varieties. The fact that this variety is classified with the *Westerkwartier* varieties in our results may be explained by the fact that Zoutkamp is a borderline case on the one hand, and transcribed by the same transcriber as the other 'real' *Westerkwartier* varieties on the other hand.

Examining the Groningen dialects, the cluster Niekerk ... Adorp represents a northern group. This cluster corresponds with the North-Groningen group in

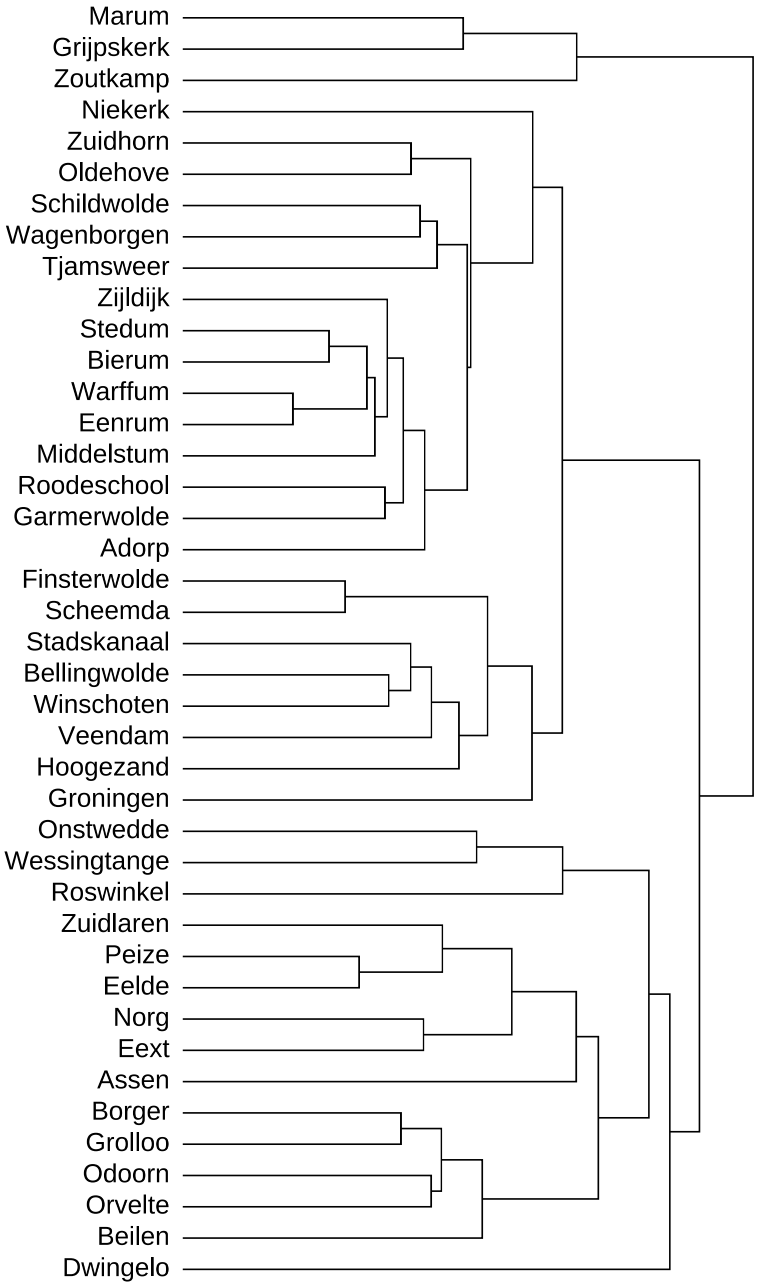


Figure 9.12: Subtree of the dendrogram in Figure 9.5, representing the Groningen group. The group contains varieties in Groningen and the northern part of Drenthe.

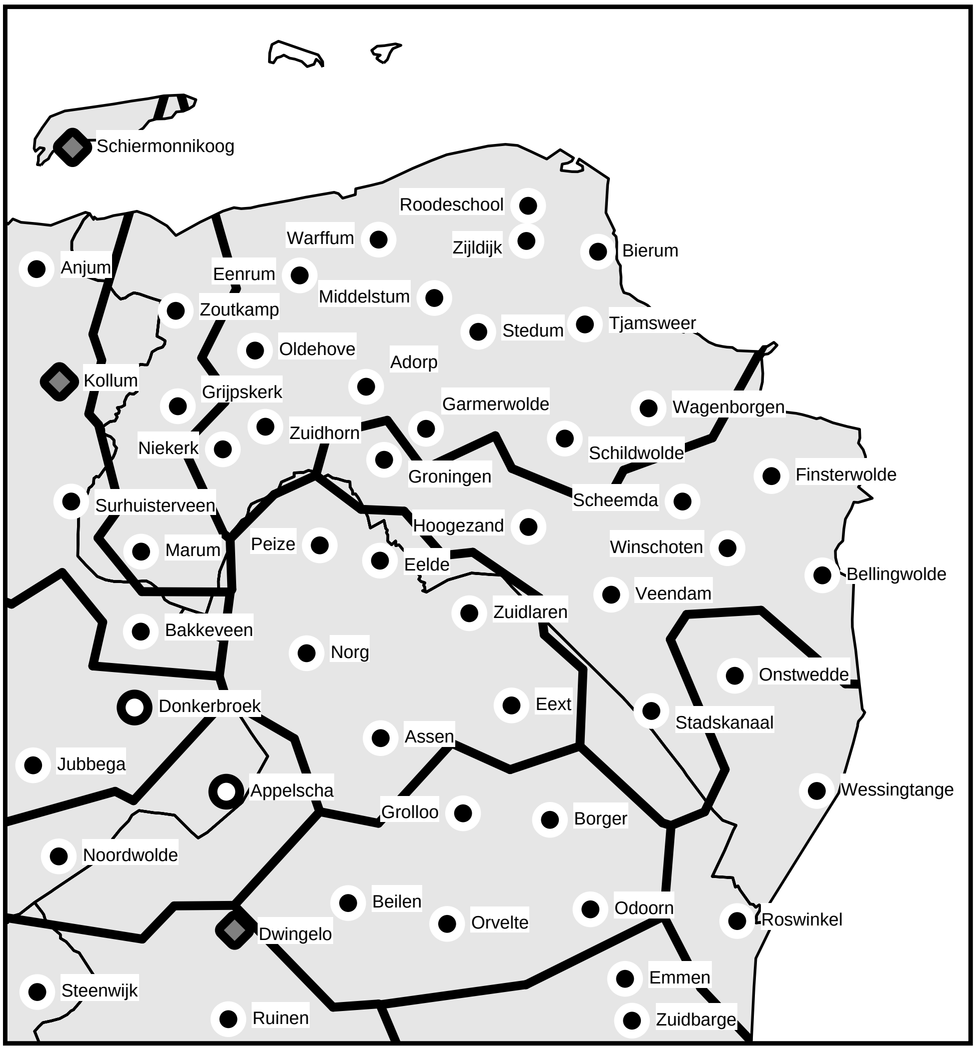


Figure 9.13: The northern half of the northeastern part of the Dutch language area. Provincial borders are represented by thinner lines and dialect cluster borders by thicker ones. The grey diamonds represent varieties that do not actually belong to the group in which they are found geographically.

the map of Reker. However, the cluster also encloses the northern part of the *Oldambt* area as given in the map of Reker. The cluster Finsterwolde ... Groningen represents an eastern group. The cluster corresponds with the *Oldambt*, the *Veenkoloniën* and the city of Groningen. A clear division between *Oldambt* and the *Veenkoloniën* is not reflected by this cluster. The dialect of the city of Groningen is most distinct in this cluster.

Looking at the other cluster that contains mostly varieties in Drenthe, we find that Dwingelo is not clustered with any of the other varieties in this group. In the map of Daan, this dialect belongs to the *Stellingwerf* varieties (group 22). Unfortunately, Dwingelo was not found in the *Stellingwerf* group in the dendrogram in Figure 9.11. The map in Figure 9.3 makes it clear that transcriber differences caused this separation. Going one level deeper we find that the cluster Onstwedde ... Roswinkel corresponds roughly with the *Westerwolde* area in the map of Reker. The other cluster contains varieties which are only found in Drenthe. The cluster Zuidlaren ... Eext is the group that is suggested to be strongly related to the Groningen varieties in the map of Heeroma and which simply belongs to the Groningen group on the map of Daan. The cluster Borger ... Beilen corresponds with the northern part of group of southern Drenthe varieties, found as group 23 on the map of Daan. However, the group halts exactly at a transcriber border.

Assen is clustered with the northern Drenthe group, although it appears as the most deviant variety in this cluster. In the map of Daan, it belongs to the central Drenthe varieties, mentioned as group 24. However, according to Daan's map, Grolloo also belongs to this group. In our dendrogram, Grolloo belongs to another cluster, namely the cluster of the southern Drenthe varieties. The transcription of Assen as given in the RND suggests that this variety is strongly influenced by Standard Dutch, which may explain why Assen is not too close to any of the other Drenthe varieties. Furthermore, the position of Grolloo does suggest that this variety fits perfectly in the southern Drenthe group rather than that it belongs to another group, namely the central Drenthe group. We found no explanation for this.

We may conclude that our results partly are in accordance with the map of Reker (1993). However, the border between the North-Groningen varieties and the *Oldambt* varieties is found further south in our results. In addition, we found no clear division between the *Oldambt* varieties and the *Veenkoloniën* varieties. To explain this difference, we should check whether the isoglosses used by Reker exist in the RND data. Furthermore, we clearly found a *Westerwolde* area, a Drenthe group and a southern Drenthe group. The varieties of the central Dutch group were either strongly influenced by Standard Dutch (Assen) or classified in the southern Drenthe group (Grolloo).

### 9.4.4 Overijssel

The Overijssel group encloses the southern part of Drenthe, Overijssel and the northern part of Gelderland. Since most varieties are found in Overijssel, we label the group simply the Overijssel group. The group corresponds with the southern part of the groups 22 and 23, the northern part of group 19, and with group 20 and 21 in Daan's map.

The dendrogram is given in Figure 9.14. The locations of the varieties are displayed in the map in Figure 9.15. In the dendrogram the deviant position of Vriezenveen immediately catches the eye. This agrees with Daan's map, in which the dialect is encircled, indicating that the dialect in this location is in strong contrast with its surrounding. Vriezenveen is an old settlement. The settlers came from the western coast area (Holland) (Entjes, 1970, pp. 2–15). The dialect has Westphalian influences.

Apart from Vriezenveen, we find two main clusters. The cluster Usselo ... Eibergen contains varieties in Bentheim and Twente while the cluster Nunspeet ... Bronkhorst contains the varieties around Bentheim and Twente.

The Bentheim/Twente cluster corresponds with group 21 in the map of Daan. In this cluster the cluster Rijssen/Eibergen and the dialect of Wierden are found to be rather apart from the other varieties in the cluster. The cluster Usselo ... Emlichheim represents varieties in or very close to German, where especially the cluster Langeveen ... Emlichheim contains varieties in and around Bentheim. Heeringa et al. (2000) report an investigation into the Dutch-German contact in and around Bentheim. The research was performed on the basis of the RND transcriptions and transcriptions of new recordings made in 1999. It appeared that the Dutch dialects shifted more towards Standard Dutch while all German dialects shifted towards Standard German. Finally the cluster Tubbergen ... Oldenzaal contains the core Twente varieties.

In the cluster with varieties around Bentheim and Twente, the dialect of Nunspeet, the cluster Vaassen/Bronkhorst and the dialect of Kuinder are found to be rather distinct. The deviant position of Nunspeet, Vaassen and Bronkhorst may be explained by the fact that they are found in a transition zone between the Low Saxon area and the Low Franconian area. According to Daan's map, the dialect of Kuinder belongs to group 22, i.e. the *Stellingwerf* varieties. Therefore, Kuinder would fit better in the *Stellingwerf* group in the dendrogram in Figure 9.11. The fact that this is not the case may be explained by transcriber differences. Apart from these four special cases, we find a cluster Emmen ... Nieuw Schoonebeek in the utmost southeast part of Drenthe, and a cluster Lochem ... Kampen containing varieties west of the Bentheim/Twente group. The southeastern Drenthe group partly covers group 23 and partly group 19 in Daan's map. In the group west of Bentheim/Twente the cluster Lochem ... Wijhe form a southern cluster partly corresponding with group 20 in the map of Daan, and the cluster Hattem ... Kampen form a northern cluster, covering the utmost

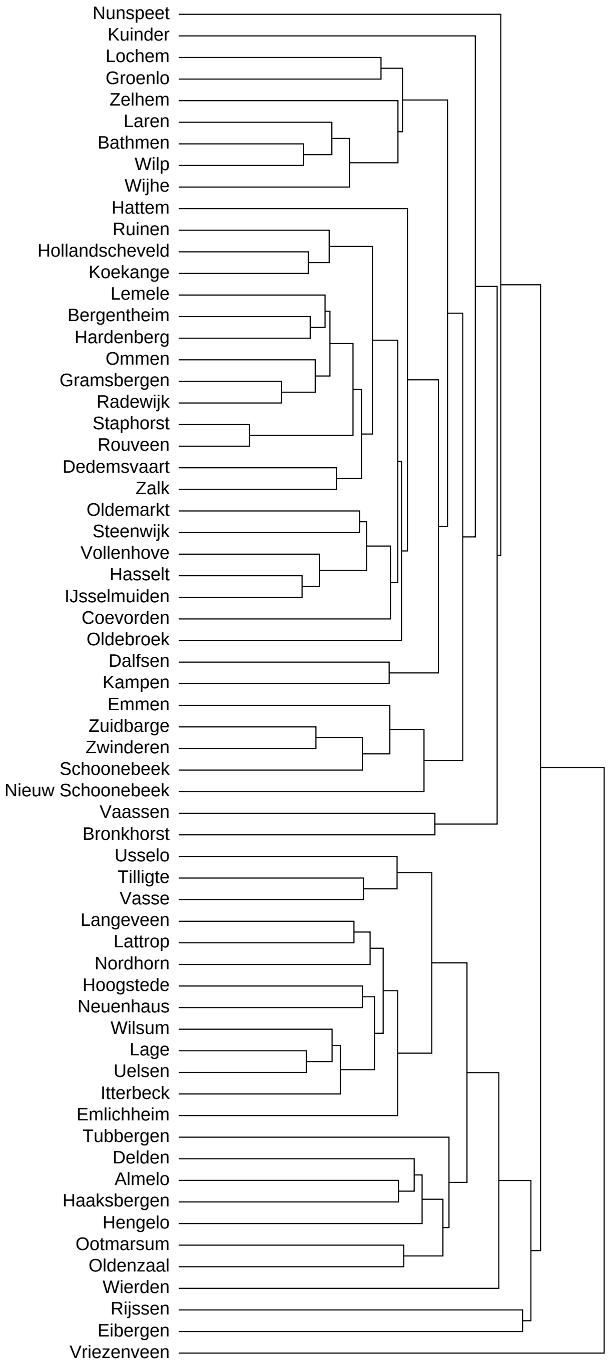


Figure 9.14: Subtree of the dendrogram in Figure 9.5, representing the Overijssel group. The group contains varieties in the southern part of Drenthe, Overijssel and the northern part of Gelderland.

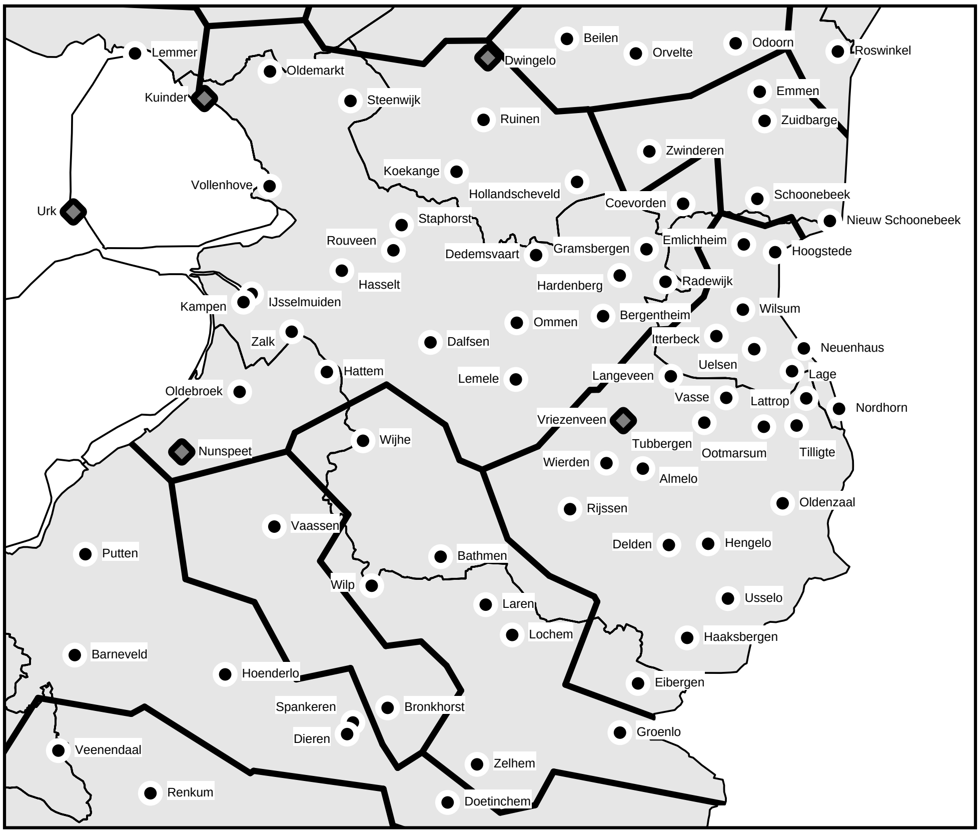


Figure 9.15: The southern half of the northeastern part of the Dutch language area. The state border and the provincial borders are represented by thinner lines and dialect cluster borders by thicker ones. The grey diamonds represent varieties that do not actually belong to the group in which they are found geographically.

southern part of group 23, and the larger northern part of group 19 on the map of Daan. In the latter cluster, it is striking that the cluster Dalfsen/Kampen is rather apart. Characteristic for these varieties is the use of the uvular [ʀ] which is uncommon for Low Saxon varieties. This uvular [ʀ] is also found in Zwolle, Deventer and Zutphen, but these varieties are not included in our data set. More about this phenomenon can be found in Hoppenbrouwers and Hoppenbrouwers (2001, p. 92).

In the large cluster Hattem ... Kampen, we find a small cluster containing the varieties of Ruinen, Hollandscheveld and Koekange. On the map of Daan, they belong to group 23, i.e. the southern Drenthe varieties. We expected them to be clustered with the southern Drenthe group in the dendrogram in Figure 9.12. The fact that they are absent there and found instead among the Overijssel varieties may be explained by transcriber differences. Furthermore, the varieties of Oldemarkt, Steenwijk and Vollenhove belong to group 22 in the map of Daan, i.e. the *Stellingwerf* varieties. Are they misclassified due to transcriber differences as well? We are not sure since they are clearly separated from Kuinder, which is also a *Stellingwerf* variety. Since Oldemarkt, Steenwijk and Vollenhove fit perfectly among the other Overijssel varieties while Kuinder is rather apart, we suspect that Kuinder is still a real *Stellingwerf* variety, while the other varieties are much more strongly related to the Overijssel varieties.

In the dendrogram discussed in this section, we found some differences with Daan's map, which cannot be explained by transcriber differences. The map of Daan is based on material from 1939 while the RND transcriptions are based on recordings made in 1974–1975 (south Drenthe and north Overijssel) and 1950–1970 (south Overijssel). Indeed, we found no explanation other than the fact that the situation has changed since 1939. In the dendrogram, we found a small southeast Drenthe group. This group suggested that the southeastern border of group 23 is shifted to the south. Furthermore, we expect that the southeast border of group 22 was shifted more to the west, since most varieties, which should belong to the *Stellingwerf* group in the map of Daan, do not appear as deviant to the other Drenthe and Overijssel varieties, with the exception of Kuinder.

### 9.4.5 Southwest Limburg

The southwest Limburg area is a small part of group 17 in Daan's map. Group 17 covers a large area, labeled as the dialect of Limburg. It is striking that our southwest Limburg area is not found as a separate group in the map of Daan. Nonetheless, it may be not surprising that a part of group 17 emerges as a separate group. It is known that the situation is complex in this area, and the varieties do not form a homogeneous group (Hoppenbrouwers and Hoppenbrouwers, 2001, p. 187). The dendrogram for our southwest Limburg area is given in Figure 9.16. The locations of the varieties can be found in the map in Figure 9.17.

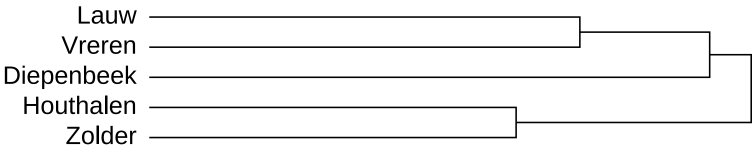


Figure 9.16: Subtree of the dendrogram in Figure 9.5, representing the Southwest Limburg group.

Although the borders of the southwest Limburg area partly coincide with transcriber borders, we think that this is not the only explanation why this set of varieties appears as one of the 13 most significant groups. The varieties distinguish themselves from the varieties in our larger Limburg group and smaller Northeast Luik group through the fact that the *r* is pronounced as an alveolar [r], and not as an uvular [ʀ]. The division as given by the dendrogram is not surprising: it follows the geography.

### 9.4.6 Brabant

The Brabant varieties roughly match group 15 in the map of Daan. However, a small number of varieties belongs to group 11, and the dialect of Geraardsbergen belongs to group 12. The division is shown in Figure 9.18 while the localities can be found in Figure 9.19.

The main division consists of the cluster Velm ... Tienen and the cluster Lot ... Geraardsbergen. All varieties in the first cluster are borderline cases. Velm just belongs to group 17 while Tienen and Diest marginally belong to group 15. On the original map of Daan, intermediate between group 15 and group 17 we find group 16: the dialect of the region between Brabant and Limburg. Since none of the varieties in our set of dialects belongs to this area, this area is not found in the map in Figure 9.3. However, if we assume that the borders in the map of Daan are drawn too narrowly or that the dialect area has been expanded since 1939, the dialects of Tienen, Diest and Velm represent a part of this dialect region.

