Measuring Dialect Pronunciation Differences using Levenshtein Distance

Heeringa, Wilbert Jan

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 [ʊin], [pɔtn̩] and [døːrə]. These pronunciations remain about the same until we arrive in Putten. In this location *deur* is pronounced as [dø˞ᵊr]. 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 [wɑ̟ˈn] in the variety of Driebergen. The monophthong [i] is replaced by the diphthong [ɑ̟ˈ]. In the following locations of Vianen and Hardinxveld the strongly related diphthongs [ɑ̟˞ⁱ] and [aⁱ] are used, but in Zevenbergen the diphthong [ɛⁱ] is used. In Oudenbosch and Roosendaal the strongly related diphthongs [æⁱ] and [ɛ˞ⁱ] are used. However in Ossendrecht we find the monophthong [ɛ˞]. 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.