Language Variation and Change, **13** (2001), 375–400. Printed in the U.S.A.

Dialect areas and dialect continua

WILBERT HEERINGA AND JOHN NERBONNE
University of Gronigen

ABSTRACT

The organizing concept behind dialect variation is still seen predominantly as the areas within which similar varieties are spoken. The opposing view—that dialects are organized in a continuum without sharp boundaries—is likewise popular. This article introduces a new element into the discussion, which is the opportunity to view dialectal differences in the aggregate. We employ a dialectometric technique that provides an additive measure of pronunciation difference: the (aggregate) pronunciation distance. This allows us to determine how much of the linguistic variation is accounted for by geography. In our sample of 27 Dutch towns and villages, the variation ranges between 65% and 81%, which lends credence to the continuum view. The borders of well-established dialect areas nonetheless show large deviations from the expected aggregate pronunciation distance. We pay particular attention to a puzzle concerning the subjective perception of continua introduced by Chambers and Trudgill (1998): a traveller walking in a straight line from village to village notices successive small changes, but seldom, if ever, observes large differences. This sounds like a justification of the continuum view, but there is an added twist. Might the traveller be misled by the perspective of most recent memory? We use the Chambers–Trudgill puzzle to organize our argument at several points.

> Accordingly, some students now despaired of all classification and announced that within a dialect area … there were no real boundaries, but only gradual transitions.… (Bloomfield, 1933:343)

The organizing concept behind dialect variation is still seen predominantly as the areas within which similar varieties are spoken. The opposing view—that dialects are organized in a continuum without sharp boundaries—is often alluded to not only by Bloomfield (1933) but also by frustrated researchers attempting to determine the boundaries predicted by the areal view (see, e.g., Tait, 1994). This article aims at introducing a new element into this traditional discussion: the opportunity to view dialectal differences in the aggregate.

Throughout, we focus on the pronunciation differences in a small sample ($N = 27$) of Dutch towns, in the hope that other levels of linguistic structure might

We thank Peter Kleiweg for his graphic programs, which were used to construct all the maps in this article. We thank Frits Steenhuisen for providing the coordinates for the maps. The article was written as a consequence of a LOT Summer School course at the Tillburg University given by Jack Chambers. We thank him for his valuable suggestions and remarks. The article is part of a larger dialect project under the supervision of John Nerbonne.

yield insight to similar analyses. We introduce a dialectometric technique that provides an additive measure of pronunciation difference when applied to varying dialectal pronunciations. We apply this technique to over 125 words at a series of towns and villages and call the result the (aggregate) pronunciation distance and also the phonological distance. This allows us not only to rise above the difficulties of identifying particular isoglosses as more significant (Bloomfield, 1933:344), but also to ask very simply how much of the linguistic variation we find is accounted for by geography. The fact that 65% of pronunciation difference is accounted for by geographic distance in this study lends credibility to the continuum view.

Our conclusion is that, while a great deal of pronunciation variation is very simply accounted for by geography, an interesting amount remains. In particular, the borders of well-established dialect areas show large deviations from the expected aggregated pronunciation distance.

Chambers and Trudgill (1998) introduced an interesting puzzle, one that is related to whether dialects should be viewed as organized by areas or via a geographic continuum. They observed that a traveller walking in a straight line notices successive small changes from village to village, but seldom, if ever, observes large differences. This sounds like a justification of the continuum view, but there is an added twist: might the traveller be misled by the perspective of most recent memory? We use the Chambers–Trudgill puzzle to organize our argument at several points.

Dialectometry

Dialectometry means literally "the measure of dialect." Jean Séguy, director of the *Atlas linguistique de la Gascogne*, coined the term. Séguy and his associates were accomplished dialectologists who published six atlas volumes containing maps of exquisite detail (Chambers & Trudgill, 1998:137). However, Séguy looked for a way to analyze the maps in a more objective way than was possible with traditional methods. Therefore, he introduced a new concept, keeping track of the points at which dialectal varieties differ and recording the differences in what amounts to a dissimilarity matrix. The number of disagreements between two neighbors was expressed as a percentage, and the percentage was treated as a measure indicating the linguistic distance between any two places (Chambers & Trudgill, 1998:138). The last ten pages of the sixth volume of the atlas contain dialectometric maps. We provide an explanation of our alternative fundamental technique, the measure of pronunciation distance, later.

