Measuring Dialect Pronunciation Differences using Levenshtein Distance

Heeringa, Wilbert Jan

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 $\hat{e}$ and $\hat{o}$ 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.

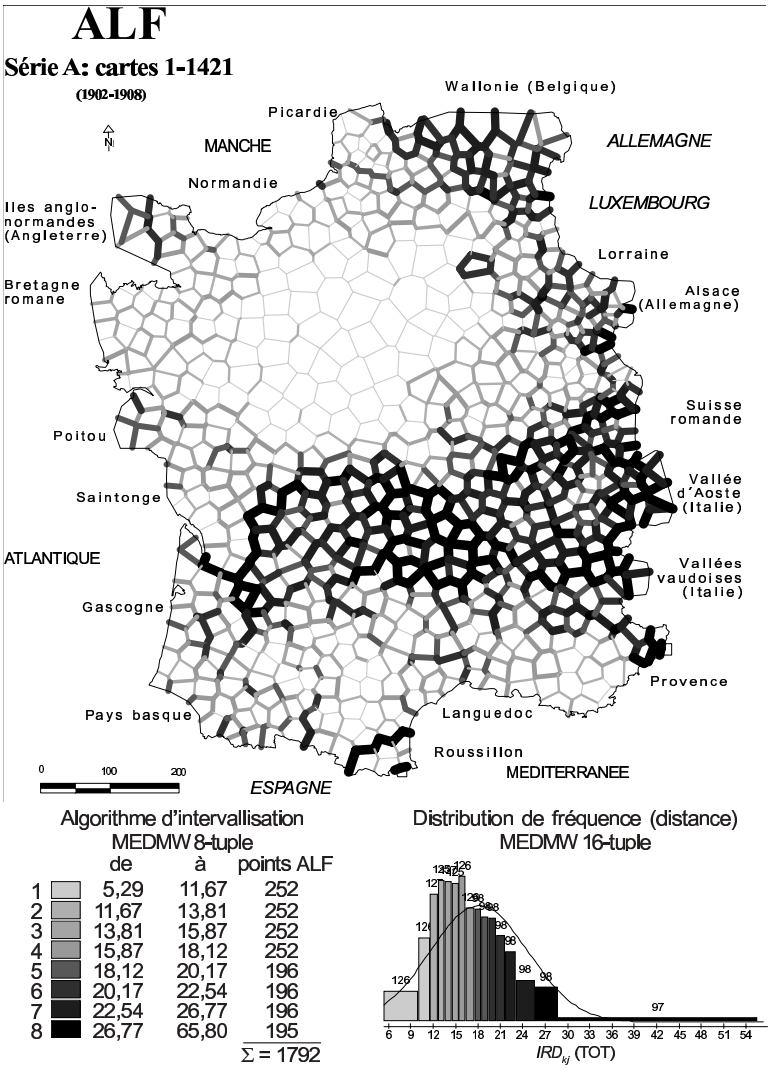

Figure 2.1: Example of an interpoint map based on 1687 working maps of the ALF and created by Goebl's research team. Darker and thicker lines separate different varieties, lighter and thinner lines separate more related ones. The 1792 distances are divided in eight classes according to the MEDMW algorithm.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

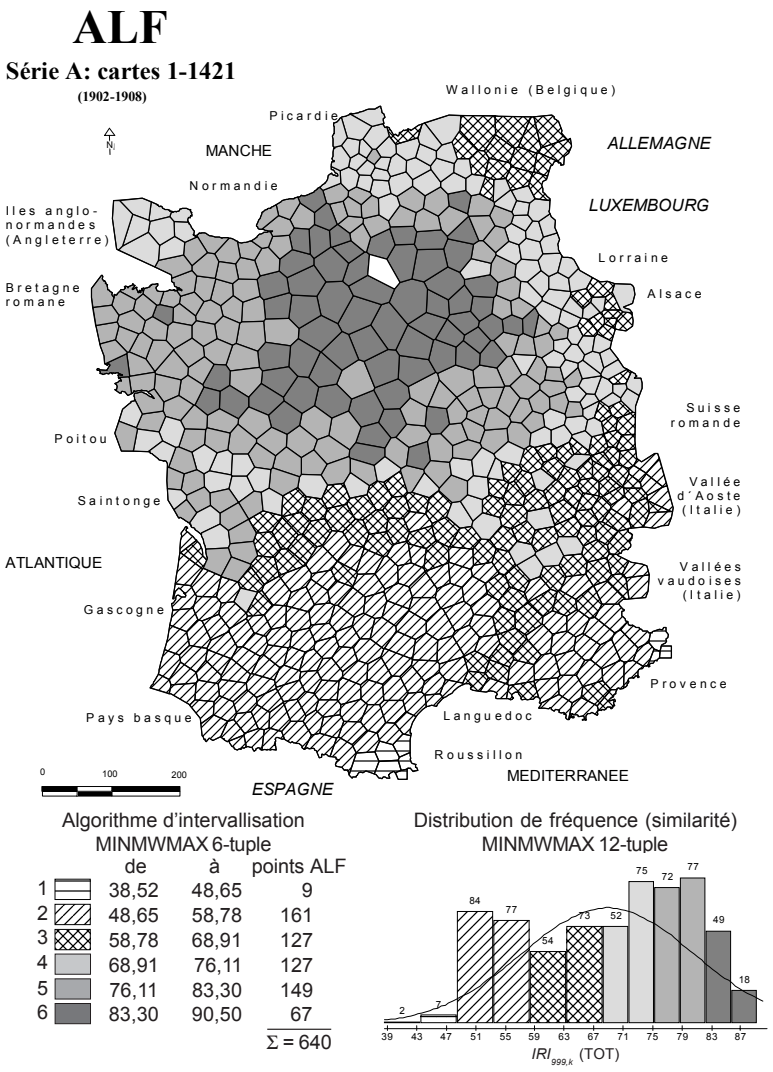

Figure 2.2: Example of an choropleth map based on 1687 working maps of the ALF and created by Goebl's research team. The map shows the distances of 640 varieties with respect to Standard French. The distances are divided in six classes according to the MINMWMAX algorithm. Each class has its own shade.

	[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 pre-

vious 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.