In the second cluster we find a subcluster Lamswaarde ... Geraardsbergen and a subcluster Lot ... Arendonk. Most varieties in the subcluster Lamswaarde ... Geraardsbergen are found in the northeast of group 11. It is striking that these varieties are not clustered with the other varieties of group 11, which are found in the dendrogram in Figure 9.23 (see Section 9.4.9). Obviously, this cannot be explained by transcriber differences. Furthermore, in this cluster we find the varieties of Geraardsbergen and Heldergem. Geraardsbergen belongs to group 12 in the map of Daan, and Heldergem is a borderline case. The dialects are geographically isolated from the other varieties in this cluster and found further south, Geraardsbergen so much so that is close to the Flemish/Walloon border.

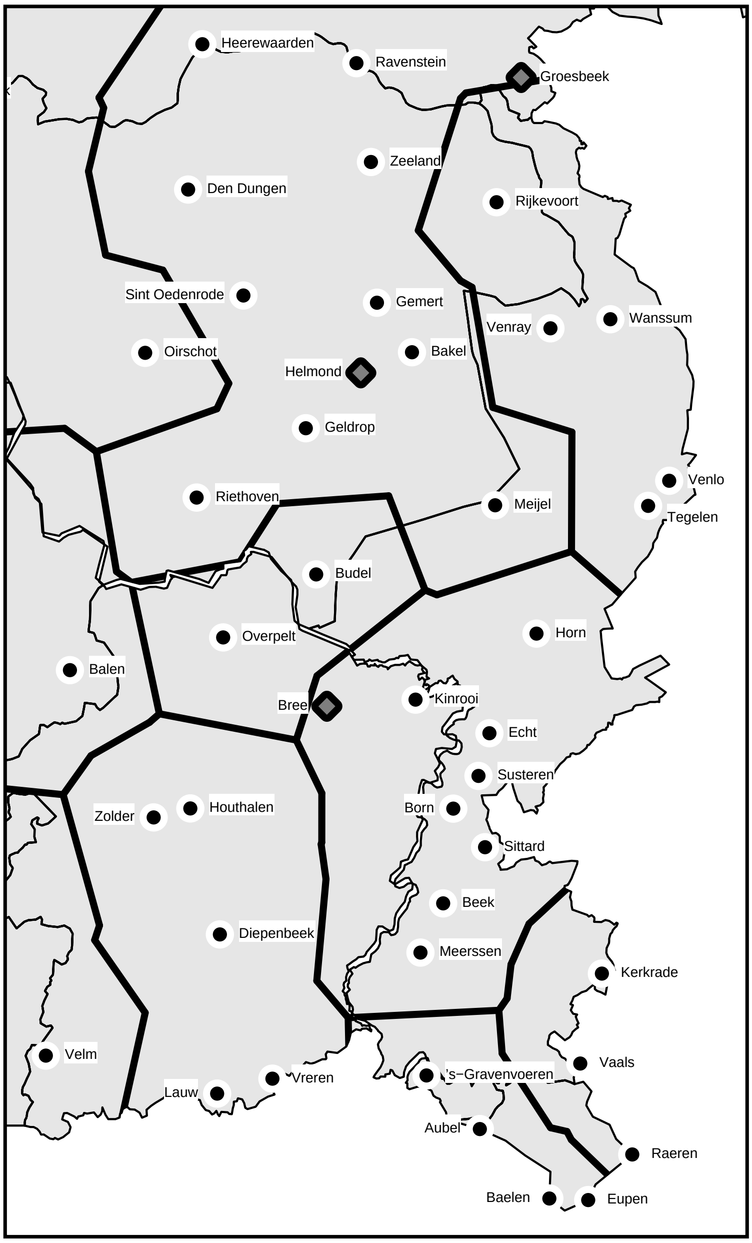


Figure 9.17: The southern half of the southeastern part of the Dutch language area. State borders and provincial borders are represented by thinner lines and dialect cluster borders by thicker ones. The grey diamonds represent varieties which, in our measurements, do not actually belong to the group in which they are found geographically.

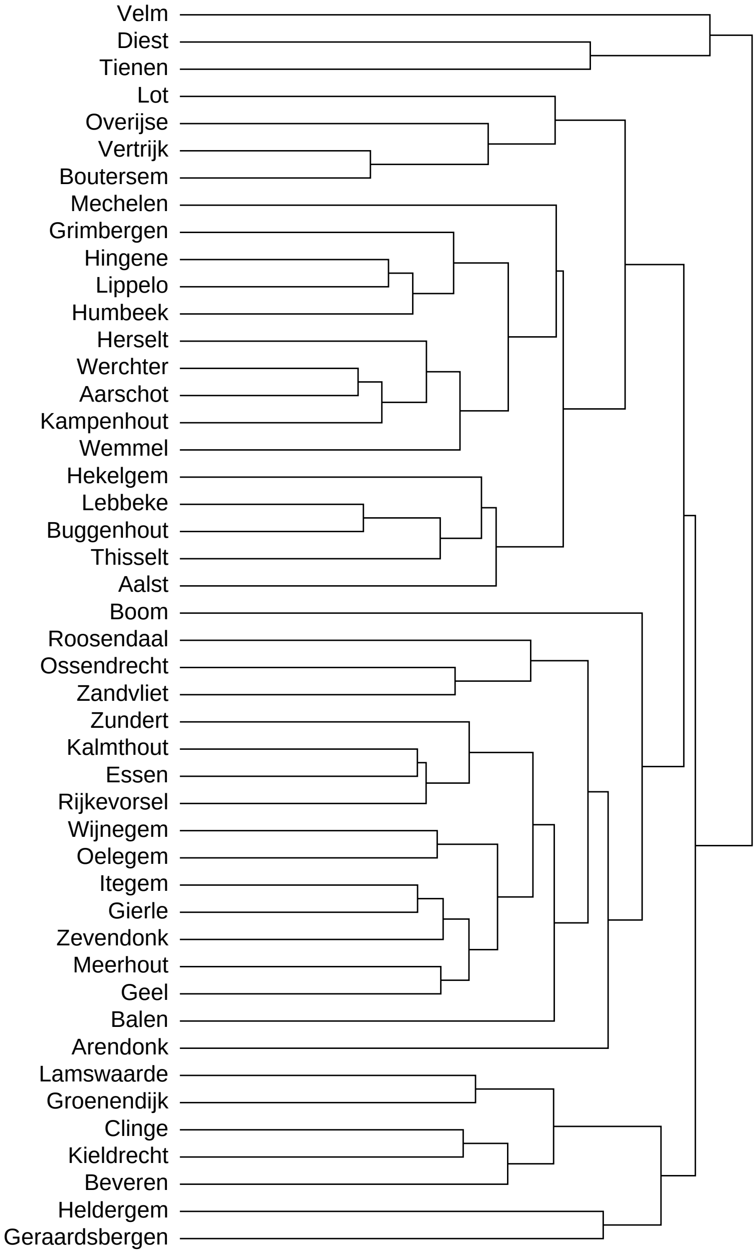


Figure 9.18: Subtree of the dendrogram in Figure 9.5, representing the Brabant group. The group contains varieties in Antwerpen and Brabant.

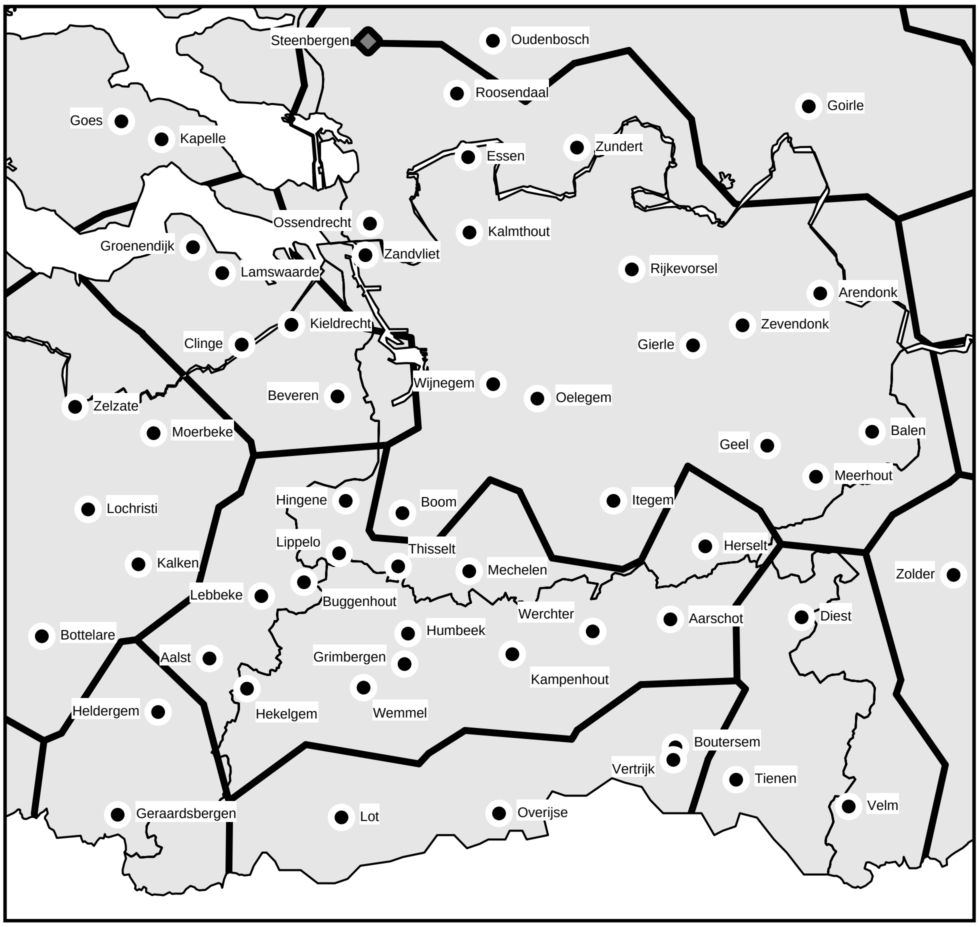


Figure 9.19: The mid-southern part of the Dutch language area. The state border and province borders are represented by thinner lines and dialect cluster borders by thicker ones. The grey diamond represents a variety that, in our measurements, does not actually belong to the group in which it is found geographically.

In the map of Daan, group 12 forms actually a small strip from the North to the South. Since this part of the map is not based on the arrow method, but on knowledge of language geographers, errors can be made because of the unexpected course of the borders. We therefore, suspect that group 12 should be expanded further to the north, covering the northeastern part of group 11.

In the subcluster Lot ... Arendonk we find a group containing southern varieties (Lot ... Aalst) and northern varieties (Boom ... Arendonk). In the group Lot ... Aalst, we in turn find a southern subgroup (Lot ... Boutersem) and a northern subgroup (Mechelen ... Aalst). These two subgroups roughly cover the province of Brabant. The group Boom ... Arendonk approximately matches the province of Antwerpen. Although the northern border roughly matches with the border as given in Daan's map, it matches perfectly with the transcriber border (see Figure 9.3).

### 9.4.7 Central Dutch varieties

The area of the central Dutch varieties is more or less found in the center of the Dutch language area, intermediate between the Frisian, Groningen and Overijssel varieties (northern) and the Zeeland, Flemish and Limburg varieties (southern). In our results this large central area is divided into a western and an eastern part only when examining the 28 most significant groups. The group encloses the southern part of the groups 18 and 19 (the central part of Gelderland) in the map of Daan, group 13 (the southern part of Gelderland) and group 14 (Noord-Brabant), group 1 (Zuid-Holland), group 6 (Utrecht) and the groups 5, 2, 3 and 4 (Noord-Holland). A division of the central Dutch varieties is given in Figure 9.20. The locations of the varieties are found in the maps in the Figures 9.21 (the eastern parts of Utrecht and Noord-Brabant and the southern part of Gelderland), 9.22 (Zuid-Holland and the western parts of Utrecht and Noord-Brabant) and 9.10 (Noord-Holland). When discussing the groups below, we will refer to the relevant map for the main groups.

Examining the dendrogram, the position of Huizen strikes the eye. This variety is most deviant from the other dialects. In Daan's map the special position of Huizen cannot be found. However, in Te Winkel's map (1901) the dialect of Huizen and its surroundings are suggested as a separate group, labeled as 'Gooisch'. Te Winkel's map is based on data obtained from questionaires. Data was obtained for 383 regions and places. An exact account was not given for the division (Daan and Blok, 1969, p. 18).

When going one level deeper, the cluster Aalten ... Woudenberg represents a central Gelderland group (see Figure 9.21) while the cluster Groesbeek ... Utrecht contains other central Dutch varieties. In the main division in Hoppenbrouwers and Hoppenbrouwers (2001, p. 58), the central Gelderland group is clustered close to the Frisian and Frisian mixed varieties. In our results, this group is much more distant from Frisian varieties, and closer to the central Dutch varieties.

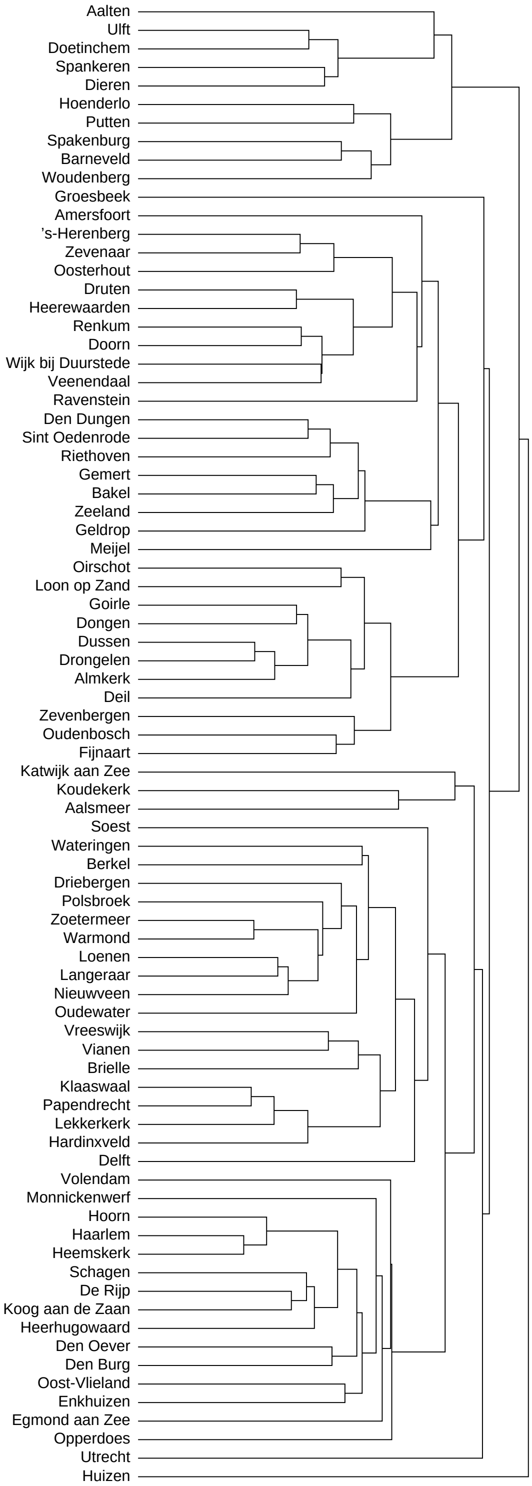


Figure 9.20: Subtree of the dendrogram in Figure 9.5, representing the Central Dutch varieties group. The group encloses varieties in south Gelderland, Noord-Brabant, Zuid-Holland, Utrecht and Noord-Holland.

The group covers the southern parts of the groups 18 and 19 of Daan's map. Unfortunately, the northern border of this group matches exactly the transcriber border (see Figure 9.3). The southern border, however, matches mainly with the border found intermediate between the groups 18 and 19 in the North and the groups 6 and 13 in the South in Daan's map (see Figure 9.2). However, our border runs west of Woudenberg, while in the map of Daan the border runs east of this location. So for this variety, the influence of transcriber differences is noticeable.

When going one level deeper again, we find a cluster Groesbeek ... Fijnaart which corresponds with southern Gelderland en Noord-Brabant, and a cluster Katwijk aan Zee ... Utrecht corresponding with Noord-Holland, Zuid-Holland and the greater part of Utrecht. In the cluster with the south Gelderland and Noord-Brabant varieties, Groesbeek is rather distinct from the other varieties. Geographically, Groesbeek is close to our Limburg group (see Figure 9.17 and Section 9.4.12). The dialect is probably an intermediate variant between the Gelderland varieties and the Limburg varieties. Apart from Groesbeek, we find a cluster Amersfoort ... Meijel, containing varieties in south Gelderland and the eastern part of Noord-Brabant (see Figure 9.21), and a cluster Oirschot ... Fijnaart containing mainly varieties in the western part of Noord-Brabant (see Figure 9.22). The south Gelderland/east Noord-Brabant cluster covers the eastern parts of the groups 6, 13 and 14 in Daan's map. The western part of the southern border matches with the the western part of the border between group 14 (north) and group 17 (south) in the map of Daan. The east Noord-Brabant cluster covers the eastern parts of the groups 13 and 14. The east/west division of the group 6, 13 and 14 in our results is clearly the result of transcriber differences. On the other hand, the east/west division of group 14 is also suggested in the map of Te Winkel (1901).

In the cluster with the varieties in Noord-Holland, Zuid-Holland and Utrecht, we found the dialect of Utrecht to be rather distinct (see Figure 9.22). The dialect of Utrecht is probably a typical town variety, contrasting with the surrounding rural dialects. Going one level deeper, we find the variaties of Katwijk aan Zee, Koudekerk and Aalsmeer as a separate cluster (see Figure 9.22). On the one hand, this is probably the result of transcriber differences (see Figure 9.3). On the other hand, Katwijk aan Zee is known to be an isolated place inhabited by fishermen, where an archaic dialect is spoken (see Hoppenbrouwers and Hoppenbrouwers, 2001, p. 152). In Hoppenbrouwers and Hoppenbrouwers (2001, p. 150) the dialect is even clustered with the Zeeland varieties, while our dendrogram suggests a stronger relation to the Holland and Utrecht varieties.

Going one level deeper again, the cluster Soest ... Delft contains varieties in Zuid-Holland and Utrecht (see Figure 9.22), while the cluster Volendam ... Opperdoes contains mainly varieties in Noord-Holland (see Figure 9.10). The cluster that contains varieties in Zuid-Holland and Utrecht, corresponds with the groups 1 (Zuid-Holland) and 6 (Utrecht) of the map of Daan. However, we found no border between these two groups. Rather we found a north/south

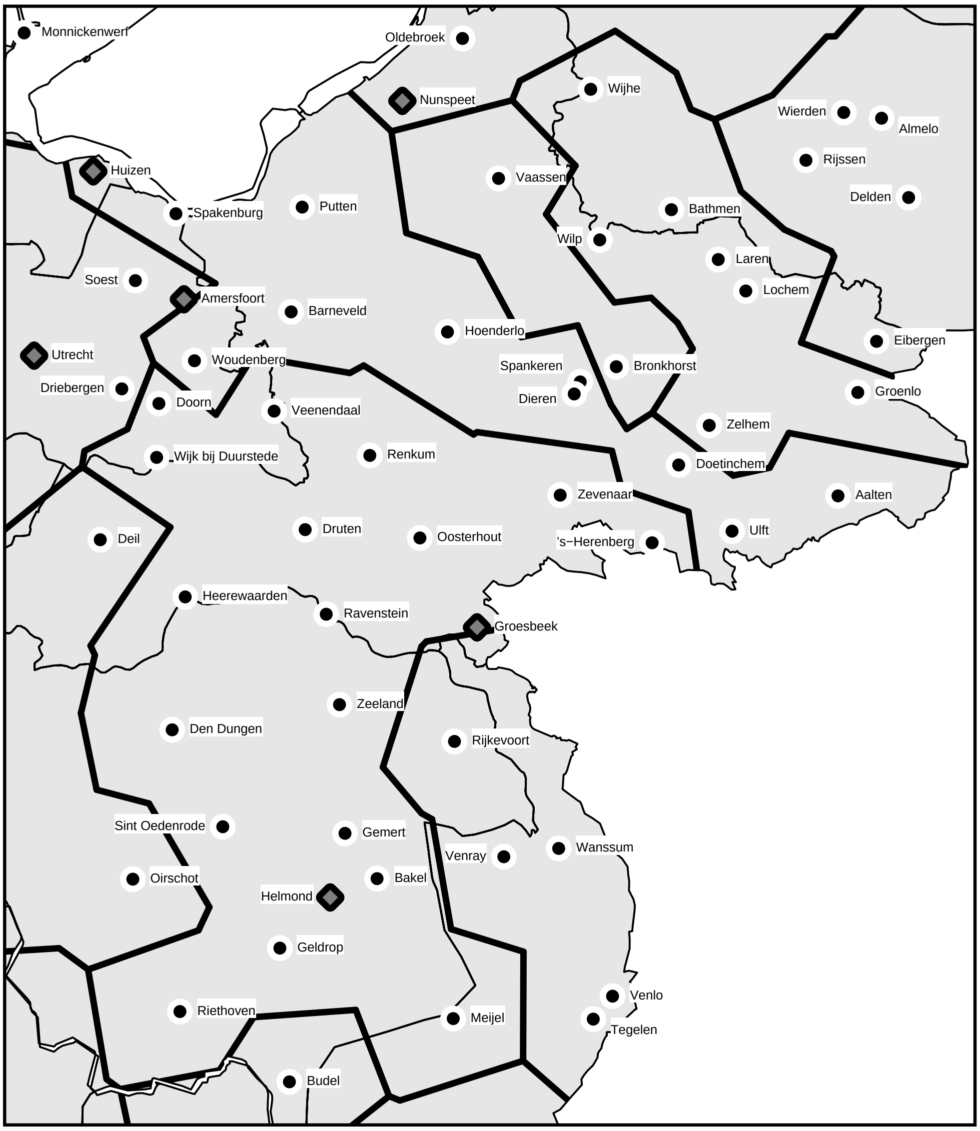


Figure 9.21: The northern half of the southeastern part of the Dutch language area. The state border and provincial borders are represented by thinner lines and dialect cluster borders by thicker ones. The grey diamonds represent varieties that do not actually belong to the group in which they are found geographically.