Areas and continua

The dialectal landscape is also often described as a continuum. Chambers and Trudgill (1998) suggested the perspective of a traveller, going from village to village in a particular direction, who would notice linguistic differences that distinguished one village from another. As Chambers and Trudgill (1998) noted, it is essential to the continuum view that these differences are "cumulative," which

means that the further we get from our starting point, the larger the differences become. Mostly the villagers of two successive villages will understand each other's dialects very well, but the longer the chain, the greater the likelihood that the dialects on the outer edges of the geographical area will not be mutually intelligible. At no point is there such a complete break that geographically adjacent dialects are not mutually intelligible, but the extent to which dialects are mutually intelligible seems to depend on their geographic distance in the continuum perspective.

Would a traveller walking in the dialect landscape notice only gradual changes? Or would one notice only abrupt changes (i.e., borders)? From the view of Chambers and Trudgill's traveller, we study the terms "dialect areas" and "dialect continuum." The main tool we use for this study is a dialectometric method: the Levenshtein distance.

If dialects were perfectly divided into areas, the distances (measured by a dialectometric method) between dialects in one area would all be zero. The traveller would not notice any difference. But then, when leaving the one area and entering the next, one would notice big differences. Somewhere between villages there is a border. Exaggerating somewhat, the traveller would get the following impression: one step and we leave, say, the Saxon area and enter the Franconian area.

If the dialect landscape were a perfect continuum, the traveller would never notice that dialects were the same, nor that there were abrupt changes, but the extent to which dialects change could be predicted by geographic distance. The further the traveller is from a starting point, the more differences accumulate. So Chambers and Trudgill (1998:5–7) described the distances as being cumulative. This may also be seen in the Rhenish Fan (Bloomfield, 1933). It falls along German–Romance language border from the Schelde (northwest) to Alsace (southeast), in which 30 parallel isoglosses can be found. A traveller travelling from the first to the thirtieth isogloss would find that differences are cumulative.

Overview

In this article, we study the concepts of dialect area and dialect continuum. In order to focus on the Chambers–Trudgill puzzle, we look at 27 dialects that lie on a straight line. Using Levenshtein distances, we calculate the linguistic distances between all pairs of these dialects. Next, we research the relation between phonological distances and geographic distances. Using regression we determine how much variation can be explained by geographic distance. We then show how dialect areas can nonetheless be identified. Like the arrow method (Daan & Blok, 1969), a distance greater than a certain threshold indicates a border. Clustering shows the dialect areas implied by the linguistic distances. Clustering also indicates that geographic information is reflected in phonological distances to a certain extent. This is the fundamental dialectological postulate, which we employ in a novel way here: our measure of pronunciation difference succeeds to an extent that allows the extraction of geographic information. This leads us back to the idea of the direct continuum. Next, we examine the relativity of the term "border." We show that distances are not completely cumulative and exhibit the

FIGURE 1. The locations of the 27 Dutch dialects studied.

shape of a continuum with respect to a starting point. Finally, we use multidimensional scaling to show how the dialects are related to each other.

DIALECT DATA

The data used for comparing dialects from the *Reeks nederlands(ch)e dialectatlassen* (RND) was compiled by Blancquaert and Peé (1925–1982). From these atlases we chose 27 sites that roughly form a straight line from northeast to southwest in the Dutch language area (see Figure 1). In the RND, the same 141 sentences were recorded and transcribed for each dialect. From these sentences we chose 125 words that we thought were representative of the range of sounds in the varieties. The word lists from the atlas usually contained one word per concept, but sometimes more. We used all the words the atlas provided (for a given concept). The different words may be a reflection of social status, but this was not usually recorded. It seems that the RND interviewers did not consciously distinguish social status. Table 1 provides information about the informants; the peri-

https://doi.org/10.1017/S0954394501133041 Published online by Cambridge University Press

TABLE 1. *Information concerning the 27 dialects from the RND*

Place	NOW	Sex	Prof.	Age	Period	Volume
Scheemda	9	m/m	n/n	64/72	1956–1961	16
Veendam	4	m/m	n/n	69/62	1956–1961	16
Eext	7	m/m/f	a/a/a	57/58/61	1956–1961	16
Driebergen	2	f/m/m/m	n/n/n/n	50/81/80/59	1950–1962	11
Koekange	0	?/m	?/a	76/73	1974–1975	14
Hasselt	0	m/m	n/a	62/66	1974–1975	14
Staphorst	1	m/m	a/a	69/47	1974–1975	14
Zalk	0	f/m	?/a	52/56	1974–1975	14
Oldebroek	0	f/m	?/n	53/32	1974–1975	14
Nunspeet	1	m/m	n/a	50/53	1950–1970	12
Putten	7	f/f/f	n/a/a	38/28/54	1950–1970	12
Amersfoort	4	m/f	n/n	71/58	1950–1970	12
Beilen	5	f/m	a/n	74/36	1956–1961	16
Ruinen	0	f/f/f	?/?/?	59/67/65	1974–1975	14
Ossendrecht	2	m/f/m	n/n/n	63/22/18	1933–1935	3
Clinge	0	m/m/m	n/s/s	39/13/12	1933–1935	3
Moerbeke	0	m/m/m	s/n/n	23/20/54	1933–1935	3
Lochristi	1	m/m/m	n/n/n	52/29/48	1933–1935	3
Vianen	1	m/m/m	a/n/a	66/30/61	1950–1962	11
Hardinxveld	1	m/m	n/n	77/80	1939–1949	9
Zevenbergen	0	m/m/m	n/?/n	36/79/41	1939–1949	9
Oudenbosch	1	m	n	46	1939–1949	9
Roosendaal	0	m/f/m	n/n/n	63/63/27	1939–1949	9
Bellegem	4	m/m	n/n	35/69	1934–1940	6
Nazareth	0	m/f/m	n/n/s	25/20/24	1927–1930	2
Waregem	4	m/m	n/n	77/63	1934–1940	6
Zwevegem	0	m/m	n/n	35/33	1934–1940	6

Note: NOW = number of words for which more than one variant was used. Prof. = professions of the informants. Here we distinguished the following categories: a = agricultural, n = nonagricultural, s = student, ? = unknown. For housewives the profession of their husband was given. Ages = ages as given in the RND. For Scheemda, Veendam, Eext, and Beilen the birth dates were given. We calculated the ages by calculating the difference between the date of birth and the mean of the first year and the last year of the recording period. Period = recording period. Volume = part of the RND in which the dialect is found.

ods of recording are taken from Wijngaard and Belemans (1997). Nunspeet, Putten, Amersfoort, and Driebergen are located in the transition zone between the Saxon and Franconian area. Nunspeet and Putten were recorded in the period 1950–1970, whereas Amersfoort and Driebergen were recorded in 1950–1962. The recordings of Nunspeet and Putten were not made by the same person, but the recordings of Amersfoort and Driebergen were.

Figures 2, 3, 4, and 5 show word variation for 4 of the 125 words. These figures are also called "display maps" (Chambers & Trudgill, 1998:25). Although the maps give the viewer an idea of the variation between dialects, it would be very difficult, perhaps impossible, to draw generalizations about the dialect gradation

FIGURE 2. Variants of *deur* ‘door’ in IPA.

from such displays. However, this is possible when using the Levenshtein distance measured over large samples of vocabulary. Using Levenshtein distance, the size of the difference between variants of words can be calculated, and the distances can be aggregated over all words.

COMPARISON OF DIALECTS

The Levenshtein distance was applied by Kessler (1995) to Irish Gaelic dialects with remarkable success and by Nerbonne, Heeringa, van den Hout, van der Kooi, Otten, and van de Vis (1996) to Dutch dialects with respectable results. The Levenshtein distance is presented in Kruskal (1999).