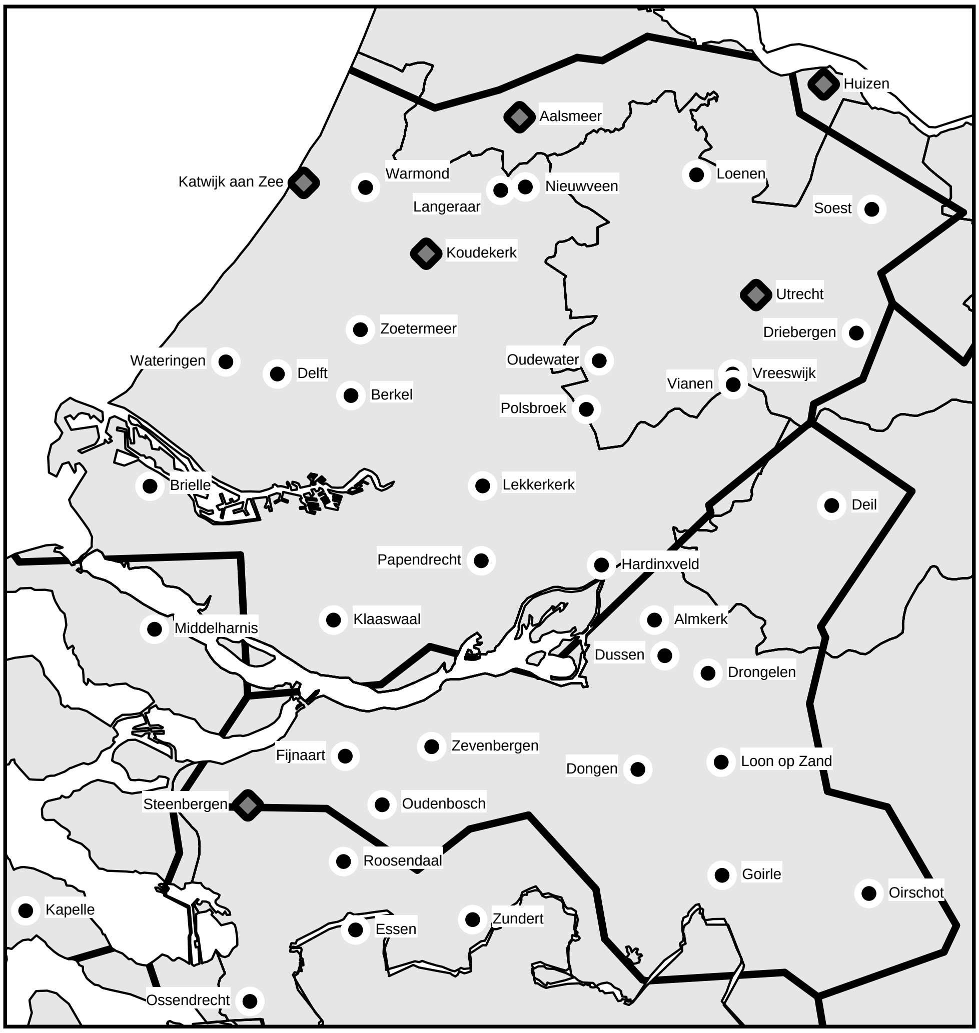


Figure 9.22: The mid-western part of the Dutch language area. The state border and province borders are represented by thinner lines and dialect cluster borders by thicker ones. The grey diamonds represent varieties that, in our measurements, do not actually belong to the group in which they are found geographically.

division across the both provinces. Apart from Delft and Soest, we find a cluster Wateringen . . . Oudewater representing the northern part, and a cluster Vreeswijk . . . Hardinxsveld representing the southern part. Daan and Blok (1969, p. 31) write that it was problematic to find the right border between Zuid-Holland and Utrecht, although they intuitively thought that there must be a border. Finally, they found that the surname *Bartels* was pronounced as [bɑrtəls] in Zuid-Holland, and as [baːrtəls] in Utrecht. However, the ɑ/a-isogloss could not be found on the basis of the data obtained from the questionaires, but was finally found on the basis of tape recordings. We doubt whether this Zuid-Holland/Utrecht border still reflects the language awareness of the dialect speakers, all the more since we did not find this border either.

The cluster containing the varieties in Noord-Holland corresponds with the groups 2, 3, 4 and 5 and a small northern part of group 1 in Daan's map. These different groups in the map of Daan cannot be clearly identified in our dendrogram. At a deeper level, we find two close clusters: Hoorn . . . Heemskerk and Schagen . . . Heerhugowaard, but neither the one nor the other corresponds clearly with one of the Noord-Holland groups in Daan's map. Possibly the situation in the time that the recordings were made (1950–1962) has changed compared to the situation at the time that the data for Daan's map was gathered (1939). On the other hand, the dialects of Egmond aan Zee, Volendam and the island Marken (representend by Monnickenwerf in our data set) are known to be independent varieties (Daan, 1956).

We observe that the border between our Zuid-Holland/Utrecht group and the Noord-Holland group exactly matches with the transcriber border (see Figure 9.3). The difference concerns Haarlem only. In our results, this dialect is clustered with the Noord-Holland varieties. In the map of Daan, this variety belongs to group 1 (Zuid-Holland). However, in Daan (1956) the conglomerate of Haarlem is mentioned as a separate group among the varieties of Noord-Holland.

Summarizing, we found five main clusters: the central Gelderland cluster Aalten . . . Woudenberg (see Figure 9.21), the south Gelderland/east Utrecht/east Noord-Brabant cluster Groesbeek . . . Meijel (see Figure 9.21), the west Noord-Brabant cluster Oirschot . . . Fijnaart (see Figure 9.22), the Zuid-Holland/Utrecht cluster Katwijk aan Zee . . . Delft (see Figure 9.22), and the Noord-Holland cluster Hoorn . . . Opperdoes (see Figure 9.10). The southern part of the border between the Zuid-Holland/Utrecht cluster and the west Noord-Brabant cluster matches the border in the map of Daan between group 1 (north) and the groups 7 and 14 (south). The Zuid-Holland/Utrecht cluster and the central-Gelderland cluster just miss bordering on each other. They are separated by the Amersfoort dialect. According to Daan's map, Amersfoort belongs to group 6 and thus was expected to be clustered among the varieties in our Zuid-Holland/Utrecht cluster. However, it is clustered with the south Gelderland/west Utrecht/west Noord-Brabant cluster, probably as the result of transcriber differences.

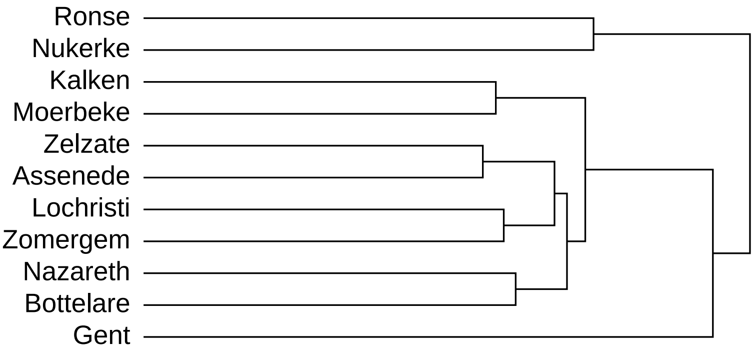


Figure 9.23: Subtree of the dendrogram in Figure 9.5, representing the East Flanders group.

### 9.4.8 Urk

It is striking that Urk belongs to none of the other groups. This can also be seen in the map in Figure 9.4, where Urk has no strong connection with other varieties. The explanation is found in the fact that Urk was an island in the past, until the Noordoostpolder was impoldered in 1942. In the map of Daan, the variety belongs to group 19. The Urk dialect has a regular vowel system where the duration of vowels is relevant. Rounded and close vowels are used more frequently. For the consonants we note *inter alia* that [sk] is pronounced where [sx] is pronounced in Standard Dutch. More about the dialect of Urk can be found in Daan (1990).

### 9.4.9 East Flanders

The group of East Flanders varieties corresponds with the groups 10 and 11 in Daan's map. The dendrogram is given in Figure 9.23. The locations of the varieties is shown in the map in Figure 9.24.

In the dendrogram, the cluster with the dialects of Ronse and Nukerke corresponds with group 10 in Daan's map. The group forms a region between the West and East Flanders dialects. The cluster Kalken ... Gent corresponds with group 11. However, in the northeast in Daan's map the border runs more easterly than it does on our map. As explained in Section 9.4.6, in our results these northeastern varieties are clustered with the varieties that belong to group 12 in the map of Daan (see Figure 9.18). Furthermore, the dendrogram shows that the town dialect of Gent differs rather strongly from the varieties in its surroundings.

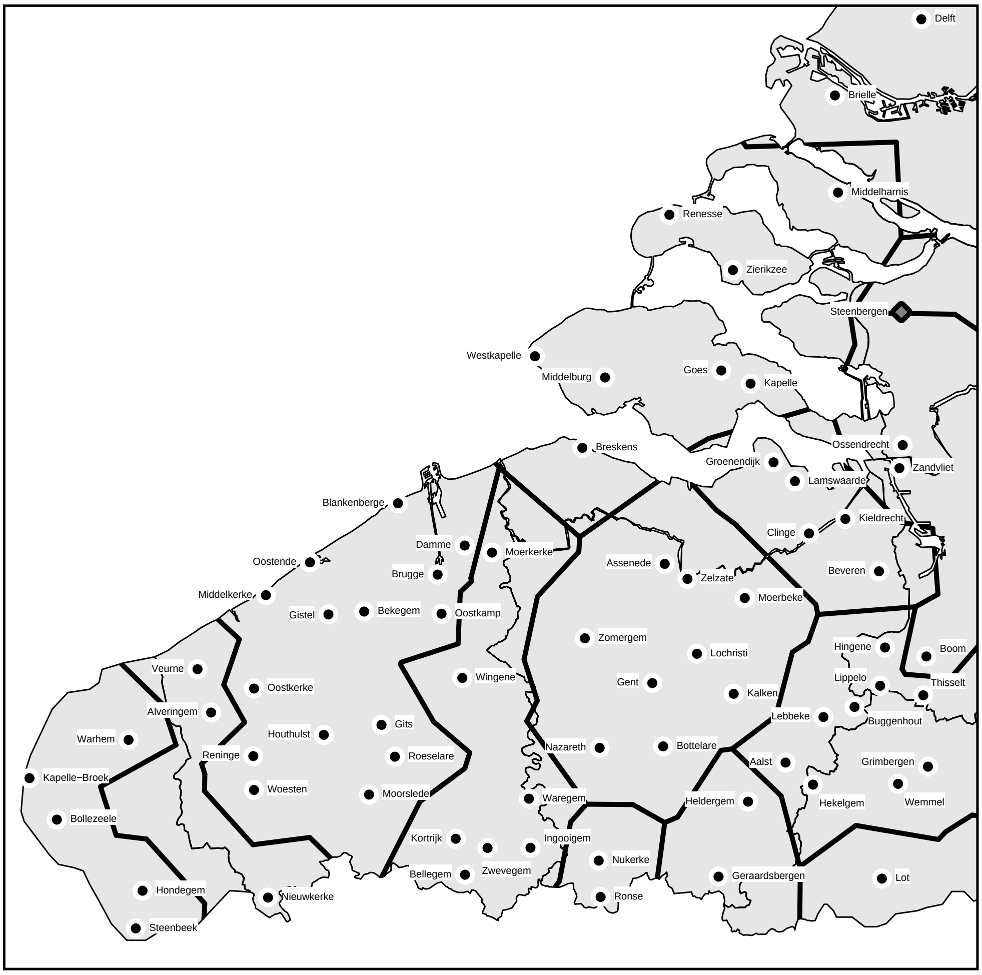


Figure 9.24: The southwestern part of the Dutch language area. National borders and provincial borders are represented by thinner lines and dialect cluster borders by thicker ones. The grey diamond represents a variety that does not actually belong to the group in which it is found geographically.

### 9.4.10 West Flanders

The varieties of West Flanders correspond with group 9 and 10 in the map of Daan. The dendrogram is given in Figure 9.25, and the locations of the dialects in the map in Figure 9.24.

On the highest level, we find a cluster Moerkerke ... Waregem containing varieties along the eastern provincial border of West Flanders, and a cluster Nieuwkerke ... Hondegem containing varieties in the remaining western part. The southern varieties in the eastern group are in or very near group 10 of Daan's map. On our version of Daan's map as given in Figure 9.2 this southern part is visualized. On the original map of Daan, a northern part of group 12 is also found, which is geographically not connected to the southern part. This northern part is not given on our map, since we have no sample sites that fall within this northern part. However, when examining the eastern group in our dendrogram, it is striking to see that it contains varieties both in the North and in the South. Especially the northern varieties of Moerkerke and Wingene are found west of the West Flanders/East Flanders provincial border, while on the map of Daan the northern part of group 10 is only found east of the provincial border. Our results suggest that the Flemish language geographers judge the borders of group 10 too narrowly.

The group of remaining western varieties corresponds with group 9 in Daan's map. A closer division of this area is not given. In our dendrogram the cluster Nieuwkerke ... Alveringem contains varieties easterly along the French Flanders/ West Flanders border. Going one level deeper, the cluster Brugge ... Houthulst contains the other West Flanders varieties, while the cluster Bollezeele ... Hondegem contains the varieties in French Flanders. On the original map of Daan the area of French Flanders is shaded, which suggested some contrast to West Flanders. Our dendrogram show that the varieties on both sides of the political border were rather strongly related when the RND recordings were made.

The cluster Nieuwkerke ... Alveringem appears as a transition zone between the Belgian and the French varieties. At first glance the special position of Veurne and Alveringem may be explained by transcriber differences (see Figure 9.3), but since these varieties form one cluster with Nieuwkerke, we rather think that this cluster represents a transition area.

### 9.4.11 Zeeland

The Zeeland varieties belong for the greater part to group 7 in the map of Daan. The dendrogram is shown in Figure 9.26. The locations of the varieties can be found in the map in Figure 9.24.

In the dendrogram, Breskens is most distinct from the other varieties. On the map of Daan, this variety belongs to group 9, i.e. the dialect of West Flanders and Zeeuws-Vlaanderen. However, it may be possible that the dialect of Breskens

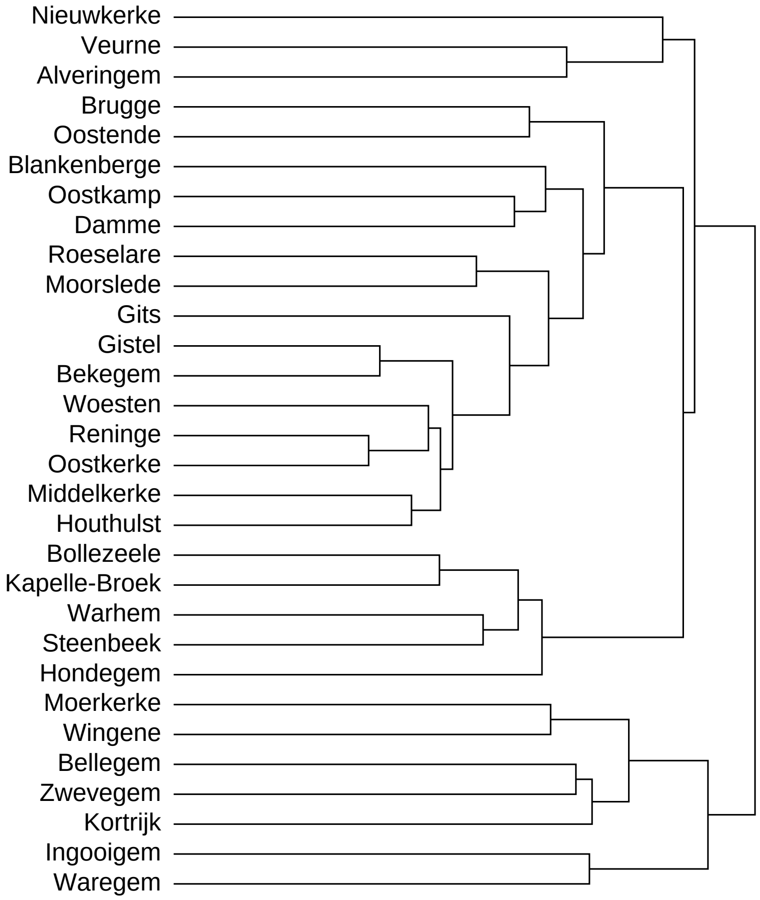


Figure 9.25: Subtree of the dendrogram in Figure 9.5, representing the West Flanders group.

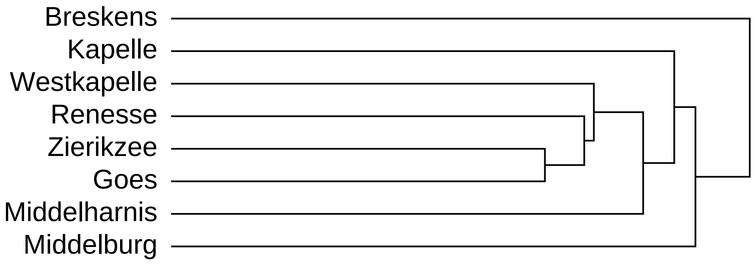


Figure 9.26: Subtree of the dendrogram in Figure 9.5, representing the Zeeland group.

was much more strongly influenced by the Zeeland varieties than other dialects in the same region. This can be explained by the regular ferry lines between Breskens and Vlissingen which started in 1828 and ended only in 2003 when a tunnel connection was completed. Apart from Breskens, the varieties of Goes and Zierikzee form the core in the dendrogram while the other varieties are clustered with them one by one. No clear division per peninsula or island can be found in the dendrogram.

### 9.4.12 Limburg

The varieties in our Limburg group are found in the western part of group 14 and in the northern and eastern part of group 17 in Daan's map. The division of the Limburg varieties is found in the dendrogram in Figure 9.27. The locations of the varieties are shown in the map in Figure 9.17.

In the dendrogram we first find a cluster that excludes the varieties of Steenbergen and Helmond. Steenbergen is found in Noord-Brabant and close to Zeeland (see the map in Figure 9.22). In the two dialects, the uvular [ʀ] is mainly used, but the varieties are geographically located among varieties in which the alveolar [r] is used. In the two varieties, the uvular [ʀ] is mainly used. In the direct surroundings of Helmond the alveolar [r] is used, but this dialect is geographically rather close to the Limburg area, where the use of the uvular [ʀ] is common. In the direct surroundings of Steenbergen the alveolar [r] is used as well, but also in geographically rather distant varieties the alveolar [r] is still used. However, the pronunciation of the /r/ cannot be a sufficient explanation. The [ʀ] is also used in the varieties of Amersfoort and Ravensbergen, which are located among varieties in which the [r] is used (see Figure 9.20). These two varieties do not deviate so strongly from their geographic neighbours in the dendrogram. Therefore, we cannot explain why the Steenbergen dialect is found in the Limburg group, but conjecture that it has to do with migration.

Going one level deeper, the position of Bree is striking. This cannot explained by transcriber differences with certainty since in that case a stronger relation with Budel was expected (compare Figure 9.3). Apart from Bree, we find a northern cluster Budel ... Tegelen and a southern cluster Horn ... Meerssen.

In the northern cluster, we find a western cluster with the varieties of Budel and Overpelt, and a northeastern cluster Rijkevoort ... Tegelen. Considering the varieties of Budel and Overpelt, we see that Budel is found at the Dutch side of the state boundary, and Overpelt at the Belgian side. The cluster of the two varieties forms the northwestern part of group 17 on the map of Daan. The border between these two varieties and the south Gelderland/east Utrecht/east Noord-Brabant cluster in the group of central Dutch varieties (see Section 9.4.7) corresponds with both the western part of the border between group 13 and 14 in Daan's map and the transcriber border.

It is striking that our northeastern cluster does not belong to group 17 (the Limburg dialects) but to group 14 (dialects of Noord-Brabant and northern Limburg) in Daan's map (see Figure 9.2). In the map of Daan this northeastern cluster is not found as a separate area, it is only a part of group 14. However, in the map of Te Winkel (1901) the eastern part of group 14 is suggested to be a separate dialect area, labeled as 'Saksisch-Oostfrankisch' (Saxon-East Franconian). The southern part of this area corresponds with our northeastern group, although the southern border of our cluster is found more south. Our more southern border coincides with a transcriber border. The western and northern border of the northeastern cluster also coincides with a transcriber border, with the exception of Meijel which was is found in the south Gelderland/east Utrecht/east Noord-Brabant cluster in the group of central Dutch varieties. In contrast to the dialects in the dendrogram of the Limburg group, in the dialect of Meijel mainly the alveolar [r] is used rather than the uvular [ʀ]. The western border is also found in the map of Te Winkel (1901).

The southern cluster is a part of group 17 of the map of Daan. The southern part of the southern cluster is bounded on the west side by the state boundary which coincides almost with the Maas river. On the west side of this boundary, our Southwest Limburg group is found (see Section 9.4.5). However, this border is not found on the map of Daan. On the one hand, this is a transcriber border, but on the other hand, this boundary is also found in the map of Te Winkel (1901). As mentioned in Section 9.4.5 the border coincides with an [r]/[ʀ]-isogloss.

### 9.4.13 Northeast Luik

The northeast Luik group covers the furthest southeastern part of group 17 in Daan's map. The dendrogram is given in Figure 9.28. The locations of the varieties can be found in the map in Figure 9.17. The varieties of Aubel, Baelen, Eupen and Raeren are actually found south of group 17 and belong to the province of Luik. Aubel and Baelen belong to the French language area, and Eupen and Raeren belong to the German language area.

When considering group 17 in the map of Daan it may be unexpected that the varieties in this group do not form one group with the varieties in the Limburg group. However, it is known that the situation in Limburg is complex, and the varieties do not form a homogeneous group (Hoppenbrouwers and Hoppenbrouwers, 2001, p. 187). Most varieties in our Northeast Luik group are found east of the isogloss that represents the opposition between *zeggen* 'to say' (west, the [ɛ] is used) and *sagen* (east, the [a] is used, see the map in Goossens (1977) on p. 21 and 60). In the dendrogram, we find a western cluster 's-Gravenvoeren ... Eupen, and an eastern cluster Raeren ... Kerkrade. The division in these two clusters perfectly reflects the Benratherlinie. This isogloss represents the opposition between Dutch/Low German *maken* 'to make' (west, the [k] is used) and High German *machen* (east, the [x] is used). The dialects west of the Benrather-

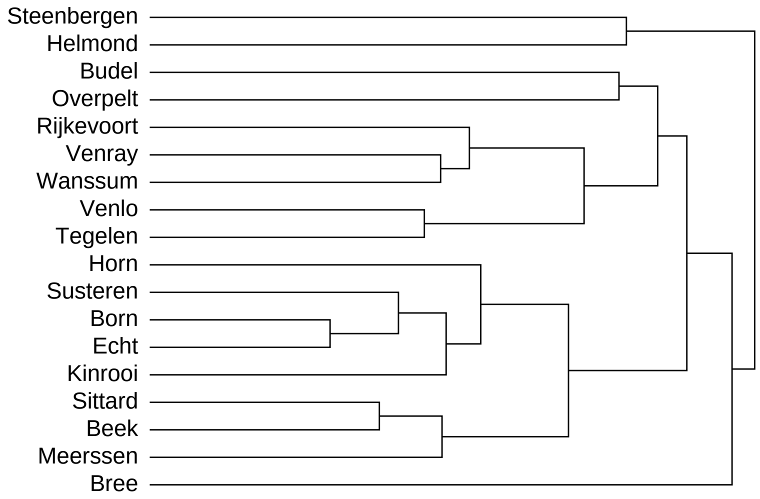


Figure 9.27: Subtree of the dendrogram in Figure 9.5, representing the Limburg group.