The Levenshtein distance may be understood as the cost of (the least costly set of) operations mapping one string to another. The basic costs are those of (single-phone) insertions, deletions, and substitutions. Insertions and deletions are half as costly as substitutions. The principle can be illustrated by a small example. In

https://doi.org/10.1017/S0954394501133041 Published online by Cambridge University Press

FIGURE 3. Variants of *potten* 'pots' in IPA.

Standard American "saw a girl" is pronounced as [sɔ:əgʌrl]. In Boston it is pronounced as [sɔ:rəgɜ:l]. Now we can change the first pronunciation into the second as follows:

sɔəgɪrl	delete r	1
sɔəgɪl	replace ɪ/ø	2
sɔəgøl	insert r	1
sɔrəgøl		
Total		4

This example has been simplified in order to clarify the fundamental idea. It is crude to treat segments as alike or different simpliciter. In fact, many sequence operations map [sɔ:əgʌrl] into [sɔ:rəgɜ:l]. In the worst case, first we delete all sounds of the first pronunciation (7 deletions), and then we insert all sounds of the second pronunciation (7 insertions). We get a total of 14. However, there is an

https://doi.org/10.1017/S0954394501133041 Published online by Cambridge University Press

FIGURE 4. Variants of *zijn* 'to be' in IPA.

algorithm that always finds the cheapest mapping. In our example this gives a cost of 4.

The simplest versions of Levenshtein distance are based on calculations of phonological distance in which phonological overlap is binary: nonidentical phones contribute to phonological distance, whereas identical ones do not. Thus the pair [a,p] counts as different to the same degree as [b,p]. In more sensitive versions phones are compared on the basis of their feature values, so that the pair [a,p] counts as much more different than [b,p]. The measurements we employ here are sensitive to segmental similarity in exactly this way.

We experimented with two systems to guard against special dependency. One was developed by Hoppenbrouwers and Hoppenbrouwers (1988) and described in Hoppenbrouwers and Hoppenbrouwers (1993) and Hoppenbrouwers (1994); the other was constructed by Vieregge, Rietveld, and Jansen (1984). The Hoppenbrouwers' system is based on Chomsky and Halle's (1968) *The Sound Pattern of English* and consists of 21 binary features that apply to all phones (vowels and

https://doi.org/10.1017/S0954394501133041 Published online by Cambridge University Press

FIGURE 5. Variants of *wijn* 'wine' in IPA.

consonants). Vieregge et al.'s system consists of 4 multivalued features only for vowels and 10 multivalued features only for consonants. We combined these systems into one system for both vowels and consonants, where default values were assigned to the vowel features of consonants and the consonant features of vowels. Vieregge et al.'s system was developed for a similar comparison task: that of checking the quality of phonetic transcriptions. This involves comparison to consensus transcriptions. The results reported in this article are based on the system of Vieregge et al.; details are given in Nerbonne and Heeringa (1998).

Assume that 125 words are transcribed from two different dialects. The Levenshtein distance can then be calculated for 125 word pairs; the total distance between the dialects is equal to the sum of the 125 Levenshtein distances. In this article, we call this total distance "linguistic distance," "pronunciation distance," or "phonological distance." Note that the distance depends only on segmental phonetics, ignoring stress and tone suprasegmentals. We are, of course, aware that morphology, syntax, and even semantics can also vary in dialects and suspect

https://doi.org/10.1017/S0954394501133041 Published online by Cambridge University Press

that interesting, perhaps related, techniques could be developed for other linguistic levels.

Nerbonne et al. (1996), Nerbonne and Heeringa (1998), and Nerbonne, Heeringa, and Kleiweg (1999) all applied Levenshtein distance to Dutch dialects.

LINGUISTIC VERSUS GEOGRAPHIC DISTANCES

If a dialect landscape were a perfect continuum with (more or less sharp) borders, then linguistic distances would completely depend on geographic distances. To find the extent to which linguistic and geographic distances were related to each other, we correlated them and performed regression analyses.