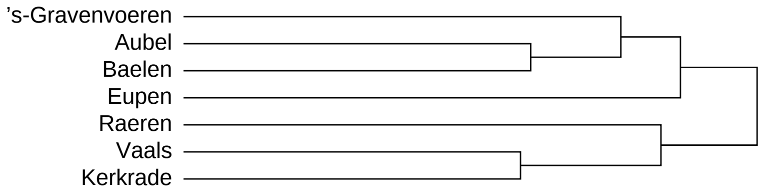


Figure 9.28: Subtree of the dendrogram in Figure 9.5, representing the Northeast Luik group.

linie belong to the East Limburg-Ripuarisch transition zone and the dialects east of the Benratherlinie belong to Ripuarisch, a dialect area around the German city of Cologne (see the maps in Goossens (1977) on p. 21 and 60 again).

## 9.5 Continuum

### 9.5.1 Multidimensional scaling

In addition to cluster analysis, we also applied multidimensional scaling to the distances between the 360 RND varieties (see Section 6.2). It appears that one dimension explains 36% of the variance, two dimensions 52%, three dimensions 88%, four dimensions 92%, five dimensions 95%, six dimensions 96% and seven dimensions 97%. These percentages make clear that a rather good representation is already obtained when using three dimensions. When using more dimensions only a rather small improvement of the variance is obtained. Therefore, we examine the three-dimensional solution.

Examining the three dimensions, we found that the first dimension distinguishes between Frisian and Low Saxon varieties on the one hand (low values), and Low Franconian varieties on the other hand (high values). The second dimension distinguishes between Frisian on the one hand (low values), and Low Franconian (high values) and Low Saxon (even higher values) on the other hand. The town Frisian varieties and the dialect of *het Bildt* are intermediate between Frisian and Low Franconian, but they are closer to Low Franconian than to Frisian in this second dimension. The *Stellingwerf* varieties are intermediate between Frisian and Low Saxon, but they are closer to Low Saxon than to Frisian. The third dimension divides the Low Franconian varieties in three groups. The first group (low values) contains the Dutch Limburg varieties. The second group (mean values) contains varieties in Belgian Limburg, Brabant, Antwerpen and the greater part of the Netherlandic Low Franconian area. The third group (high values) contains varieties in East Flanders, West Flanders and also some Zeeland varieties. The third dimension does not divide either the Frisian or the Low Saxon varieties, all of them belong to the second group.

We investigate which phenomena are especially responsible for each dimension. For this purpose we calculate distances between varieties per dimension. When e.g. two varieties have respectively the values $x$ and $x'$ in a dimension, the distance is equal to $|x - x'|$. In this way, for each pair of varieties the distance for one dimension is found. Having 360 varieties, we get $(360 \times 359)/2$ distances. In Section 5.1 we explained how we applied Levenshtein distance using transcriptions. Using Levenshtein distance, a distance matrix can be obtained, containing Levenshtein distances between the different pronunciations of one particular word. This matrix also contains $(360 \times 359)/2$ distances.

When we have calculated multidimensional scaling distances per dimension on the one hand, and Levenshtein distances per word on the other hand, the two sorts of distances can be correlated. The stronger the Levenshtein distances correlate with the distances of one dimension, the more the variation of the corresponding word contributed to the values of that dimension. For each of the 125 words we calculated the Levenshtein distances between the 360 varieties. This results in 125 matrices. Subsequently, each of the matrices was correlated with the distances derived from the first, second and third dimension, respectively.

It appears that the distances in the first dimension correlate strongest with distances obtained on the basis of equivalents for *waren* 'were' ($r = 0.70$). The variation of this word is shown in the map in Figure 9.29. In the north we find forms like [ʋaːʳn̩] or [ʋɑdn̩] and in the south forms like [ʋa·rə], [wʊ·rə] or [wa·ʀə]. In the furthest southwest, we find forms like [wɑ̝rn] or [wʊ·ʀn]. So the forms in the north and in the furthest southwest end on [n] or [n̩] and the forms in the south on [ə]. In some other strongly correlating words, the same phenomenon was found. We conjecture therefore that the treatment of the weak syllable /ən/ is the single most significant dialect marker in Dutch. The two types of endings corresponds with the division in Frisian and Low Saxon varieties on the one hand (north), and Low Franconian varieties on the other hand (south), as represented by the first dimension. However, the stronger relation between the Low Saxon varieties and the Low Franconian varieties in the furthest southwest was not found in the first dimension, which is possibly the main explanation for the fact that a perfect correlation was not found.

Distances in the second dimension correlate strongest with distances obtained on the basis of equivalents for *vader* 'father' ($r = 0.64$). The variation of this word is shown in the map in Figure 9.30. In the northwest we find the Frisian forms like [hɛⁱt] and [hɑⁱt], and the town Frisian form [vɔ̞ᵊdər]. In the remaining part we find forms like [va·dər], [vʊ·dər] and [vɔ·dəʀ]. The two different lexical forms clearly correspond with the division between Frisian and non-Frisian varieties as represented by the second dimension. Other strongly correlating words represent both lexical and phonological differences.

Distances in the third dimension correlate strongest with distances obtained on the basis of equivalents for *breder* 'broader' ($r = 0.56$). The variation of this word is shown in the map in Figure 9.31. At first glance the word seems to divide the area in north (using [d]) and south (the [d] is substituted by the [j] or even deleted). However, the correlation with the distances in the first dimension was much lower ($r = 0.27$). Another phenomenon that catches the eye is that in most varieties the alveolar [r] is used in forms like [bre·dər], [bre·jər] or [briːᵊ]r], while the uvular [ʀ] was used in the Limburg varieties (southeast) in forms as [bʀɛ·ⁱəʀ]. From Figure 4.7 we may conclude that the difference between [r] and [ʀ] weight

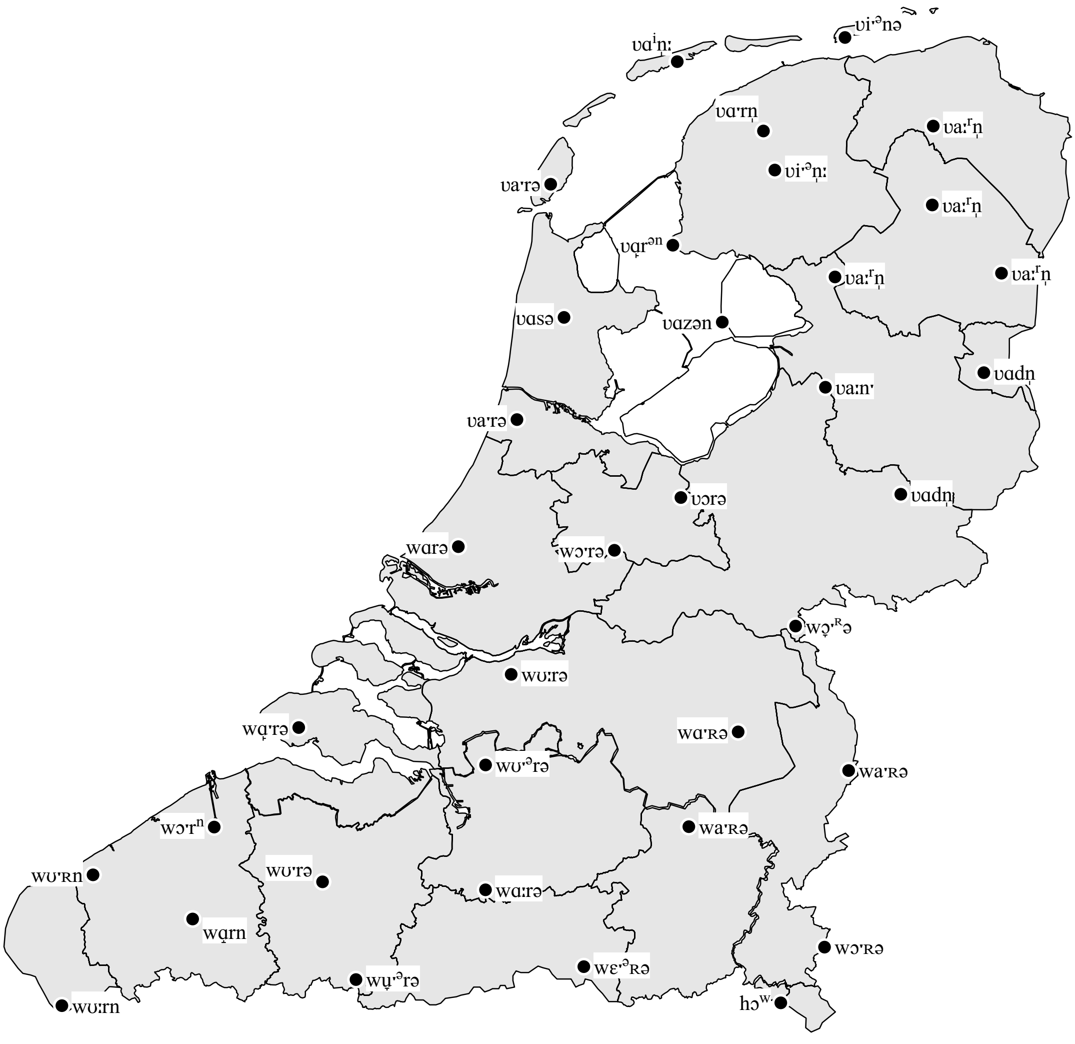


Figure 9.29: Variation of the equivalents for *waren* ‘waren’. The transcriptions correspond with the labels in Figures 9.32 and 9.33. Extra-short sounds are noted in superscript. Distances among 360 Dutch varieties as found on the basis of the first dimension of a three-dimensional MDS solution correlate most strongly with distances obtained on the basis of pronunciations of this word ($r = 0.70$).

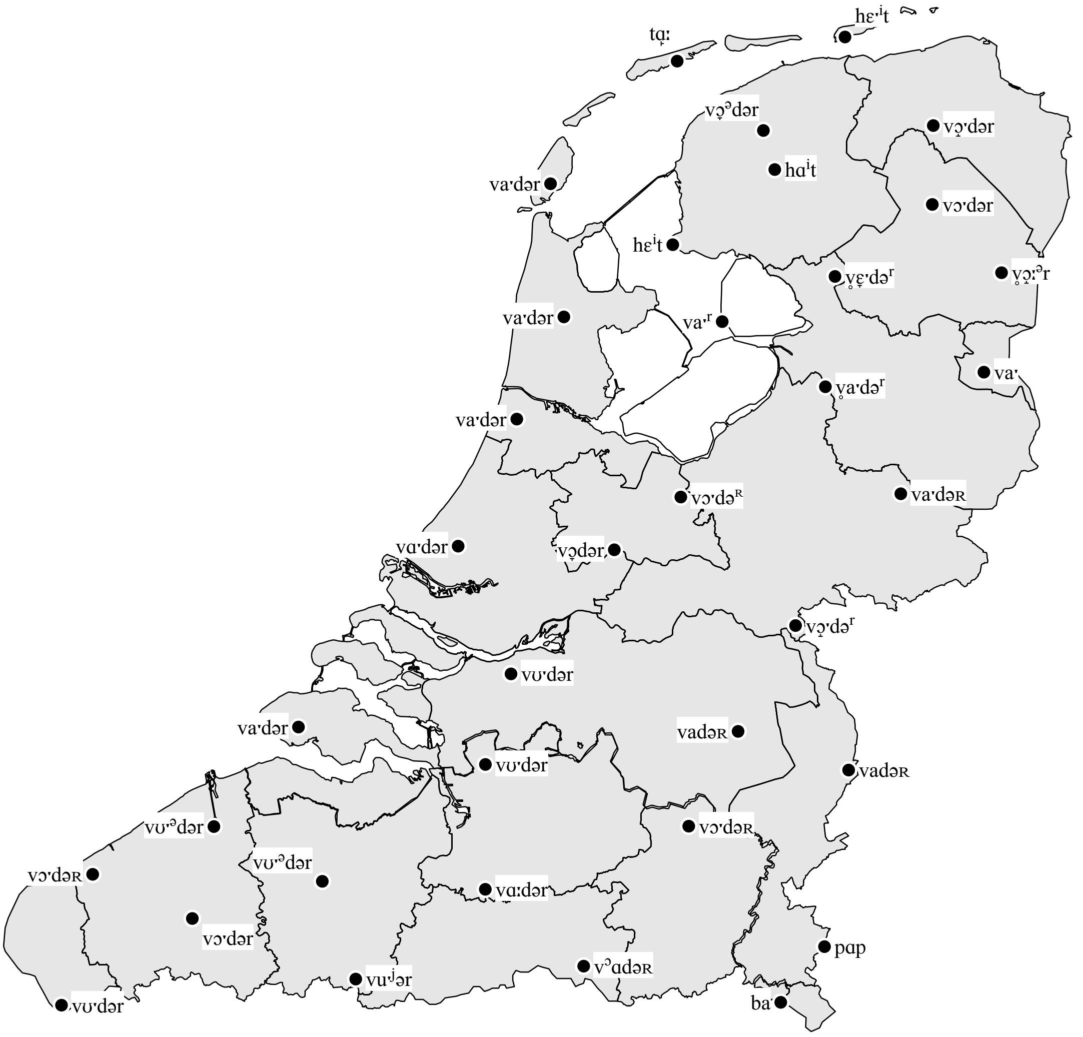


Figure 9.30: Variation of the equivalents for *vader* 'father'. The transcriptions correspond with the labels in Figures 9.32 and 9.33. Extra-short sounds are noted in superscript. Distances among 360 Dutch varieties as found on the basis of the second dimension of a three-dimensional MDS solution correlate most strongly with distances obtained on the basis of pronunciations of this word ($r = 0.64$).

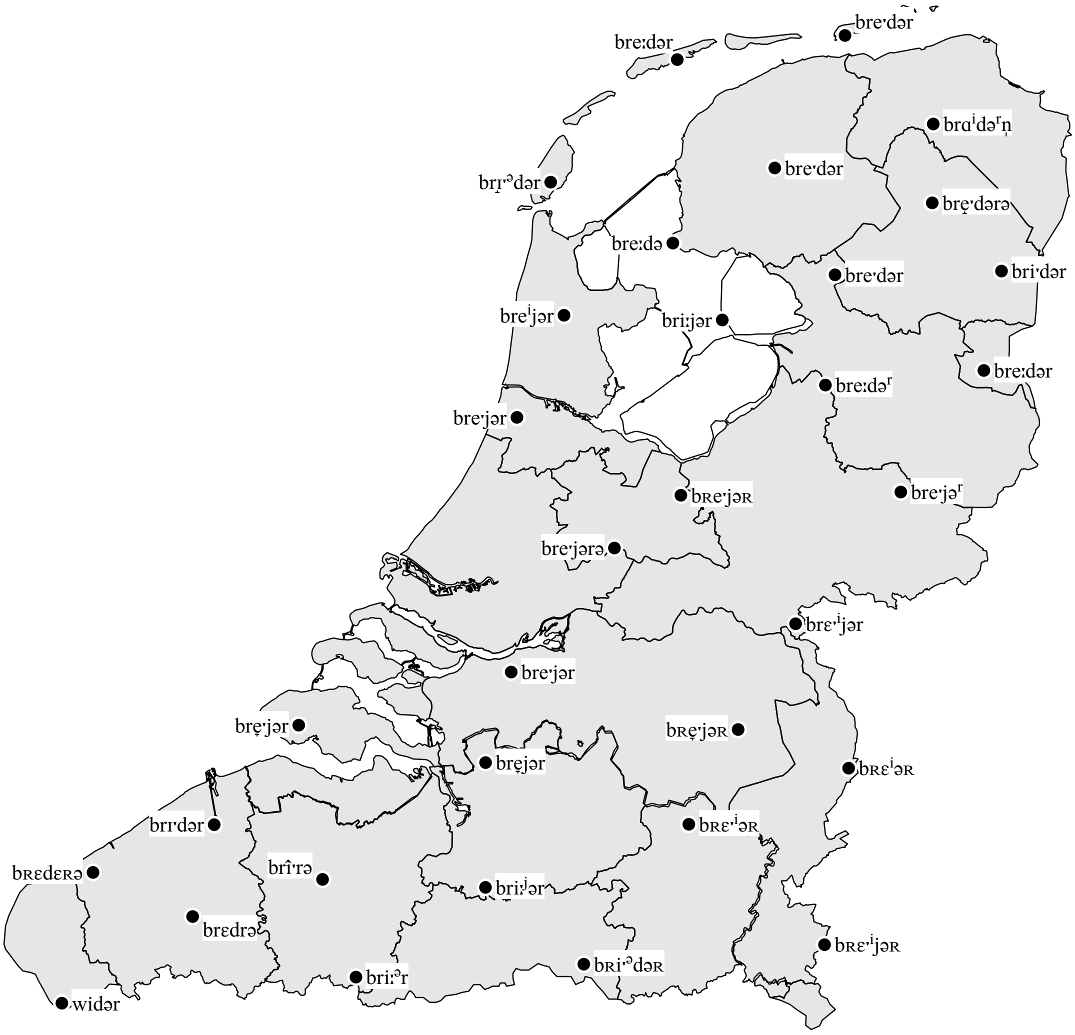


Figure 9.31: Variation of the equivalents for *breder* 'broader'. The transcriptions correspond with the labels in Figures 9.32 and 9.33. Extra-short sounds are noted in superscript. Distances among 360 Dutch varieties as found on the basis of the third dimension of a three-dimensional MDS solution correlate most strongly with distances obtained on the basis of pronunciations of this word ($r = 0.56$).

more heavily than the difference between [d] and [j]. The [r]/[ʀ] distinction divides only the southern part of the Dutch language area into two parts: an eastern and a western part. This accords with the third dimension that represents an east-west dimension as well. However, the third dimension distinguishes between an east, central and west group, while the [r]/[ʀ] difference distinguishes only between an east and west area. This difference can be explained by the fact that the third dimension is based not only on the [r]/[ʀ] difference, but rather on the aggregate of several phenomena. Nonetheless, the [r]/[ʀ] difference is one of the most important phenomena as appears from the strong correlation. In other strongly correlating words the same phenomenon was found as well.

### 9.5.2 Continuum map

As described in Section 6.2.4 on the basis of the three dimensions of the three-dimensional multidimensional scaling solution, each variety can be represented by a color. The three dimensions determine the intensities of red, green and blue. We used this approach to create a map in which each variety gets its own unique color. We assign the colors to the three dimensions so that the color scheme of our map approaches the color scheme in Daan's map maximally. The first dimension represents the intensity of red, the second dimension inversely the intensity of blue, and the third dimension inversely the intensity of green.

In Figure 9.32 a color map based on three MDS dimensions is shown. On this map dialect points are blown up to small areas until they border each other (see Section 6.2.4). However, dialect islands are not blown up, but represented by diamonds, just as in the map in Figure 9.6. For an explanation about the dialect islands see Section 9.3. To keep a clear picture, the same restricted set of labels of (in general) better-known locations is printed in the map as in the map in Figure 9.6.

On the map, the Frisian area (northwest) is represented by bright blue, the Low Saxon area (northeast) is green. The Netherlandic Low Franconian area is colored by different grey shades. Note that the town Frisian varieties (most diamonds and the island Ameland north of Leeuwarden) have a color intermediate between grey and blue. The same color is found for *het Bildt* (northwest of Leeuwarden). The Frisian mixed varieties of Tjalleberd (the higher diamond south of Grouw) and Donkerbroek (the diamond intermediate between Grouw and Assen) are more greenish. The 'pure' Frisian variety of Appelscha (west of Assen) is colored blue, just as are the other 'pure' Frisian varieties. Turning now to the southern part of the Low Franconian area, the West Flanders varieties (in the furthest southwest) are colored darker red, the East Flanders varieties

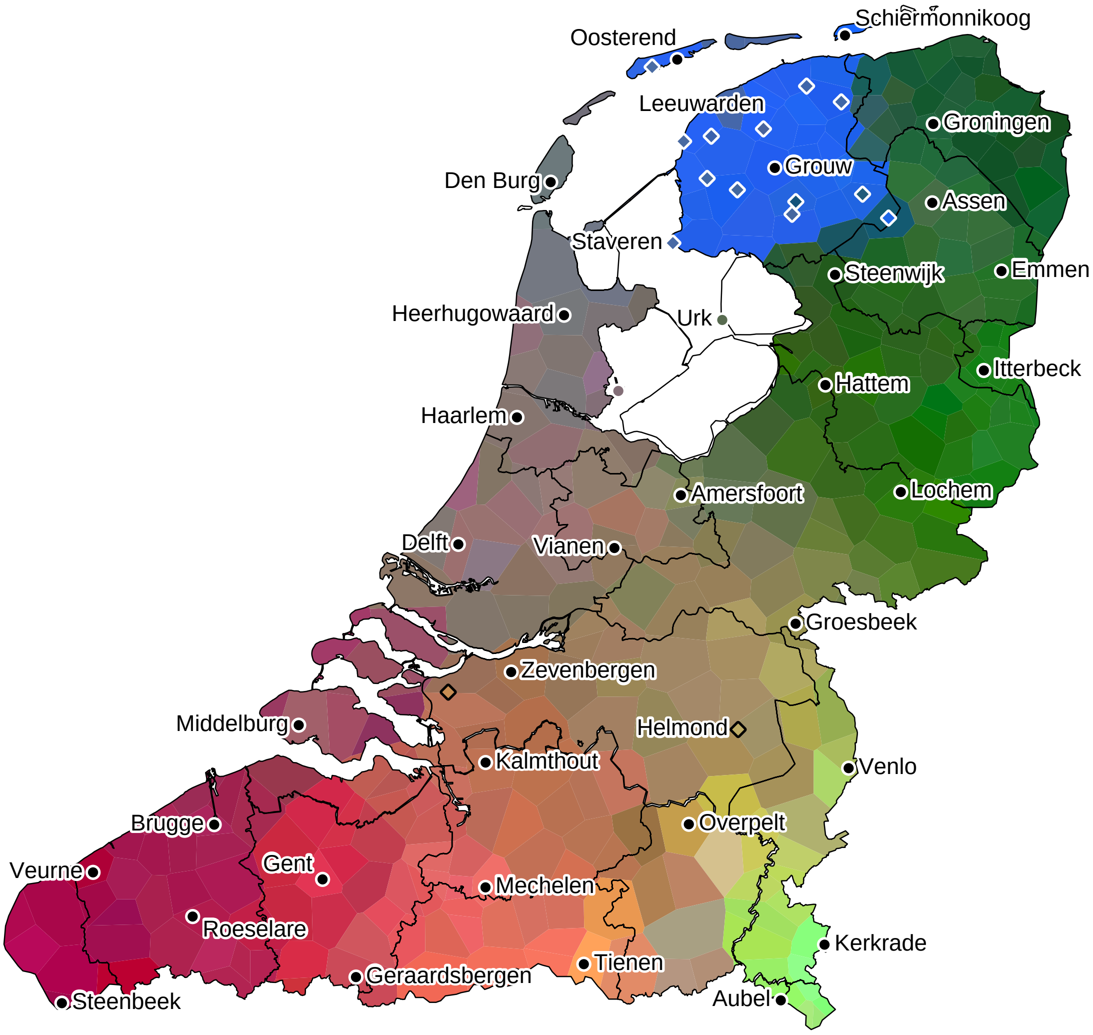


Figure 9.32: Dialect variation represented by color. The first MDS dimension is mapped to red, the second inversely to blue and the third inversely to green. Kruskal's Non-metric MDS is used (see Section 6.2.2). Each of the dialect points is blown up to a small area, except the dialect islands that are marked with diamonds.