As the basis for the following discussion, we calculated the geographic distances on the basis of coordinates given by Map Blast, a mapping program at http://www.mapblast.com. A coordinate pair consists of a latitude (north-south axis) and longitude value (east-west axis), where degrees are given as decimal values. We calculated the Euclidean distance between any two points as the square root of the sum of the square of the latitude value and the square of the longitude value. Using this coordinate system gives some distortion, because it ignores the Earth's curvature, but since the area under consideration is small, the distortion is minimal.

Correlation

The correlation coefficient between the phonological distances and the geographic distances turned out to be equal to $r = .8054$, which is highly significant.[1] This means that 65% ($r^2 \times 100$) of the aggregate phonological variation is accounted for by distance, with no particular appeal to discrete areas. We also calculated pronunciation distances on the basis of each word separately, thus obtaining 125 distance matrices. We correlated each of them to the geographic distances. The word *groen* 'green' has the highest correlation: .6842, which is significant. The word *zoon* 'son' has the lowest correlation: −.0885. The mean of all separate word correlation values is .3372, with a standard deviation of .1784.

Note that the highest correlation with distances on the basis of one word (.6842) is lower than the correlation with distances that are equal to the mean of 125 word distances (.8054). It is usually the case that averages show higher correlations than component scores, although theoretically this need not be the case. We chose to focus on average pronunciation distance since this represents the distance between (aggregate) varieties. When travelling from village to village along a chain, the gradual change we notice is not based on a single word but on many words. Each word separately may change at several different positions in the chain, and the number of variants per word may be different. So while a single word may have a lower correlation, the combination of the words will typically have a higher correlation.

https://doi.org/10.1017/S0954394501133041 Published online by Cambridge University Press

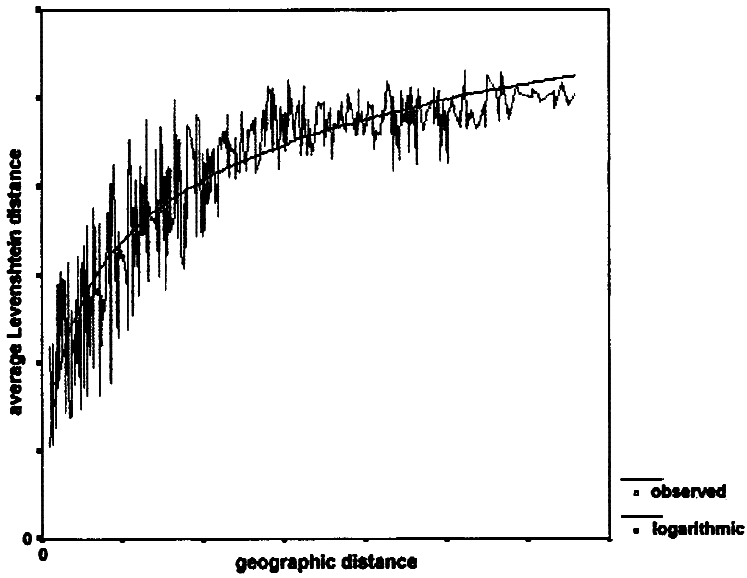

FIGURE 6. Geographic distances vs. average Levenshtein distances. Two successive points are connected by a straight line, illustrating the range of variation for average Levenshtein (pronunciation) distance. In SPSS the logarithmic regression line was drawn. Note that the logarithmic line seems to overestimate the pronunciation differences associated with greater distances.

It would be fallacious to conclude that dialect areas cannot account for more than 35% of the variation in linguistic distance. Especially seeing that areas and geography are strongly associated, the explained variation might be larger.

Explaining linguistic distances using geography

We used regression to fit the relation between phonology and geography into a formula. The formula may represent a linear relation, but several more complex relations are also possible. Using SPSS, we found that the logarithmic regression line represents the relation between the phonological and geographic distances fairly well. This is represented in Figure 6. The logarithmic correlation coefficient is equal to .899, which means that 81% of the variation in the phonological distances is explained by the logarithmic geographic distances. The logarithmic regression line represents the relation between phonological and geographic distances well because local distances are more significant than remote ones. At more remote places, phonological distances increase more slowly with respect to the starting point (in our study this is Scheemda). For dialects far away it matters more that they are far away and less how far away they are.

Once the relation is fixed in a formula, the expected phonological distances can be calculated on the basis of the geographic distances. The dialects of our

https://doi.org/10.1017/S0954394501133041 Published online by Cambridge University Press

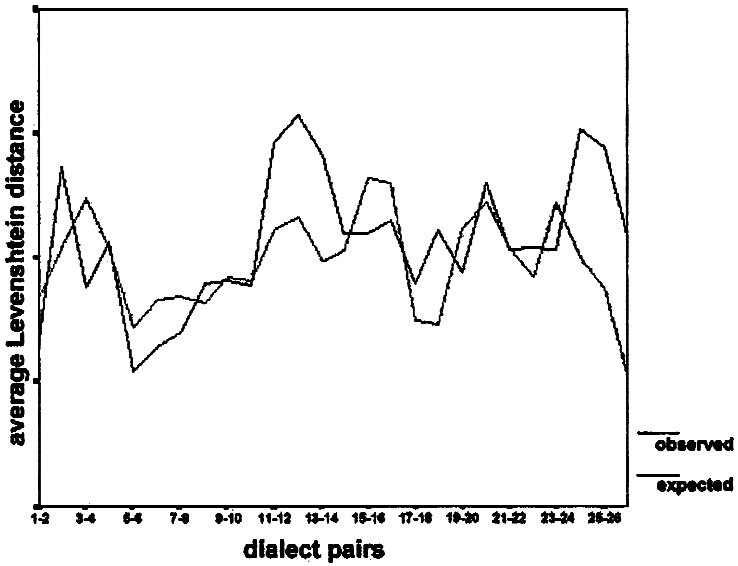

FIGURE 7. The observed and expected Levenshtein distances between successive dialects on the path of Chambers and Trudgill's traveller. The dialects are numbered from northwest to southeast in the same order as on the geographic map (Figure 1). The points at which observed and expected distances differ greatly suggest themselves as candidates as borders of distinct areas.

dataset lie on the straight line. Now we are especially interested in distances between two geographically successive dialects. In Figure 7 the real intermediate phonological distances are compared to the expected intermediate phonological distances (those predicted by the regression analysis).

Even though geographic distance is an excellent predictor of phonological distance, subtracting the expected values from the observed values leaves residues: that is, differences between actual and predicted values. If a residue is positive, the phonological distance between two successive points is greater than we would expect on the basis of their geographic distance. Large positive residues are points at which we may suspect a dialect area border. If a residue is negative, the distance between two successive points is smaller than we would expect on the basis of their geographic distance. In order to focus on significant values, we transformed the residues to *z* values (i.e., standard deviations). We calculated the mean and the standard deviation on the basis of the residues of all possible dialect pairs (regardless of whether they were adjacent). Next we calculated *z* values (differences expressed in terms of standard deviations) and determined the accompanying statistical significance. We found that the distances between Putten and Amersfoort, Amersfoort and Driebergen, Oudenbosch and Roosendaal, Nazareth and Waregem, Waregem and Zwevegem, and Zwevegem

https://doi.org/10.1017/S0954394501133041 Published online by Cambridge University Press