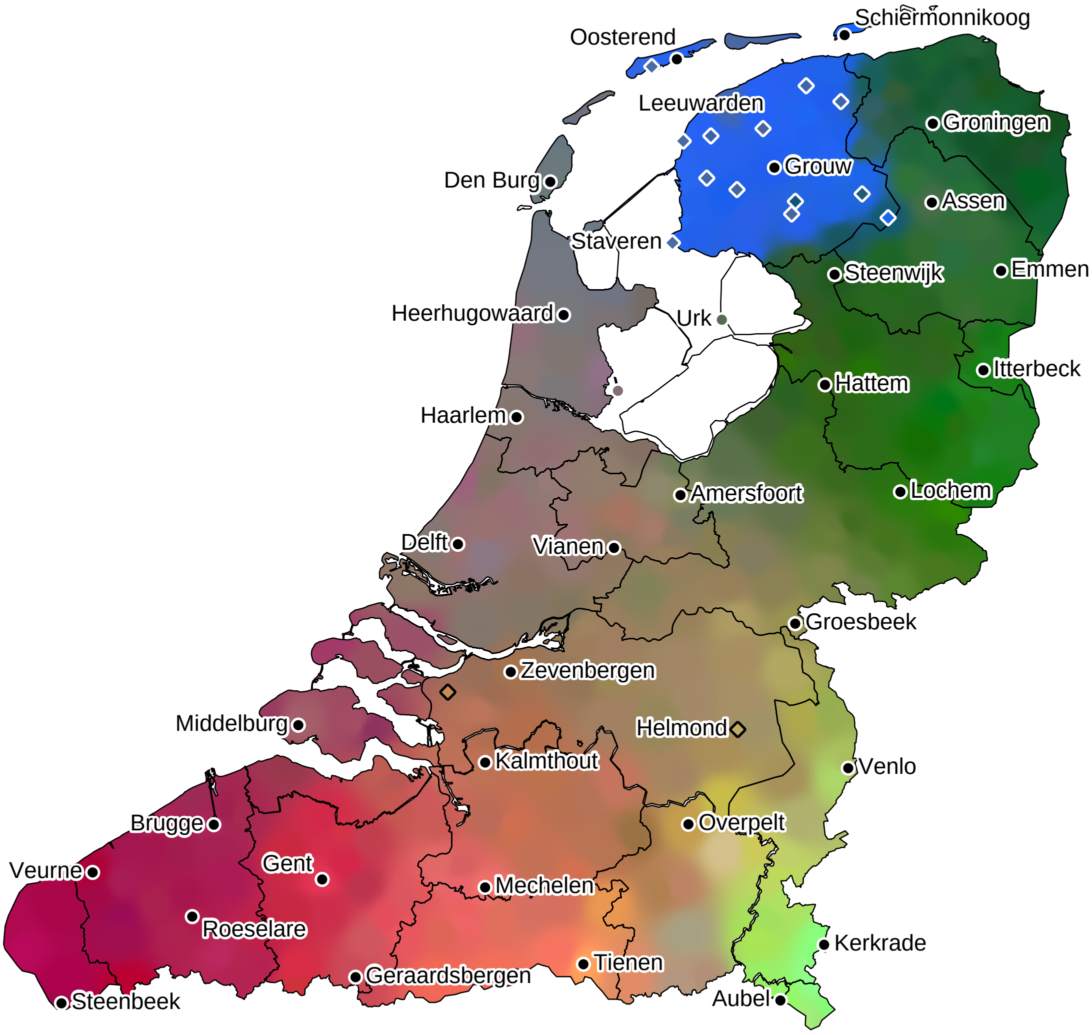


Figure 9.33: Dialect variation represented by color. The first MDS dimension is mapped to red, the second inversely to blue and the third inversely to green. Kruskal's Non-metric MDS is used (see Section 6.2.2). The color of intermediate points is determined by interpolation using Inverse Distance Weighting. Dialect islands are marked with a diamond. They are not involved in the interpolation process.

(around Gent) bright red. The Antwerpen area (around and south of Kalmthout) is greyish red, the Belgian Brabant area lighter red. The varieties in the southern part of Limburg are colored lighter green. In the map the colors of the diamonds of Steenbergen (west of Zevenbergen) and Helmond are a little bit more like the colors of the Limburg varieties. However, they do not contrast as strongly with their surroundings as might be expected on the basis of the findings in Section 9.4.12. Although results of cluster analysis and multidimensional scaling are in accordance with each other in general, this example shows that minor differences can be found. Therefore, it is useful to show results obtained with both techniques. Broadly speaking our color map is similar to the map of Daan.

In Figure 9.33 a map is given that is related to the map in Figure 9.32. However, in this map the space between the points is colored on the basis of MDS values which are found by interpolation using Inverse Distance Weighting (see Section 6.2.4 again). This map consequently does justice to the idea that the dialect landscape may be regarded as a continuum. Dialect islands are of course not involved in the interpolation process.

## 9.6 Relation to Standard Dutch

In the same way as distances are calculated between dialects, distances between dialects and a standard language can be calculated. For the RND material we calculated distances with respect to Standard Dutch. The results are shown in Figure 9.34. In the map the dialects are colored according to the rainbow. The most similar dialects are red, followed by orange, yellow, green and lighter blue. The dark blue dialects are most distant.

Exactly in accordance with the general opinion, Haarlem is closest to Standard Dutch. The dialect has a distance of 14.7%. The way in which percentages are found is explained in the Sections 5.1.8 and 5.1.10. Haarlem was followed by Brielle (16.8%, south of Delft), Hoorn (16.9%, east of Heerhugowaard), Warmond (17.0%, south of Haarlem), Heemskerk (17.2%, north of Haarlem) and Vianen (17.5%).

Most distant was the dialect on the island Schiermonnikoog, north of the border between Friesland and Groningen. The variety has a distance of 44.9%. Although the Schiermonnikoog variety is Frisian, we found in Section 9.4.1 that it has a distinct position among the other Frisian varieties. However, Schiermonnikoog is followed by the other ('pure') Frisian varieties. The 'pure' Frisian variety most like Standard Dutch was West-Terschelling (40.9%, west of Oosterend). This may be explained by the fact that the island Terschelling belonged to the province of Noord-Holland until 1942. The high percentages for the Frisian varieties may justify the fact that Frisian is recognized as a second official language in the Netherlands. The dialects of *het Bildt*, the Frisian cities, Midsland, and Ameland Island are clearly less distant from Standard Dutch. Most distant was Franeker (33.0%, west of Leeuwarden), most similar was Hollum (29.5%, on the island Ameland north of Leeuwarden).

Besides Frisian, four Limburg dialects were also very distant to Standard Dutch: Vaals (43.2%, south of Kerkrade), Kerkrade (42.2%), Raeren (41.8%, south of Vaals) and Aubel (41.6%). Also in the West-Flemish dialect area we found distant dialects: Alveringem (42.3%, south of Veurne), Warhem (41.0%, southwest of Veurne), Reninge (40.9%, southeast of Alveringem) and Veurne (40.7%). In the province of Overijssel, we found the dialects of Vriezenveen (41.8%), Rijssen (40.8%) and Wierden (40.4%, the three dialects are found southwest of Itterbeck) to be rather distant. The special position of Vriezenveen was already found in Section 9.4.4. Although the Groningen dialects belong to the more distant dialects (indicated by lighter blue), they are not as distinct as for example the Frisian varieties and the four Limburg varieties that we mentioned above. Most distant were Finsterwolde (41.0%, east of the city of Groningen) and Onstwedde (40.4%, east of Assen).

In Hoppenbrouwers and Hoppenbrouwers (2001) the authors also discuss a ranking with respect to Standard Dutch (pp. 124–131). Distances are obtained by means of the feature frequency method (see Section 2.3.2). The scores in the rank order list are divided into 23 intervals. Frisian is found in interval 5 and 6, which suggest that Frisian is rather close to Standard Dutch. Town Frisian varieties are not obviously closer to Standard Dutch than ‘pure’ Frisian varieties, both are found in interval 5 and 6. This is clearly different from our results, where the ‘pure’ Frisian varieties form a group of the most distant varieties, while the town Frisian varieties are obviously more related to Standard Dutch (see Figure 9.34: darker blue versus green). In the results of Hoppenbrouwers and Hoppenbrouwers the Groningen dialects are found in the intervals 13, 15 and 16, suggesting that they are much more distant than the Frisian varieties. In our results, both, Frisian and Groningen dialects are rather distant, but Frisian is more distant (Figure 9.34: darker blue versus lighter blue). Our results agree with those of Hoppenbrouwers and Hoppenbrouwers that the dialect of Kerkrade and neighboring varieties belong to the most distant ones.

The comparison of our results with those of Hoppenbrouwers and Hoppenbrouwers shows that the use of Levenshtein distance gives different results than the feature frequency method. Especially when looking at the Frisian dialects, our results are much more in accordance with the prevailing opinion and especially the opinion of the Frisians themselves. Using the Levenshtein distance, words are regarded as linguistic units while the order of segments in a word is also considered. The example of the Frisian underscores the importance of these two aspects.

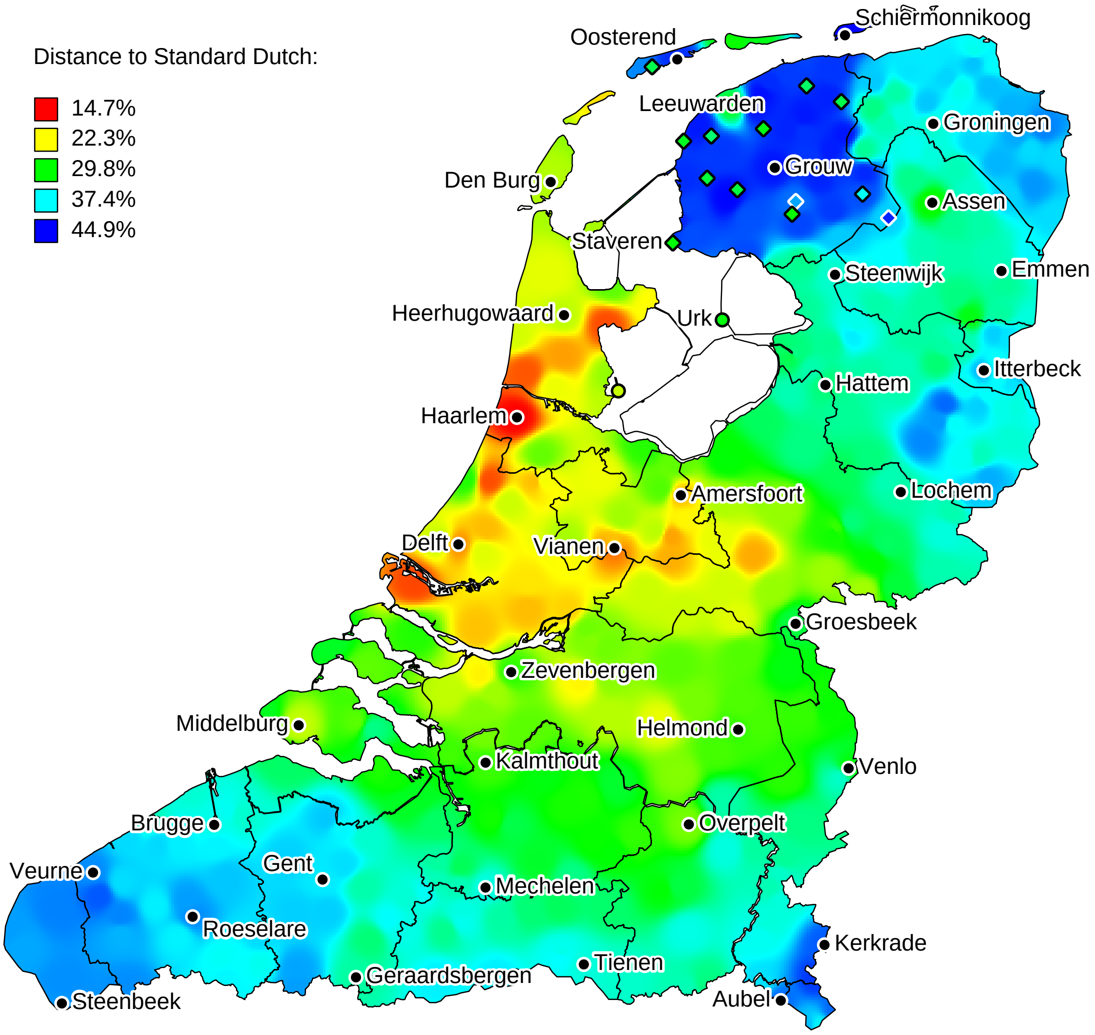


Figure 9.34: Distances with respect to Standard Dutch. The color between the sample points was found by interpolation. Diamonds represent dialect islands. Closest to Standard Dutch is Haarlem (14.7%), most distant was Schiermonnikoog (44.9%).

## 9.7 Conclusions

We calculated distances between 360 Dutch varieties. Data was taken from the RND. We compared the results with Daan's map (see Figure 9.2). However, since the RND was compiled by 16 different transcribers, we also kept track of transcriber borders (see Figure 9.3).

Comparing our division of the main 13 dialect groups with the map of Daan, we found similarities and differences. We found that especially our Groningen and Overijssel group and our group of central Dutch varieties were divided into different groups in the map of Daan. The opposite result was also found. The Limburg group in the map of Daan was divided into three different groups on our map. We explain the differences by the different weighting of dialect contrasts by dialect speakers, whose judgements formed the basis of Daan's divisions. Social and economic structures may influence judgements of dialect similarity. Furthermore, in a homogeneous area differences can be more easily identified than in a heterogeneous area. This may explain why e.g., the Overijssel area is divided into several groups in the map of Daan while the Limburg area is found to be one group.

Both when examining the 13 groups and when examining each of the groups in more detail, we find real dialect borders and transcriber borders. Sometimes, dialect borders and transcriber borders nearly coincide. For these borders no obvious conclusion can be drawn. Although transcriber differences were normalized to some extent, we did not succeed in eliminating them fully, as appears in the results. Normalizing transcriber differences is risky, since it is not always clear whether differences in notation reflect transcriber differences or dialect differences. Advisable future work would be to calculate distances between Dutch varieties again on the basis of data from the *Fonologische Atlas Nederlandse Dialecten* (FAND) (Goeman and Taeldeman, 1996; Goossens et al., 1998, 2000). This material is known to be of higher quality.

We also calculated distances with respect to Standard Dutch. Just as in Hoppenbrouwers and Hoppenbrouwers (2001) we found the dialect of Haarlem to be most similar to Standard Dutch. In our results, the Frisian varieties were most distant, while in the results of Hoppenbrouwers and Hoppenbrouwers (2001) the Frisian varieties were relatively related to Standard Dutch. We think that our results are most in accordance with linguistic reality and with general opinion. This difference shows that it is important to regard words as linguistic units and to consider the order of segments in word pronunciation. With the Levenshtein distance used in our research these two aspects are taken into account.

# Chapter 10

# Conclusions and future prospects

## 10.1 Conclusions

The goal of this thesis is to explore whether Levenshtein distance could be a useful tool for measuring dialect word pronunciation distances, and thus for measuring dialect distances. Since we want to be able to compare the Levenshtein distance with the corpus frequency method and the frequency per word method, the two latter methods were also involved in the research. In Section 2.4 we concluded that the frequency per word method is methodologically better than the corpus frequency method, and the Levenshtein distance is methodologically better than the frequency per word method.

When attempting to quantify distances in pronunciation between dialects, we need to determine the relations between different speech segments. For this purpose we investigated discrete representations of segments (Chapter 3) and acoustic representations of segments (Chapter 4). The *phone* representation is the least discriminating discrete representation. Two segments are equal or unequal. Using feature representations gradual segment distances can be obtained. We examined the feature systems of Hoppenbrouwers & Hoppenbrouwers (H & H), Vieregge & Cucchiarini (V & C) and Almeida & Braun (A & B). When correlating segment distances obtained on the basis of these systems we found that the systems of V & C and A & B appeared to be most similar, although the correlations between these two systems were not significantly stronger than comparable correlations between any other pair of systems. Even these systems are different, however, as indicated by the rather low, but significant correlations. The use of the different systems will yield different results. We also investigated different metrics for comparing feature histograms (used in frequency-based methods) and feature bundles (used in Levenshtein distance). The correlations between the Manhattan metric and the Euclidean metric were stronger than the corresponding correlations between any other pair of metrics. Partly they were also significantly stronger. This indicates that results obtained on the basis of

Manhattan distance will not strongly differ from results obtained on the basis of Euclidean distance. The Pearson correlation coefficient appeared to be rather different from the two other metrics.

The acoustic segment representations we examined were the Barkfilter representation, the cochleagram representation and the formant track representation. These representations are more perceptually oriented than the commonly used type of spectrogram which has a Hertz-scale. On the basis of the distances obtained by the different representations we applied multidimensional scaling and scaled the distances to two dimensions. For the vowels we obtained a vowel quadrilateral (Barkfilter and cochleagram) or vowel triangle (formant tracks) and for the consonants a distinction between the different manners of articulation. These were reasonable results, although they were based on only two speakers. We conclude that the use of acoustic representations is useful, but recommend future work to verify the conclusion on the basis of more speakers, and if necessary to refine the acoustic processing. When comparing the present results of the different representations, we found that the Barkfilter results and the cochleagram results correlate significantly more strongly than the other pairs of representations. The formant track results appeared to be more different, so the use of the formant track representation will yield significantly different results than when using the Barkfilter or cochleagram representation.

We compared the acoustic representations with the feature-based representations. For vowels we found that the Barkfilter distances and the cochleagram distances correlate strongest with the A & B distances, but correlations with other feature systems were not significantly weaker. The formant track distances correlate strongest with the V & C distances. However the correlations were mostly not significantly higher than comparable ones of other feature systems. For consonants the Barkfilter distances and the cochleagram distances correlate strongest with the V & C distances, but only significantly stronger than with the A & B distances. The formant track distances correlate strongest with the V & C distances (RND) or H & H distances (IPA). The correlation coefficients were only significantly higher than the comparable ones of A & B. The correlations between acoustic distances and feature-based distances were not extremely high, although they were mostly significant. Therefore, both types of segment representations were considered in validation work.

In Chapter 5 we described Levenshtein distance. The Levenshtein distance of two word pronunciations is equal to the set of operations with the least cost which changes the one pronunciation into the other. We used insertions, deletions and substitutions. Future work may be to add the swap operation and to find the correct weight for this operation. The distance between two dialects is equal to the average word distance. We apply the Levenshtein distance to transcriptions where operations are applied to the transcription segments, and to recordings where operations are applied to spectra or formant bundles.

Once the distances between dialects are obtained, the dialects can be classified. Cluster analysis and multidimensional scaling are explained in Chapter 6. We examined several cluster methods and found that *Unweighted Pair Group Method using Arithmetic averages* to be methodologically superior to the other methods. We examined three different multidimensional scaling algorithms and found *Kruskal's Non-metric Multidimensional Scaling* preferable, since the results of this procedure represent the original distances with the greatest fidelity.

In Chapter 7 we validated different versions of the corpus frequency method, the frequency per word method and the Levenshtein distance on the basis of 15 Norwegian varieties. We used recordings and transcriptions of the fable 'The North Wind and the Sun', in Norwegian: 'Nordavinden og sola' (NOS). The data was compiled by Jørn Almberg. In the text we found 58 different words. In advance consistency was checked for the word-based methods. We calculated Cronbach's $\alpha$ values and found that the 58 words were enough to obtain reliable results. For one particular Levenshtein variant we found that the use of only 25 words gave already an acceptable degree of consistency ($\alpha = 0.70$).

Subsequently, for both transcription-based methods and recording-based methods we compared the measurements with the results of a perception experiment in which dialect speakers themselves judge the distances between the varieties. Examining the transcription-based methods it appeared that results obtained by methods using phones and the logarithmic Levenshtein distances using acoustic representations correlate most strongly with the perceptual distances. At first glance, this may be a partly unexpected outcome, but the methods share the property that small segment distances are relatively heavily weighted, which is perhaps also the case in perception. Among the feature representations, the H & H system yields the best results. Among the acoustic representations we found the Barkfilter representation better than the other two representations, but only when using the linear Levenshtein distance. Furthermore, we found that the use of 4 length gradations is preferable to 2 length gradations in general.[1] The computations did not clarify whether two-segmental representations of diphthongs are better than one-segmental representations, or the other way round. When representing speech segments by features, Manhattan is mostly preferable when using the corpus frequency method, and Euclidean is the better candidate when applying word-based methods. Using the Euclidean metric larger differences are weighted relatively more heavily than smaller differences. When using the corpus frequency method, dialect distances are measured with the metrics. Using the frequency per word method and Levenshtein distance, respectively word distances and segment distances are calculated with these metrics. This indicates that on the highest level (comparison of dialects) differences should

[1] When using 4 length gradations extra-short, short, half-long and long are represented by multiplying segments in the transcription. When using 2 length gradations only extra-short and non-extra-short are represented by multiplying segments in the transcription.

be weighted equally, but on the lower levels (comparison of words or segments) larger differences should be weighted relatively more heavily than smaller ones. The best method is a variant of the transcription-based Levenshtein distance, where segment distances are found on the basis of Barkfilter segment distances, four length gradations are used, diphthongs are represented as a sequence of two segments, and logarithmic segment distances are used.

Examining the recording-based methods we found that the three Levenshtein variants gave less satisfying results. We explained this by the rough way in which word length was normalized and by diversity in voice quality. For the first problem solutions may be found in the field of automatic speech recognition (ASR). The second problem may be solved by using a large number of speakers per variety instead of exactly one speaker, as we did until thus far.

Other future work may consist of carrying out a perception experiment with many more varieties. When validating on the basis of a denser sampling, minor differences may nonetheless emerge clearly.

The best method we found in Chapter 7 was applied to a larger set of 55 Norwegian varieties in Chapter 8. The results were analyzed by clustering and multidimensional scaling. When comparing our results to the authoritative map of Skjekkeland (1997), we found some minor and some major differences. On the one hand, this may be the result of the choice of the 58 words. A better approach would be to select the words randomly from a corpus as Bolognesi and Heeringa (2002) did. On the other hand, the map of Skjekkeland is based on a restricted number of phenomena. We are not sure in how far the map accords with the perception of the speakers. Creating a new Norwegian dialect map on the basis of the arrow method, as done by Daan and Blok (1969) for the Netherlandic part of the Dutch language area, would be interesting.

In Chapter 9 we calculated distances between 360 Dutch variants with the same Levenshtein variant as used for the Norwegian data. Data was taken from the *Reeks Nederlandse Dialectatlassen* (RND). On the basis of these distances cluster analysis was applied and multidimensional scaling was performed. We compared the results to the map of Daan and Blok (1969). We found similarities and differences. Larger groups in our results were divided into smaller groups in the map of Daan, and a larger group in the map of Daan was divided into smaller groups in our results. This suggests that not all borders are equally significant on the map of Daan. When analyzing our results we also examined dendrograms. The benefit of a dendrogram is that groups and borders can be found at any level of significance.

We found it to be a big disadvantage that the RND transcriptions are made by different transcribers. Although we normalized transcriber differences to some extent, we did not succeed eliminating them all, as appeared in our results. We would like to calculate distances between Dutch varieties again on the basis of data from the *Fonologische Atlas Nederlandse Dialecten* (FAND) (Goossens et al., 1998, 2000). For the compilation of this atlas also different transcribers were involved. Nonetheless, the transcriptions are known to be of excellent quality.

Besides examining the relations between varieties, we also compared the varieties with respect to Standard Dutch. We found results which accord rather well with linguistic reality and with general opinion. The Frisian varieties appeared to be most distant.

From validation work on the one hand, and results from application to Norwegian and Dutch varieties on the other hand, the Levenshtein distance appeared to be a useful tool for finding dialect distances. Differences between our results and existing maps may be explained mostly by shortcomings in our data or in the traditional maps.

## 10.2 Applications

In this thesis we applied Levenshtein distance to the Norwegian NOS data and to the Dutch RND data. In Bolognesi and Heeringa (2002) Levenshtein distance is applied to a set of 54 Sardinian varieties, Latin, Italian, Genoese, Spanish, Catalan and Dutch. Latin was included since all Romance languages originated from Latin. Italian, Genoese, Spanish and Catalan were included since this languages influenced Sardinian in the past. Dutch was included to show the relative closeness of the Romance languages compared to the Germanic Dutch language. On the basis of Levenshtein distances the varieties were classified. The classification of the Sardinian varieties accorded with dialectological opinion. Since the Sardinian varieties are known to be relatively archaic with respect to the other Romance varieties, they were expected to be very close to Latin. It appeared that Italian was most close to Latin, followed by two Sardinian dialects, Spanish, 39 Sardinian varieties, Catalan, 13 Sardinian varieties, Genoese and – obviously at the end – Dutch. The authors found none of the Sardinian varieties to be obviously more conservative than any of the other Romance varieties in the investigation.

As mentioned above it would be interesting to study the Dutch language area again on the basis the FAND data. However it is a pity that many dialect atlasses or data sets are bounded by political borders, and not by linguistic borders. Inspired by the traditional map in Niebaum and Macha (1999, p. 193), we would like to create a new dialect map of the continental West Germanic language area, including the Netherlands, Flanders, Luxemburg, Germany, Switzerland, Liechtenstein and Austria. Similar to this, it should be interesting to investigate the whole Scandinavian language area, including Iceland, Faroe Islands, Norway, Sweden, Denmark and the Swedish speaking part of Finland. However we would like to enlarge the bounds even more, creating a map of Europe, where each of the five continua as shown in the map in Chambers and Trudgill (1998, p. 6) and the English continuum are represented in sufficient detail.[2] Possibly the *Atlas Linguarum Europae* (ALE) may be suitable for this purpose (Weijnen et al., 1973–1997).

Besides synchronic measurements, Levenshtein is also useful for diachronic research. In Heeringa and Nerbonne (2000) distances are calculated between 41 Dutch varieties on the basis of old and new transcriptions. The old transcriptions were based on translations of the parable 'The prodigal son' which were compiled by Winkler (1874). The new transcriptions were based on translations of the same text which were compiled by Harrie Scholtmeijer in 1996. Heeringa and Nerbonne used 41 varieties which appeared in both the set of Winkler and in the set of Scholtmeijer. The old and the new varieties were classified. On the basis of the 1874 data a rather sharp division in Frisian, Low Saxon and Low Franconian varieties was found, but for 1996 varieties a division in Frisian, Western Dutch and Eastern Dutch varieties was found. Further an old and a new version of Standard Dutch was added. The old varieties were compared to old Standard Dutch, the new ones to new Standard Dutch. It appeared that the majority of dialects converged to Standard Dutch. Only the dialects along the South-West coast line and in the Middle-East diverged somewhat from Standard Dutch.

In Heeringa et al. (2000) a study about Dutch-German Contact in and around Bentheim is presented. Although the RND mainly contains varieties in the Netherlands and North Belgium, 8 varieties in the German county of Bentheim were also included (see the map in Figure 9.15). The recordings of the varieties in and around Bentheim are made in 1974–1975. In 1999 Heeringa et al. made new recordings of the same Bentheim varieties and 9 varieties at the Dutch side around Bentheim. Standard Dutch and Standard German were added. There were minor differences between the older and the newer version of Standard Dutch, but the older and newer version of Standard German were the same. Just as for the data source mentioned above, the older and newer varieties were classified. The classification results showed that some dialects in the German part, which could be regarded as Dutch Low Saxon dialects in 1974–1975, were found to be German dialects in 1999. On the other hand, Dutch dialects which were grouped among German Low Saxon dialects in 1974–1975, were found to be grouped among the other Dutch dialects in 1999. All Dutch dialects shifted towards Standard Dutch while all Low German dialects shifted towards Standard German. From the results it was concluded that the political border nowadays has got a significant influence on the graduality of the dialect continuum, acting as a separator between Dutch and German dialects.

[2] It is striking that the English continuum including England, Ireland and Scotland is not shown on the map of Chambers and Trudgill (1998, p. 6), although the caption of the figure is: 'European dialect continua'.

# Appendix A

# Figures

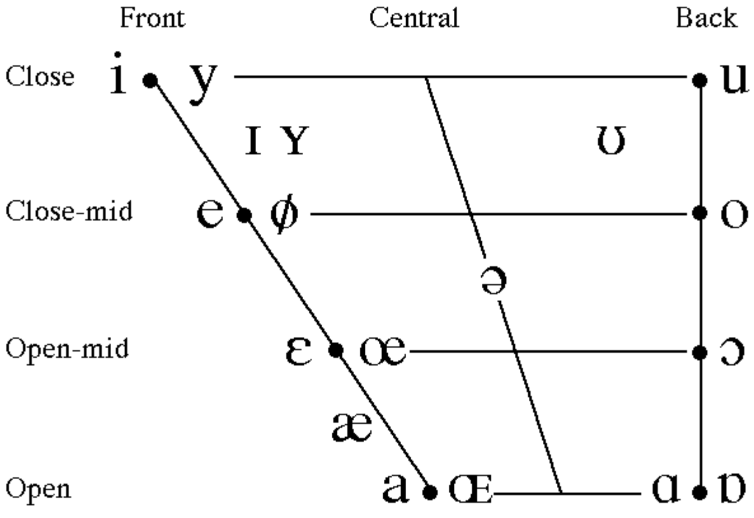


Figure A.1: The RND vowels located in the IPA vowel quadrilateral. Where symbols appear in pairs, the one to the right represents a rounded vowel.

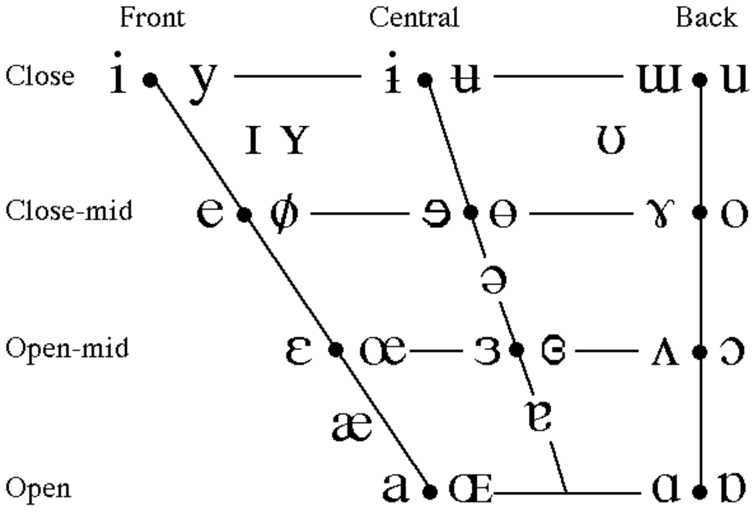
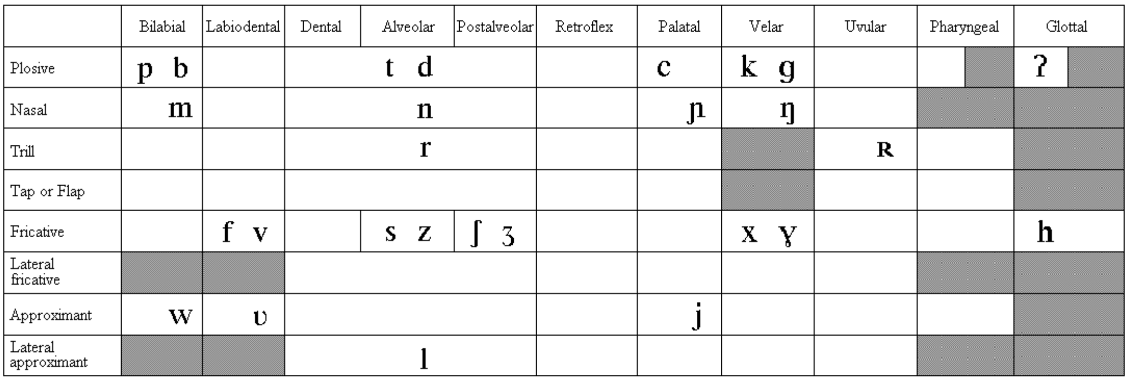
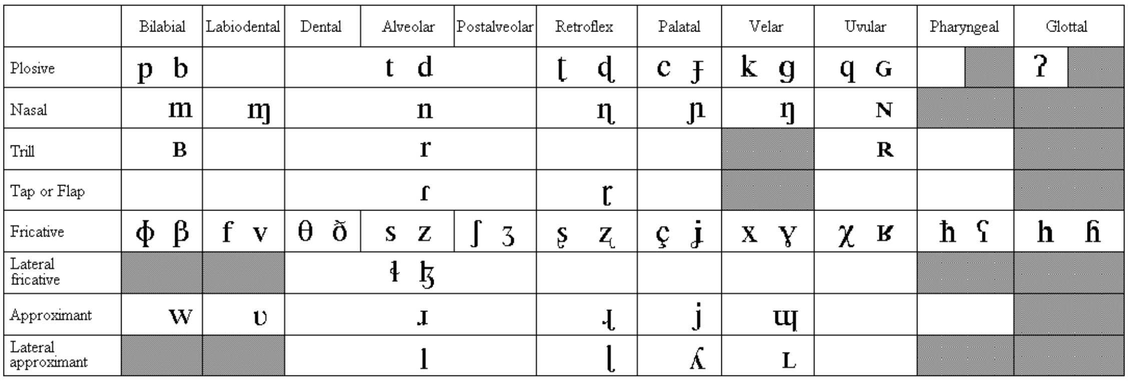


Figure A.2: The IPA vowel quadrilateral. Where symbols appear in pairs, the one to the right represents a rounded vowel. We interpret the [æ] and [ɐ] as not rounded, the [ə] as half rounded and the [ʊ] as rounded.



| | Bilabial | Labiodental | Dental | Alveolar | Postalveolar | Retroflex | Palatal | Velar | Uvular | Pharyngeal | Glottal |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Plosive | p b | | | t d | | | c | k ɡ | | ▨ | ʔ ▨ |
| Nasal | m | | | n | | | ɲ | ŋ | | ▨ | ▨ |
| Trill | | | | r | | | | ▨ | ʀ | | ▨ |
| Tap or Flap | | | | | | | | ▨ | | | ▨ |
| Fricative | | f v | | s z | ʃ ʒ | | | x ɣ | | | h |
| Lateral fricative | ▨ | ▨ | | | | | | | | ▨ | ▨ |
| Approximant | w | ʋ | | | | | j | | | | ▨ |
| Lateral approximant | ▨ | ▨ | | l | | | | | | ▨ | ▨ |

Figure A.3: The RND consonants in the IPA consonant table. Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.



| | 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 A.4: The IPA consonant table. Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

# Appendix B

# Tables

| | Norwegian | English | NOS |
|---|---|---|---|
| 1 | nordavinden | the northwind | 1, 4, 4, 6 |
| 2 | og | and | 1, 4, 5, 5, 6 |
| 3 | sola | the sum | 1, 5, 6 |
| 4 | kjekla / kjeklet | quarrel | 1 |
| 5 | om | about | 1, 3 |
| 6 | kven / hvem | who | 1 |
| 7 | av | of | 1, 3, 6 |
| 8 | dei / dem | them | 1, 3, 6 |
| 9 | som | who | 1, 3 |
| 10 | var | was | 1, 6 |
| 11 | den | the | 1, 3, 6 |
| 12 | sterkaste / sterkeste | strongest | 1, 3, 6 |
| 13 | da | then | 2, 5 |
| 14 | kom | came | 2 |
| 15 | det | there | 2 |
| 16 | en | a | 2, 2 |
| 17 | mann | man | 2 |
| 18 | gaaande / gaaende | going | 2 |
| 19 | med | with | 2 |
| 20 | varm | warm | 2 |
| 21 | frakk | coat | 2 |
| 22 | pa | around | 2 |
| 23 | seg | himself | 2, 3, 4, 5 |
| 24 | dei / de | they | 3 |
| 25 | vart / blei | were | 3 |
| 26 | samde / enige | agreed | 3 |
| 27 | at | that | 3, 6 |
| 28 | han / den | he | 3 |
| 29 | foerst | first | 3 |
| 30 | kunne | could | 3 |
| 31 | faa | get | 3 |
| 32 | mannen | the man | 3, 4, 5 |
| 33 | til aa | till | 3 |
| 34 | ta | take | 3 |
| 35 | av | off | 3, 5 |
| 36 | frakken | the coat | 3, 4, 5 |
| 37 | skulle gjelde for | would apply for | 3 |
| 38 | saa | so | 4, 5, 6 |
| 39 | blaaste | blew | 4, 4 |
| 40 | av all si makt | with might and main | 4 |

Table B.1: List of 58 words which appeared in the text ‘Nordavinden og sola’. The third column gives the sentence number(s) in which the word can be found. All instances of one word are used.

|  | **Norwegian** | **English** | **NOS** |
|---|---|---|---|
| 41 | men | but | 4 |
| 42 | dess | the | 4, 4 |
| 43 | meir / mer | more | 4 |
| 44 | han | he | 4 |
| 45 | tettare / tettere | tighter | 4 |
| 46 | trakk | draw | 4 |
| 47 | rundt | around | 4 |
| 48 | til sist | finally | 4 |
| 49 | gav | gave | 4 |
| 50 | opp | up | 4 |
| 51 | skein / skinte | shone | 5 |
| 52 | fram | to the front | 5 |
| 53 | godt | good | 5 |
| 54 | varmt | warm | 5 |
| 55 | straks | at once | 5 |
| 56 | tok | took | 5 |
| 57 | maatte | must | 6 |
| 58 | innroemme | admit | 6 |

Table B.2: Table B.1 continued.

| | Dutch | English | RND |
|---|---|---|---|
| 1 | mijn | my | 2 |
| 2 | vriend | friend | 2 |
| 3 | werk | work | 4 |
| 4 | op | on | 5 |
| 5 | schip | ship | 5 |
| 6 | kregen | got | 5 |
| 7 | brood | bread | 5 |
| 8 | vinger | finger | 6 |
| 9 | vier | four | 10 |
| 10 | bier | beer | 10 |
| 11 | twee | two | 11 |
| 12 | drie | three | 12 |
| 13 | hij | he | 13 |
| 14 | knuppel | cudgel | 13 |
| 15 | ik | I | 14 |
| 16 | knie | knee | 14 |
| 17 | gezien | seen | 14 |
| 18 | kerel | fellow | 21 |
| 19 | stenen | stones | 25 |
| 20 | breder | broader | 25 |
| 21 | duivel | devil | 28 |
| 22 | gebleven | stayed | 28 |
| 23 | meester | master | 29 |
| 24 | zee | sea | 29 |
| 25 | graag | gladly | 31 |
| 26 | steel | handle | 33 |
| 27 | bezem | broom | 33 |
| 28 | geroepen | called | 35 |
| 29 | peer | pear | 36 |
| 30 | rijp | ripe | 36 |
| 31 | geld | money | 38 |
| 32 | ver | far | 39 |
| 33 | brengen | bring | 39 |
| 34 | zwemmen | swim | 42 |
| 35 | bed | bed | 45 |
| 36 | springen | spring | 47 |
| 37 | vader | father | 53 |
| 38 | zes | six | 53 |
| 39 | jaar | year | 53 |
| 40 | school | school | 53 |
| 41 | laten | let | 53 |

Table B.3: List of 125 words selected from the 141 RND sentences. The third column gives the sentence number from which the word usually was taken.

| | Dutch | English | RND |
|---|---|---|---|
| 42 | gaan | go | 53 |
| 43 | potten | jars | 56 |
| 44 | zijn | are | 56 |
| 45 | veel | much | 56 |
| 46 | maart | March | 58 |
| 47 | nog | yet | 58 |
| 48 | koud | cold | 58 |
| 49 | kaars | candle | 59 |
| 50 | geeft | gives | 59 |
| 51 | licht | light | 59 |
| 52 | paard | horse | 60 |
| 53 | tegen | against | 63 |
| 54 | kaas | cheese | 66 |
| 55 | dag | day | 68 |
| 56 | avond | evening | 68 |
| 57 | barst | crack | 70 |
| 58 | brief | letter | 71 |
| 59 | hart | heart | 72 |
| 60 | spannen | put | 74 |
| 61 | nieuwe | new | 74 |
| 62 | kar | cart | 74 |
| 63 | zoon | son | 76 |
| 64 | koning | king | 76 |
| 65 | ook | also | 76 |
| 66 | geweest | been | 76 |
| 67 | lange | long | 78 |
| 68 | woord | word | 79 |
| 69 | kindje | baby | 80 |
| 70 | was | was | 80 |
| 71 | dochtertje | little daughter | 82 |
| 72 | bos | wood | 82 |
| 73 | ladder | ladder | 83 |
| 74 | mond | mouth | 86 |
| 75 | droog | dry | 86 |
| 76 | dorst | thirst | 86 |
| 77 | weg | way | 87 |
| 78 | krom | curved | 87 |
| 79 | liedje | ditty | 90 |
| 80 | goed | good | 92 |
| 81 | kelder | cellar | 95 |
| 82 | voor | for | 95 |
| 83 | moest | must | 96 |

Table B.4: Table B.3 continued.

| | **Dutch** | **English** | **RND** |
|---|---|---|---|
| 84 | drinken | drink | 96 |
| 85 | broer | brother | 98 |
| 86 | moe | tired | 98 |
| 87 | dun | thin | 100 |
| 88 | zuur | sour | 100 |
| 89 | put | well | 101 |
| 90 | uur | hour | 101 |
| 91 | vuur | fire | 104 |
| 92 | duwen | push | 105 |
| 93 | hebben | have | 106 |
| 94 | stuk | piece | 106 |
| 95 | brug | bridge | 106 |
| 96 | veulen | foal | 107 |
| 97 | komen | come | 107 |
| 98 | deur | door | 109 |
| 99 | gras | grass | 111 |
| 100 | bakken | bake | 113 |
| 101 | je | you | 116 |
| 102 | eieren | eggs | 116 |
| 103 | krijgen | get | 116 |
| 104 | waren | were | 119 |
| 105 | vijf | five | 119 |
| 106 | hooi | hay | 122 |
| 107 | is | is | 122 |
| 108 | groen | green | 122 |
| 109 | boompje | little tree | 124 |
| 110 | wijn | wine | 125 |
| 111 | huis | house | 126 |
| 112 | melk | milk | 127 |
| 113 | spuit | spouts | 127 |
| 114 | koe | cow | 127 |
| 115 | koster | sexton | 128 |
| 116 | buigen | bend | 129 |
| 117 | blauw | blue | 131 |
| 118 | geslagen | struck | 131 |
| 119 | saus | sauce | 132 |
| 120 | flauw | flat | 132 |
| 121 | sneeuw | snow | 133 |
| 122 | doen | do | 136 |
| 123 | dopen | baptize | 137 |
| 124 | dorsen | thresh | 138 |
| 125 | binden | bind | 139 |

Table B.5: Table B.3 continued.

# Samenvatting

## Inleiding

Volgens het bewustzijn van dialectsprekers bestaan er dialectgrenzen in het dialectlandschap. Dit blijkt uit de dialectkaart van Daan, waarop grenzen zijn getekend op basis van het dialectbewustzijn van de sprekers. Het dialectlandschap kan echter ook als een continuüm beschouwd worden. Wanneer we langs een rechte lijn reizen van dorp naar dorp, bemerken we slechts geleidelijke veranderingen. Om dialectgrenzen en dialectcontinua te verkennen op elke niveau van gedetailleerdheid, hebben we een 'liniaal' nodig waarmee de taalkundige afstand voor een willekeurig dialectpaar op een objectieve manier gevonden kan worden.

De eerste die een methode ontwikkelde voor het meten van dialectafstanden was Jean Séguy. Hij berekende de afstand tussen twee dialecten als het aantal keren dat de twee dialecten voor een bepaald item verschilden. Het aantal verschillende items werd uitgedrukt in een percentage. Een vergelijkbare aanpak werd ook toegepast door Hans Goebl.

De gebroeders Hoppenbrouwers introduceerden in 1988 twee frequentie-gebaseerde methoden waarmee dialectafstanden gevonden kunnen worden op basis van fonetische teksten. Bij de eerste methode worden per tekst de frequenties van de klanken bepaald en die frequenties worden gedeeld door het totale aantal klanken in de tekst. De afstand tussen twee variëteiten is gelijk aan de som van de frequentieverschillen. Bij de tweede methode worden frequenties van features (kenmerken van klanken) bepaald. De afstand tussen twee variëteiten is in het eenvoudigste geval opnieuw gelijk aan de som van de frequentieverschillen. Beide methoden duiden we aan als varianten van de *corpus-frequentie-methode*.

De beide frequentie-gebaseerde methoden onderscheiden geen woorden in de tekst. Dit kan opgelost worden door foon- of featurefrequenties per woord te bepalen. De afstand tussen twee woorduitspraken, corresponderend met twee dialecten, is opnieuw gelijk aan de som van de frequentieverschillen. De afstand tussen twee dialecten is gelijk aan de som van de woordafstanden. We noemen deze aanpak de *frequentie-per-woord-methode*.

In 1995 gebruikte Kessler de *Levenshtein afstand* voor het bepalen van taalkundige afstanden tussen dialecten. Met deze afstandsmaat wordt de afstand tussen twee woorduitspraken bepaald door de kosten te bepalen van de minimaal vereiste verzameling van toevoegingen, verwijderingen en vervangingen die nodig is om de ene uitspraak te veranderen in de andere. Kessler paste de afstandsmaat toe op Ierse dialecten. Dit bleek succesvol. Wij gebruikten deze afstandsmaat eveneens omdat de methode objectief is, graduele woordafstanden berekent, woorden als taalkundige eenheden verwerkt, en de volgorde van klanken in een woord in beschouwing neemt. De Levenshtein-afstand staat centraal in dit proefschrift.

## Het meten van segmentafstanden

Als we taalvariëteiten op basis van woordtranscripties willen vergelijken, moeten vooraf de afstanden tussen de segmenten bekend zijn. Deze afstanden zijn afhankelijk van de manier waarop spraaksegmenten zijn gerepresenteerd. We onderzochten de foonrepresentatie, de featurerepresentatie en de akoestische representatie.

In het eenvoudigste geval is een spraaksegment of foon niet verder gedefinieerd: twee fonen zijn gelijk of verschillend. Nadeel van de foonrepresentatie is dat bijvoorbeeld de afstand tussen de [ɪ] en de [e] even groot is als de afstand tussen de [ɪ] en de [ɒ]. Dit probleem wordt opgelost door klanken te representeren door een reeks van onderscheidende kenmerken oftwel features. Featurewaarden representeren de mate waarin een feature geldig is. Bijvoorbeeld een feature *lang* is 0 voor een korte klank, 0.5 voor een halflange klank en 1 voor een lange klank.

We experimenteerden met drie featuresystemen. Het eerste werd in 1988 ontwikkeld door de gebroeders Hoppenbrouwers (H & H). Het betreft een articulatie-gebaseerd systeem dat de auteurs gebruikten voor het vergelijken van dialecten in het Nederlandse dialectgebied. Het tweede systeem is gebaseerd op twee andere systemen. Het ene systeem werd ontwikkeld door Vieregge in 1984. Vieregge ontwikkelde zijn systeem voor de controle van de kwaliteit van transcripties. Dit systeem is gedeeltelijk gebaseerd op metingen van perceptieve klankafstanden. Het andere systeem werd ontwikkeld door Cucchiarini 1993. Het systeem van Cucchiarini is een aangepaste versie van het systeem van Vieregge. We definieerden de klinkers in de lijn van Vieregge, en de medeklinkers in de lijn van Cucchiarini. Het derde systeem in ons onderzoek is ontwikkeld door Almeida en Braun in 1986 (A & B). Evenals het tweede systeem is ook dit systeem bedoeld voor de controle van de kwaliteit van transcripties. In het systeem worden op een heel directe manier de afstanden afgeleid uit het IPA-systeem.

Featuresystemen zijn vaak niet gebaseerd op fysische metingen. Alleen het systeem van V & C is gedeeltelijk gebaseerd op afstanden die gemeten werden in een perceptieëxperiment. We hebben daarom ook klankafstanden gemeten op basis van akoestische representaties van samples van de IPA klanken. We gebruikten samples van de geluidsband *The Sounds of the International Phonetic Alphabet* waarop alle IPA klanken uitgesproken worden door twee sprekers.

De klinkers werden geïsoleerd uitgesproken, en de medeklinkers knipten we uit de context waarin ze werden uitgesproken.

We experimenteerden met twee spectrogram-gebaseerde representaties en met een representatie door formantsporen. Een spectrogram is een grafiek waarin de frequentie gerepresenteerd wordt door de x-as en de tijd door de y-as, en waarin de grijswaarde voor ieder punt in de grafiek de intensiteit representeert. We gebruikten niet de standaard-spectrogrammen, maar meer perceptief gemotiveerde modellen: het Barkfilter en het cochleagram. Essentieel voor de waarneming van klinkers is dat spectrale pieken door het oor worden herkend. Hetzelfde geldt voor sonorante medeklinkers. Deze pieken heten *formanten*, en een reeks van formanten in het verloop van de tijd heet een *formantspoor*. We experimenteerden ook met de formantsporenrepresentatie.

Zowel op basis van feature-representaties als op basis van akoestische representaties berekenden we de segmentafstanden. Omdat in onze perceptie kleine klankverschillen soms een relatief sterke rol spelen ten opzichte van grote klankverschillen, experimenteerden we ook met een aanpak waarbij de logaritmen van de klankafstanden gebruikt werden. Omdat de logaritme van 1 gelijk is aan 0, berekenden we die als: $ln(\text{afstand} + 1)$.

# Het meten van dialectafstanden

Wanneer de afstanden tussen spraaksegmenten vastgesteld zijn, kunnen we de afstanden tussen woorduitspraken bepalen en vervolgens de afstanden tussen taalvariëteiten. We bepaalden de afstand tussen een woorduitspraak uit de ene variëteit en de corresponderende woorduitspraak uit de andere variëteit met de Levenshtein-afstand. Dit algoritme bepaalt hoe zo eenvoudig mogelijk het ene woord kan worden veranderd in het andere woord door klanken toe te voegen, te verwijderen of te vervangen. Aan de operaties worden gewichten toegekend. In de eenvoudigste vorm van het algoritme hebben alle operaties hetzelfde gewicht, bijvoorbeeld 1. We illustreren het gebruik van de gewichten met een voorbeeld. Het woord *konijn* wordt uitgesproken als [kənɛːn] in het dialect van Amsterdam, en als [kniːnə] in het dialect van Zwollekerspel.[1] Het veranderen van de ene variant in de andere gaat als volgt:

| | | |
|---|---|---|
| kənɛːn | verwijder ə | 1 |
| knɛːn | vervang ɛː door iː | 1 |
| kniːn | voeg toe ə | 1 |
| kniːnə | | |
| | | 3 |

[1] De woorduitspraken werden opgenomen en transcribeerd in 2000 door Renée van Bezooijen, Katholieke Universiteit Nijmegen.

Voor de bepaling van deze afstand met het Levenshtein-algoritme worden beide woorden onder elkaar gezet, waarbij een keuze gemaakt wordt welke segmenten uit de ene variant corresponderen met welke segmenten uit de andere variant. Met andere woorden: de varianten worden *opgelijnd*. De kracht van het Levenshtein-algoritme is nu dat dit algoritme de woordafstand altijd berekent op basis van de oplijning waarin klankcorrespondenties *zodanig* zijn gekozen, dat de som van de operaties minimaal is. In ons voorbeeld ziet de oplijning er als volgt uit:

| | | | | | |
|---|---|---|---|---|---|
| k | ə | n | ɛː | n | |
| k | | n | iː | n | ə |
| 0 | 1 | 0 | 1 | 0 | 1 |

Wanneer woorduitspraken op deze manier met elkaar vergeleken worden, zal de afstand tussen langere woorden gemiddeld genomen groter zijn dan de afstand tussen kortere woorden. Hoe langer een woord is, hoe groter de kans dat er verschillen zijn ten opzichte van het corresponderende woord in een andere taalvariëteit. Omdat dit niet overeenstemt met het idee dat een woord een taalkundige eenheid is, ongeacht het aantal segmenten waaruit het bestaat, wordt de Levenshtein-afstand gedeeld door de lengte van de oplijning (de gecombineerde woordlengte). In ons voorbeeld is deze gelijk aan 6. De woordafstand, genormaliseerd over de lengte, is nu gelijk aan $3/6 = 0.5$.

Bij gebruik van de foonrepresentatie zijn de gewichten van de operaties gelijk aan 1. Gebruiken we echter een featurerepresentatie of een akoestische representatie, dan zullen de gewichten gradueel variëren.

Op basis van de gemiddelde Levenshtein-afstanden tussen variëteiten kunnen de variëteiten geclassificeerd worden. We maakten gebruik van *cluster-analyse* en *multidimensionale schaling*, twee technieken die elkaar aanvullen. Het resultaat van cluster-analyse is een dendrogram, een boom waarin de variëteiten de bladeren zijn. Het resultaat van multidimensionale schaling is een plot waarop sterk verwante verwante variëteiten dicht bij elkaar zijn geplaatst, en sterk verschillende variëteiten juist ver uit elkaar. We schaalden zowel naar twee als naar drie dimensies. In de plot worden de eerste en tweede dimensie gerepresenteerd door respectievelijk de x-as en de y-as, en de derde dimensie door de grijswaarde van de stippen.

# Validatie

In een validatie-onderzoek vergeleken we de corpus-frequentie-methode, de frequentie-per-woord-methode en de Levenshtein-afstand met elkaar. Voor elk van de drie methoden testten we de verschillende segmentrepresentaties: de foonrepresentatie en de featurerepresentatie. Voor de Levenshtein-afstand testten we ook de akoestische segment-representaties. Verder werden voor de Levenshtein-afstand zowel lineaire als logaritmische segment-afstanden in beschouwing genomen.

Het validatie-onderzoek voerden we uit op basis van 15 Noorse dialecten. Digitale opnamen en transcripties werden gemaakt door Jørn Almberg. De opnamen bestaan uit vertalingen van de fabel 'De noordenwind en de zon'. De tekst bestond (gewoonlijk) uit 58 woorden.[2] Op basis van de opnamen voerde Charlotte Gooskens een perceptie-experiment uit in de lente van 2000. In elk van de 15 plaatsen beluisterde een groep leerlingen op de middelbare school een band met daarop de opnames van alle 15 dialecten. Voor elk van de dialecten moesten de leerlingen op een schaal van 1 tot en met 10 de mate van verwantschap met hun eigen dialect geven, waarbij 1=gelijk aan eigen dialect en 10=ongelijk aan eigen dialect. De gemiddelde scores van de leerlingen in een plaats geven de afstanden van de 15 dialecten op de band ten opzichte van het dialect in die plaats. Omdat het experiment in elk van de 15 plaatsen werd uitgevoerd, kregen we een afstandenmatrix van 15 × 15 afstanden. We correleerden de resultaten van onze methoden (of varianten daarvan) met deze perceptieve afstanden. Hoe hoger de correlatie, hoe beter de methode de perceptie benadert.

De methoden op basis van de foonrepresentatie bleken het sterkste te correleren met de perceptieve afstanden, direct gevolgd door de Levenshtein-afstanden op basis van de akoestische segmentrepresentatie. De methoden op basis van featuresystemen waren beduidend slechter. Bekijken we resultaten per representatie, dan zien we zowel bij de foonrepresentatie als bij de featurerepresentatie dat de frequentie-per-woord-methode even goed is als, of beter is dan de corpus-frequentie-methode, de Levenshtein-afstand altijd beter is dan de frequentie-per-woord-methode, en de Levenshein-afstand op basis van logaritmische segmentafstanden even goed is als, of beter is dan de Levenshtein-afstand op basis van lineaire segmentafstanden. Dit is ook wat we op methodologische gronden verwachtten. Het feit dat de foonrepresentatie erg goed werkt (voor alle drie methoden), en dat logaritmische segmentafstanden vaak betere resultaten geven dan lineaire segmentafstanden, lijkt erop te wijzen dat het in de perceptie vooral belangrijk is *dat* twee segmenten verschillend zijn, en dat *de mate* waarin ze van elkaar verschillen veel minder belangrijk is. Van de drie akoestische segmentrepresenaties blijkt het Barkfilter beter te zijn dan de twee andere representaties wanneer lineaire segmentafstanden worden gebruikt. Bij gebruik van logaritmische afstanden is er geen verschil.

# Resulaten

Hoewel de Levenshtein-afstand op basis van de foonrepresentatie iets sterker correleerde dan de Levenshtein-afstand op basis van de logaritmische akoes-

[2] De opnamen en transcripties zijn gratis beschikbaar via `http://www.ling.hf.ntnu.no`.

tische segmentafstanden, genereerden we de resultaten toch met de variant van de Levenshtein-afstand die gebruik maakt van logaritmische akoestische segmentafstanden. Voor een kleine gegevensverzameling van 15 dialecten werkt de foonrepresentatie-gebaseerde aanpak weliswaar goed, maar bij gebruik van een dichter net van plaatsen kunnen kleinere verschillen een sterkere rol spelen. Met de akoestische maat worden die verschillen in sterkere mate verwerkt. We pasten de afstandsmaat toe op dialecten in het Noorse en het Nederlandse dialectgebied.

De 15 Noorse dialecten van het perceptie-onderzoek maken deel uit van een grotere gegevensverzameling. We berekenden afstanden tussen 55 Noorse dialecten. Afgezien van enkele taaleilanden kregen we op basis van cluster-analyse een hoofdindeling bestaande uit zes groepen: noord, centraal, west, oost, zuidwest en zuidoost. Op basis van multidimensionale schaling kregen we een indeling bestaande uit ruwweg 5 groepen: noord, centraal, west, oost, zuid. De laatste indeling komt iets beter overeen met de traditionele indeling van Skjekkeland. Verschillen kunnen verklaard worden door beperkingen van onze woordenlijst enerzijds, en de keuze van de isoglossen door Skjekkeland anderzijds.

De *Reeks Nederlandse Dialectatlassen* werd samengesteld in de periode 1925–1982 door E. Blancquaert and W. Pée. Van de 1956 beschikbare dialecttranscripties kozen we er 360. We berekenden de afstanden tussen de dialecten op basis van 125 woorden. Met cluster-analyse kregen we een indeling in Fries, Nedersaksisch, Nederfrankisch en Limburgs. We onderzochten elk van de vier groepen meer gedetailleerd en vergeleken de indeling met de kaart van Daan. Verschillen konden soms verklaard worden uit notatieverschillen van de verschillende transcribenten in de RND, en een enkele keer uit een (vermoedelijke) tekortkoming van de kaart van Daan. Op basis van multidimensionale schaling kregen we de klassieke indeling in Fries, Nedersaksisch en Nederfrankisch. We vergeleken de dialecten ook ten opzichte van het Standaard Nederlands. Het dialect van Haarlem bleek het sterkst verwant, en de Friese variëteiten bleken het meest afwijkend.

## Conclusie

In dit proefschrift ontwikkelden we verschillende varianten van de Levenshtein-afstand en onderzochten of deze afstandsmaat bruikbaar is voor het berekenen van afstanden tussen taalvariëteiten. Uit validatie-onderzoek bleek dat de Levenshtein-afstand betere resultaten geeft dan de corpus-frequentie-methode en de frequentie-per-woord-methode. Ook bij toepassing van de Levenshtein-afstand op Noorse en Nederlandse gegevens bleek de methode een geschikt gereedschap voor het vinden van afstanden tussen taalvariëteiten.

# Bibliography

Almberg, J. (2001). The circumflex tone in a Norwegian dialect. In Van Dommelen, W. A. and Fretheim, T., editors, *Nordic Prosody: Proceedings of the VIIIth Conference, Trondheim 2000*, pages 9–22, Frankfurt am Main. Peter Lang.

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

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

Almeida, A. and 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):158–172.

Barbujani, G., Whitehaed, G. N., Bertorelle, G., and Nasidze, I. S. (1994). Testing hypotheses on processes of genetic and linguistic change in the Caucasus. *Human Biology: the International Journal of Population Biology and Genetics*, 66(5):843–864.

Blancquaert, E. (1939). *Tekstboekje.* De Sikkel, Antwerpen, 2nd edition. Nederlandse Fonoplaten van Blancquaert en van der Plaetse, Eerste Reeks.

Blancquaert, E. (1948). *Na meer dan 25 jaar Dialect-onderzoek op het Terrein.* Nr. 28; Reeks III. Koninklijke Vlaamse Academie voor Taal- en Letterkunde, Gent.

Blancquaert, E. and Peé, W., editors (1925–1982). *Reeks Nederlands(ch)e Dialectatlassen.* De Sikkel, Antwerpen.

Bloomfield, L. (1933). *Language.* Holt, Rinehart and Winston, New York.

Bolognesi, R. and Heeringa, W. (2002). De invloed van dominante talen op het lexicon en de fonologie van Sardische dialecten. *Gramma/TTT: tijdschrift voor taalwetenschap*, 9(1):45–84.

Bonnet, E. and Van de Peer, Y. (2002). zt: a software tool for simple and partial Mantel tests. *Journal of Statistical Software*, 7(10):1–12. Available via: `http://www.jstatsoft.org/`.

Booij, G. E. (1995). *The phonology of Dutch*. Oxford University Press, Oxford.

Breuker, P. (1993). *Normaspecten fan it hjoeddeiske Frysk*. PhD thesis, Rijksuniversiteit Groningen, Stichting FFYRUG, Groningen.

Bürkle (1986). Zur Validität eines Maßes zur Reliabilitätsbestimmung phonetisch-segmenteller Transkriptionen. *Zeitschrift für Dialektologie und Linguistik*, LIII(2):173–181.

Chambers, J. K. and Trudgill, P. (1998). *Dialectology*. Cambridge University Press, Cambridge, 2nd edition.

Chomsky, N. and Halle, M. (1968). *The Sound Pattern of English*. Harper & Row, New York.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16:297–334.

Cucchiarini, C. (1993). *Phonetic Transcription: a Methodological and Emperical Study*. PhD thesis, Katholieke Universiteit Nijmegen, Nijmegen.

Daan, J. (1956). Noordhollandse dialekten. *Taal en Tongval*, 8:113–121.

Daan, J. (1990). *Urk: het dialect van Urk*, volume 4 of *Flevo Profiel*. De Walburg Pers, Zutphen.

Daan, J. and Blok, D. P. (1969). *Van Randstad tot Landrand; toelichting bij de kaart: Dialecten en Naamkunde*, volume XXXVII of *Bijdragen en mededelingen der Dialectencommissie van de Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam*. Noord-Hollandsche Uitgevers Maatschappij, Amsterdam.

Dalbor, J. B. (1969). *Spanish Pronunciation: Theory and Practice*. Holt, Rinehart and Winston, New York, Toronto and London.

Elert, C.-C. (1964). *Phonologic Studies of Quantity in Swedish*. PhD thesis, University of Stockholm, Uppsala.

Entjes, H. (1970). *Die Mundart des Dorfes Vriezenveen*. PhD thesis, Westfälische Wilhelms-Universität Münster, Groningen.

Farris, J. S. (1969). On the cophenetic correlation coefficient. *Systematic Zool*, 18:279–285.

Fintoft, K. (1961). The duration of some Norwegian speech sounds. *Phonetica*, 7:19–39.

Foerste, W. (1960). *Einheit und Vielfalt der niederdeutschen Mundarten*, volume 4 of *Schriften zur Heimatkunde und Heimatpflege.* Aschendorffsche Verlagsbuchhandlung, Münster Westfalen.

Fowlkes, E. B. and Mallows, C. L. (1983). A method for comparing two hierarchical clusterings. *Journal of the American Statistical Association*, 78:553–569.

Goebl, H. (1982). *Dialektometrie; Prinzipien und Methoden des Einsatzes der numerischen Taxonomie im Bereich der Dialektgeographie*, volume 157 of *Philosophisch-Historische Klasse Denkschriften.* Verlag der Österreichischen Akademie der Wissenschaften, Vienna. With assistance of W.-D. Rase and H. Pudlatz.

Goebl, H. (1984). *Dialektometrische Studien. Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*, volume 191, 192, 193 of *Beihefte zur Zeitschrift für romanische Philologie.* Max Niemeyer Verlag, Tübingen. With assistance of S. Selberherr, W.-D. Rase and H. Pudlatz.

Goebl, H. (1993). Probleme und Methoden der Dialektometrie: Geolinguistik in globaler Perspektive. In Viereck, W., editor, *Proceedings of the International Congress of Dialectologists*, volume 1, pages 37–81, Stuttgart. Franz Steiner Verlag.

Goebl, H. (2002). Analyse dialectométrique des structures de profondeur de l'ALF. *Revue de linguistique romane*, 66(261-262):5–63.

Goeman, A. and Taeldeman, J. (1996). Fonologie en morfologie van de Nederlandse dialecten; een nieuwe materiaalverzameling en twee nieuwe atlasprojecten. *Taal en Tongval*, XLVIII(1).

Gooskens, C. (1997). *On the Role of Prosodic and Verbal Information in the Perception of Dutch and English Varieties.* PhD thesis, Katholieke Universiteit Nijmegen, Nijmegen.

Gooskens, C. (2002). How well can Norwegians identify their dialects? *Nordic Journal of Linguistics.* Submitted.

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

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

Goossens, J. (1977). *Inleiding tot de Nederlandse Dialectologie.* Wolters-Noordhoff, Groningen.

Goossens, J. (1997). Wat heeft de RND de dialectoloog in 1997 nog te vertellen? In Van de Wijngaard, H. H. A. and Belemans, R., editors, *Nooit verloren werk; terugblik op de Reeks Nederlandse Dialectatlassen*, volume 4 of *Het dialectenboek*, pages 65–72. Stichting Nederlandse Dialecten, Groesbeek.

Goossens, J., Taeldeman, J., and Verleyen, G. (1998). *Fonologische Atlas van de Nederlandse Dialecten (F.A.N.D.) Deel I*, volume 1 of *Bouwstenen op het gebied van de Nederlandse Naamkunde, Dialectologie en Filologie.* Koninklijke Academie voor Nederlandse Taal- en Letterkunde, Gent.

Goossens, J., Taeldeman, J., and Verleyen, G. (2000). *Fonologische Atlas van de Nederlandse Dialecten (F.A.N.D.) Deel II + III*, volume 5 of *Bouwstenen op het gebied van de Nederlandse Naamkunde, Dialectologie en Filologie.* Koninklijke Academie voor Nederlandse Taal- en Letterkunde, Gent.

Gusfield, D. (1999). *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology.* Cambridge Univesity Press, Cambridge.

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

Heeringa, W. (2002). Over de indeling van de Nederlandse streektalen. Een nieuwe methode getoetst. *Driemaandelijkse bladen voor taal en volksleven in het oosten van Nederland*, 54(1-4):111–148.

Heeringa, W. and Braun, A. (2003). The use of the Almeida-Braun system in the measurement of Dutch dialect distances. *Computers and the Humanities*, 37(3):257–271.

Heeringa, W. and Gooskens, C. (2003). Norwegian dialects examined perceptually and acoustically. *Computers and the Humanities*, 37(3):293–315.

Heeringa, W. and Nerbonne, J. (2000). Change, convergence and divergence among Dutch and Frisian. In Boersma, P., Breuker, P. H., Jansma, L. G., and Van der Vaart, J., editors, *Philologia Frisica Anno 1999. Lêzingen fan it fyftjinde Frysk filologekongres*, pages 88–109, Leeuwarden. Fryske Akademy.

Heeringa, W. and Nerbonne, J. (2001). Dialect areas and dialect continua. *Language Variation and Change*, 13:375–400.

Heeringa, W., Nerbonne, J., and Kleiweg, P. (2002). Validating dialect comparison methods. In Gaul, W. and Ritter, G., editors, *Classification, Automation, and New Media; Proceedings of the 24th Annual Conference of the Gesellschaft für Klassifikation e. V.*, pages 445–452, Berlin, Heidelberg and New York. Springer.

Heeringa, W., Nerbonne, J., Niebaum, H., Nieuweboer, R., and Kleiweg, P. (2000). Dutch-German contact in and around Bentheim. In Gilbers, D., Nerbonne, J., and Schaeken, J., editors, *Languages in Contact. Studies in Slavic and General Linguistics*, volume 28, pages 145–156. Rodopi, Amsterdam and Atlanta GA.

Heeroma, K. (1961). De Oostnederlandse langevocalensystemen. In *Structuurgeografie*, volume XXXVII of *Bijdragen en mededelingen der Dialectencommissie van de Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam*, pages 1–15. Noord-Hollandsche Uitgevers Maatschappij, Amsterdam.

Heeroma, K. (1963). De geografische indeling der Oostnederlandse volkstaal. *Taal en Tongval*, 15:175–181.

Hof, J. J. (1933). *Friesche dialectgeographie*, volume 3 of *Noord- en Zuid-Nederlandse dialectbibliotheek*. Nijhoff, 's-Gravenhage.

Hogg, R. V. and Ledolter, J. (1992). *Applied Statistics for Engineers and Physical Scientists*. Macmillan Publishing Company, New York, 2nd edition.

Hoppenbrouwers, C. and Hoppenbrouwers, G. (1988). De featurefrequentiemethode en de classificatie van Nederlandse dialecten. *TABU: Bulletin voor taalwetenschap*, 18(2):51–92.

Hoppenbrouwers, C. and 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.

Hunt, M. J., Lennig, M., and Mermelstein, P. (1999). Use of dynamic programming in a syllable-based continuous speech recognition system. In Sankoff, D. and Kruskal, J., editors, *Time Warps, String edits, and Macromolecules. The Theory and Practice of Sequence Comparison*, pages 163–187. CSLI, Stanford, 2nd edition. 1st edition appeared in 1983.

Inoue, F. (1996a). Computational dialectology (1). *Area and Culture Studies*, 52:67–102.

Inoue, F. (1996b). Computational dialectology (2). *Area and Culture Studies*, 53:1–20.

IPA (1949). *The Principles of the International Phonetic Association: being a Description of the International Phonetic Alphabet and the Manner of Using it, illustrated by Texts in 51 Languages.* International Phonetic Association, London.

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

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

Jellinghaus, H. (1892). *Die Niederländischen Volksmundarten; nach den Aufzeichnungen der Niederländer.* D. Soltau's Verlag, Norden and Leipzig.

Johnson, K. (1997). *Acoustic and Auditory Phonetics.* Blackwell Publishers, Cambridge, etc.

Johnson, S. C. (1967). Hierarchical clustering schemes. *Psychometrika*, 32:241–254.

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

King, B. (1967). Step-wise clustering procedures. *Journal of the American Statistical Association*, 69:86–101.

Kocks, G. H. (1970). *Die Dialekte von Südostdrente und anliegenden Gebieten. Eine strukturgeographische Untersuchung.* PhD thesis, Rijksuniversiteit Groningen, Groningen.

König, W. and Paul, H. J. (1991). *dtv-Atlas zur deutschen Sprache.* Deutscher Taschenbuch Verlag, München.

Krämer, J. (1995). *Delaunay Triangulation in Two and Three Dimensions.* PhD thesis, University of Tübingen, Institut für Informatik, Tübingen, Germany.

Kristoffersen, G. (2000). *The Phonology of Norwegian.* The Phonology of the World's Languages. Oxford University Press, Oxford.

Kruskal, J. B. (1964). Multidimensional Scaling by Optimizing Goodness-of-Fit to a Nonmetric Hypothesis. *Psychometrika*, 29:1–28.

Kruskal, J. B. (1999). An overview of sequence comparison. In Sankoff, D. and Kruskal, J., editors, *Time Warps, String edits, and Macromolecules. The Theory and Practice of Sequence Comparison*, pages 1–44. CSLI, Stanford, 2nd edition. 1st edition appeared in 1983.

Kruskal, J. B. and Liberman, M. (1999). The symmetric time-warping problem: from continuous to discrete. In Sankoff, D. and Kruskal, J., editors, *Time Warps, String edits, and Macromolecules. The Theory and Practice of Sequence Comparison*, pages 125–161. CSLI, Stanford, 2nd edition. 1st edition appeared in 1983.

Kruskal, J. B. and Wish, M. (1978). *Multidimensional Scaling.* Number 07–011 in Sage University Paper Series on Quantitative Applications in the Social Sciences. Sage Publications, Newbury Park.

Ladefoged, P. (1975). *A Course in Phonetics.* Harcourt Brace Jovanovich, Inc., New York, Chicago, San Francisco and Atlanta.

Lecoutere, C. P. F. and Grootaers, L. (1926). *Inleiding tot de Taalkunde en tot de Geschiedenis van het Nederlandsch.* Wolters, Leuven, Den Haag and Groningen, 3rd edition.

Levenshtein, V. I. (1966). Binary codes capable of correcting deletions, insertions, and reversals. *Cybernetics and Control Theory*, 10(8):707–710.

Manly, B. F. J. (1997). *Randomization, Bootstrap and Monte Carlo Methods in Biology.* Chapman and Hall, London, 2nd edition.

Manni, F. (2001). *Strutture genetiche e differenze linguistiche: Un approccio comparato a livello micro e macro regionale.* PhD thesis, University of Ferrara, Ferrara.

Manning, C. D. and Schütze, H. (1999). *Foundations of Statistical Natural Language Processing.* The MIT Press, Cambridge, etc.

Mantel, N. (1967). The detection of disease clustering and a generalized regression approach. *Cancer Research*, 27:209–220.

Matthews, P. H. (1997). *The Concise Oxford Dictionary of Linguistics.* Oxford Paperback Reference. Oxford University Press, Oxford, etc.

Moulton, W. G. (1960). The short vowel systems of northern Switzerland. *Word*, 16:155–182.

Moulton, W. G. (1962). The vowels of Dutch: phonetic and distributional classes. *Lingua*, 11:294–312.

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

Nerbonne, J. and Heeringa, W. (2001). Computational comparison and classification of dialects. *Dialectologia et Geolinguistica. Journal of the International Society for Dialectology and Geolinguistics*, 2001(9):69–83.

Nerbonne, J. and Kleiweg, P. (2003). Lexical distance in LAMSAS. *Computers and the Humanities*, 37(4).

Niebaum, H. and Macha, J. (1999). *Einführung in die Dialektologie des Deutschen*. Niemeyer, Tübingen.

Nooteboom, S. G. (1971). Over de lengte van korte klinkers, lange klinkers en tweeklanken van het Nederlands. *Nieuwe Taalgids*, 64:396–402.

Nooteboom, S. G. (1972). *Production and Perception of Vowel Duration, a Study of Durational Properties of Vowels in Dutch*. PhD thesis, Rijksuniversiteit Utrecht, Utrecht.

Norušis, M. J. (1997). *SPSS Professional Statistics 7.5*. SPSS Inc, Chicago.

Nunnally, J. C. (1978). *Psychometric Theory*. McGraw-Hill, New York.

Oh, M. S. and Raftery, A. (2001). Bayesian multidimensional scaling and choice of dimension. *Journal of the American Statistical Association*, 10:1–28.

Omdal, H. (1995). Attitudes toward spoken and written Norwegian. *International Journal of the Sociology of Language*, 115:85–106.

OUP (1998). *Oxford Paperback Encyclopedia*. Oxford Paperback Reference. Oxford University Press, Oxford, etc.

Pols, L. C. W. (1977). *Spectral Analysis and Identification of Dutch Vowels in Monosyllabic Words*. PhD thesis, Vrije Universiteit Amsterdam, Amsterdam.

Potter, R. K., Kopp, G. A., and Green, H. C. (1947). *Visible Speech*. The Bell Telephone Laboratories Series. Van Nostrand, New York.

Pountain, C. J. (2001). *A History of the Spanish Language through Texts*. Routledge, London and New York.

Rand, W. M. (1971). Objective criterion for evaluation of clustering methods. *Journal of the American Statistical Association*, 66:846–850.

Reker, S. (1993). *Zakwoordenboek Gronings-Nederlands Nederlands-Gronings*. Staalboek, Veendam.

Reker, S. (1997). Ter plaatse gewonnen: de RND-inleidingen gelezen. In Van de Wijngaard, H. H. A. and Belemans, R., editors, *Nooit verloren werk; terugblik op de Reeks Nederlandse Dialectatlassen*, volume 4 of *Het dialectenboek*, pages 11–51. Stichting Nederlandse Dialecten, Groesbeek.

Rensink, W. G. (1955). Dialectindeling naar opgaven van medewerkers. *Mededelingen der Centrale Commissie voor Onderzoek van het Nederlandse Volkseigen*, 7:20–23.

Rietveld, A. C. M. (1979). Judgements on the articulatory similarity of Dutch vowels. In *IFN-Proceedings*, pages 79–88. University of Nijmegen, Institute of Phonetics.

Rietveld, A. C. M. and Van Heuven, V. J. (1997). *Algemene fonetiek*. Coutinho, Bussum.

Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Transactions on Computers*, C 18:401–409.

Sassen, A. (1953). *Het Drents van Ruinen*. PhD thesis, Rijksuniversiteit Groningen, Assen.

Schroeder, M. R., Atal, B. S., and Hall, J. L. (1979). Optimizing digital speech coders by exploiting masking properties of the human ear. *Journal of the Acoustical Society of America*, 66:1647–1652.

Séguy, J. (1973). La dialectométrie dans l'Atlas linguistique de la Gascogne. *Revue de linguistique romane*, 37:1–24.

Shepard, R. N. (1962). The analysis of proximities: multidimensional scaling with an unknown distance function. *Psychometrika*, 27:125–140, 219–246.

Skjekkeland, M. (1997). *Dei norske dialektane: tradisjonelle særdrag i jamføring med skriftmåla*. Høyskoleforlaget, Kristiansand.

Sneath, P. H. A. and Sokal, R. R. (1973). *Numerical Taxonomy. The Principles and Practice of Numerical Classification*. A Series of Books in Biology. W. H. Freeman and Company, San Francisco.

Sokal, R. R. and Michener, C. D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas science bulletin*, 38:1409–1438.

Sokal, R. R. and Rohlf, F. J. (1962). The comparison of dendrograms by objective methods. *Taxon*, 11:33–40.

Stevens, K. N. (1998). *Acoustic Phonetics*, volume 30 of *Current Studies in Linguistics*. The MIT Press, Cambridge, etc.

Takane, Y., Young, F. W., and DeLeeuw, J. (1977). Nonmetric individual differences multidimensional scaling: an alternating least squares method with optimal scaling features. *Psychometrika*, 42:7–67.

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.

Ten Bosch, L. (2000). ASR, dialects, and acoustic/phonological distances. In *ICSLP2000*, Beijing.

Togerson, W. S. (1952). Multidimensional scaling. i. Theory and method. *Psychometrika*, 17:401–419.

Togerson, W. S. (1958). *Theory and Methods of Scaling.* Wiley, New York.

Trask, R. L. (1996). *A Dictionary of Phonetics and Phonology.* Routledge, London and New York.

Traunmüller (1990). Analytical expressions for the tonotopic sensory scale. *Journal of the Acoustical Society of America*, 88:97–100.

Twilhaar, J. N. (1990). *Generatieve fonologie en de studie van Oostnederlandse dialecten.* PhD thesis, Rijksuniversiteit Utrecht, Utrecht.

Van der Veen, K. F. (1986). Yndieling en relative ôfstân fan Fryske plattelânsdialekten. In *Philologia Frisica Anno 1984, Lêzingen en neipetearen fan it tsiende Frysk Filologekongres, oktober 1984*, pages 10–60, Ljouwert. Fryske Akademy.

Van der Veen, K. F. (1994). Yndielingskaarten fan Fryske plattelânsdialekten. *It Beaken*, 56:1–23.

Van Ginneken, J. (1913). *De sociologische structuur der Nederlandsche Taal*, volume I of *Handboek der Nederlandsche Taal.* Malmberg, Nijmegen.

Vieregge, W. H. (1987). Basic Aspects of Phonetic Segmental Transcription. In Almeida, A. and Braun, A., editors, *Probleme der phonetischen Transkription*, Zeitschrift für Dialektologie und Linguistik, Beihefte, pages 5–55. Franz Steiner Verlag Wiesbaden, Stuttgart.

Vieregge, W. H., Rietveld, A. C. M., and 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., editors, *Proceedings of the 10th International Congress of Phonetic Sciences*, pages 654–659, Dordrecht and Cinnaminson. Foris Publications.

Von Trimberg, H. (1970). *Der Renner.* Texte des Mittelalters. De Gruyter, Berlin. Republished by G. Ehrismann and G. Schweikle, Deutsche Neudrucke.

Weijnen, A. (1941). *De Nederlandse dialecten.* Noordhoff, Groningen and Batavia.

Weijnen, A. (1946). De grenzen tussen de oost-noord-Brabantse dialecten onderling. In *Oost-Noordbrabantsche dialectproblemen: lezingen gehouden voor de dialecten-commissie der Koninklijke Nederlandsche Akademie van Wetenschappen op 12 april 1944*, volume VIII of *Bijdragen en meededelingen der Dialectencommissie van de Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam*, pages 1–15. Noord-Hollandsche Uitgevers Maatschappij, Amsterdam.

Weijnen, A. (1958). *Nederlandse dialectkunde.* Taalkundige bijdragen van Noord en Zuid. Van Gorcum, Assen.

Weijnen, A. (1966). *Nederlandse dialectkunde.* Studia Theodisca. Van Gorcum, Assen, 2nd edition.

Weijnen, A. A., Alinei, M. L., Kruijsen, T. J. W. M., and Brozović, D. (1973–1997). *Atlas linguarum Europae.* Van Gorcum/Musumeci Éditeur/Istituto Poligrafico e Zecca dello Stato, Libreria dello Stato, Assen/Quart/Roma.

Wells, J. and House, J. (1995). *The Sounds of the International Phonetic Alphabet.* Department of Phonetics and Linguistics University College London, London. booklet with cassette or tape.

Weng, Z. and Sokal, R. R. (1995). Origins of Indo-Europeans and the spread of agriculture in Europe: comparison of lexicostatistical and genetic evidence. *Human Biology: the International Journal of Population Biology and Genetics*, 67(4):577–594.

Wilks, D. S. (1995). *Statistical Methods in the Atmospheric Sciences: an Introduction*, volume 59 of *International Geophysics Series.* Academic Press, San Diego.

Winkler, J. (1874). *Algemeen Nederduitsch en Friesch Dialecticon.* Martinus Nijhoff, 's-Gravenhage.

Wortmann, F. (1960). Zur Geschichte der ê- und ô-laute in Niederdeutschland, besonders in Westfalen. In Wortmann, F., Møller, R., Andersson-Schmitt, M., et al., editors, *Münstersche Beitrage zur niederdeutschen Philologie*, pages 1–23. Böhlau Verlag, Graz and Cologne.

Zwaardemaker, H. and Eykman, L. P. H. (1928). *Leerboek der phonetiek: inzonderheid met betrekking tot het standaard-Nederlandsch.* Bohn, Haarlem.

Zwicker, E. and Fastl, H. (1990). *Psychoacoustics and Models.* Springer Verlag, Berlin.

Zwicker, E. and Feldtkeller, R. (1967). *Das Ohr als Nachrichtenempfänger*, volume 19 of *Monographien der elektrischen Nachrichtentechnik.* Hirzel, Stuttgart, 2nd revised edition.

# Groningen dissertations in linguistics
## (GRODIL)

1. Henriëtte de Swart (1991). *Adverbs of Quantification: A Generalized Quantifier Approach.*

2. Eric Hoekstra (1991). *Licensing Conditions on Phrase Structure.*

3. Dicky Gilbers (1992). *Phonological Networks. A Theory of Segment Representation.*

4. Helen de Hoop (1992). *Case Configuration and Noun Phrase Interpretation.*

5. Gosse Bouma (1993). *Nonmonotonicity and Categorial Unification Grammar.*

6. Peter I. Blok (1993). *The Interpretation of Focus: an epistemic approach to pragmatics.*

7. Roelien Bastiaanse (1993). *Studies in Aphasia.*

8. Bert Bos (1993). *Rapid User Interface Development with the Script Language Gist.*

9. Wim Kosmeijer (1993). *Barriers and Licensing.*

10. Jan-Wouter Zwart (1993). *Dutch Syntax: A Minimalist Approach.*

11. Mark Kas (1993). *Essays on Boolean Functions and Negative Polarity.*

12. Ton van der Wouden (1994). *Negative Contexts.*

13. Joop Houtman (1994). *Coordination and Constituency: A Study in Categorial Grammar.*

14. Petra Hendriks (1995). *Comparatives and Categorial Grammar.*

15. Maarten de Wind (1995). *Inversion in French.*

16. Jelly Julia de Jong (1996). *The Case of Bound Pronouns in Peripheral Romance.*

17. Sjoukje van der Wal (1996). *Negative Polarity Items and Negation: Tandem Acquisition.*

18. Anastasia Giannakidou (1997). *The Landscape of Polarity Items.*

19. Karen Lattewitz (1997). *Adjacency in Dutch and German.*

20. Edith Kaan (1997). *Processing Subject-Object Ambiguities in Dutch.*

21. Henny Klein (1997). *Adverbs of Degree in Dutch.*

22. Leonie Bosveld-de Smet (1998). *On Mass and Plural Quantification: The case of French 'des'/'du'-NPs.*

23. Rita Landeweerd (1998). *Discourse semantics of perspective and temporal structure.*

24. Mettina Veenstra (1998). *Formalizing the Minimalist Program.*

25. Roel Jonkers (1998). *Comprehension and Production of Verbs in aphasic Speakers.*

26. Erik F. Tjong Kim Sang (1998). *Machine Learning of Phonotactics.*

27. Paulien Rijkhoek (1998). *On Degree Phrases and Result Clauses.*

28. Jan de Jong (1999). *Specific Language Impairment in Dutch: Inflectional Morphology and Argument Structure.*

29. H. Wee (1999). *Definite Focus.*

30. E.-H. Lee (2000). *Dynamic and Stative Information in Temporal Reasoning: Korean tense and aspect in discourse.*

31. Ivilin P. Stoianov (2001). *Connectionist Lexical Processing.*

32. Klarien van der Linde (2001). *Sonority substitutions.*

33. Monique Lamers (2001). *Sentence processing: using syntactic, semantic, and thematic information.*

34. Shalom Zuckerman (2001). *The Acquisition of "Optional" Movement.*

35. Rob Koeling (2001). *Dialogue-Based Disambiguation: Using Dialogue Status to Improve Speech Understanding*

36. Esther Ruigendijk(2002). *Case assignment in Agrammatism: a cross-linguistic study*

37. Tony Mullen (2002). *An Investigation into Compositional Features and Feature Merging for Maximum Entropy-Based Parse Selection.*

38. Nanette Bienfait (2002). *Grammatica-onderwijs aan allochtone jongeren.*

39. Dirk-Bart den Ouden (2002). *Phonology in Aphasia: Syllables and segments in level-specific deficits.*

40. Rienk Withaar (2002). *The Role of the Phonological Loop in Sentence Comprehension.*

41. Kim Sauter (2002). *Transfer and Access to Universal Grammar in Adult Second Language Acquisition.*

42. Laura Sabourin (2003). *Grammatical Gender and Second Language Processing: An ERP Study.*

43. Hein van Schie (2003). *Visual Semantics.*

44. Lilia Schürcks-Grozeva (2003). *Binding and Bulgarian.*

45. Stasinos Konstantopoulos (2003). *Using ILP to Learn Local Linguistic Structures.*

46. Wilbert Heeringa (2004). *Measuring Dialect Pronunciation Differences using Levenshtein Distance.*

GRODIL, secretary Department of General Linguistics
P.O. Box 716
9700 AS Groningen