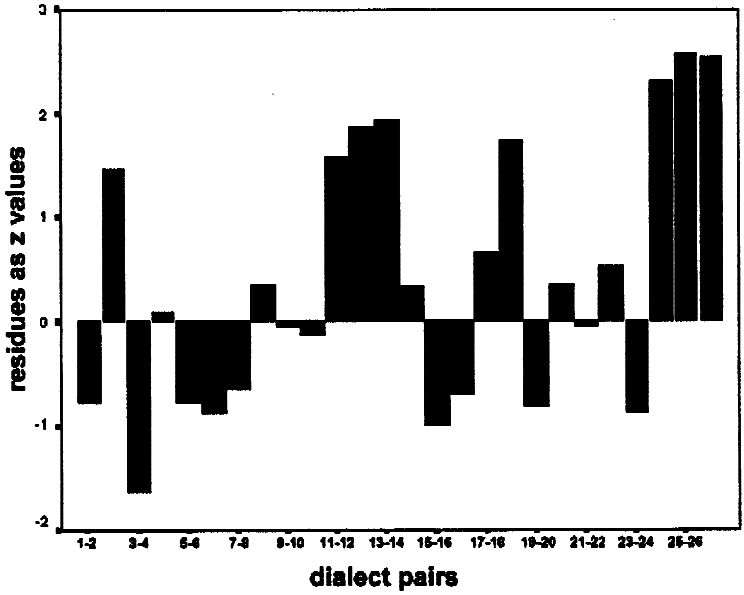

FIGURE 8. Differences between observed and expected average Levenshtein distances for all pairs of two successive dialects, given as *z* values (standard deviations). The dialects are numbered from northwest to southeast in the same order as on the geographic map (Figure 1). The large positive residues mark points at which one might expect borders between distinct areas.

and Bellegem were significantly higher than one would expect on the basis of their geographic distance. This is shown in Figure 8.

Looking at the residues, we see that phonological distances can mostly be explained by geographic distance. This justifies the continuum perspective. In those cases where a dialect distance between successive points is significantly higher than would be expected on the basis of geographic distances, we may encounter a dialect border. This justifies the area perspective.

DIALECT AREAS

It would also be possible to apply regression to determine the extent to which dialect areas might explain linguistic distance. To be convincing, this would involve a large, carefully chosen sample of varieties. Our sample here was chosen to investigate areas and continua from the perspective of Chambers and Trudgill's traveller. We therefore postpone the determination of the contribution of areas to phonological distance until future work.

https://doi.org/10.1017/S0954394501133041 Published online by Cambridge University Press

Arrow method

In traditional dialectology, researchers sought dialect areas, trying to find borders that separate one area from another. As an example, we mention the dialect map in Daan and Blok (1969). For the Netherlandic part of the Dutch language area Daan and Blok used the arrow method to find dialect borders. Dialects that speakers judge to be similar are connected by arrows. Bare strips, where no arrows are placed, show dialect area borders.

The arrow method focuses on when a speaker judges a dialect as (nearly) the same as one's own and when not. When walking from northeast to southwest, when will Chambers and Trudgill's traveller judge a change as a border? This will be the case when the difference exceeds some threshold.

With Levenshtein distance the distance between each pair of two contiguous sites can be measured. Perhaps this can be construed as quantifying the arrow method. When the Levenshtein distance between two dialects exceeds some threshold, we might hypothesize that these dialects are separated by a dialect border.

How do we fix the threshold? Earlier we described how to split up phonological distances into a geographic component and a residual part. Converting the residues to z values, it is possible to calculate the likelihood that the residual distance between two dialects is equal to or greater than the observed distance. When the likelihood is lower than a reasonable value α, then the residue represents a significant deviation from the distance that would be expected on the basis of the geographic distance. The α value is the threshold. We use $\alpha = .05$.

Looking at Figure 8 we see that the Saxon dialects (numbers 1–12) and the Franconian dialects (numbers 14–27) were separated by two borders, namely between 12 and 13 (Putten and Maersfoort) and 13 and 14 (Amersfoort and Driebergen). The p values were, respectively, .0294 and .0262. So the distinction between both areas was very clear. Between 18 and 19 (Oudenbosch and Roosendaal), we also found a border. The p value was .0409. Furthermore there were borders between 24 and 25 (Nazareth and Waregem), 25 and 26 (Waregem and Zwevegem), and 26 and 27 (Zwevegem and Bellegem). The p values were, respectively, .0102, .0049, and .0057. Possibly this can be explained by the fact that Nazareth, Waregem, and Zwevegem approach the Flemish–French border, and so they belong to the French–Flemish transition zone.

Clustering distances

If dialect areas exist, we can find them by applying clustering (Jain & Dubes 1988). The result is an hierarchically structured tree in which the dialects are the leaves.

Calculating the distances among the 27 dialects gives a 27×27 matrix, on the basis of which we can cluster the dialects. Clustering here is most easily understood procedurally. In the matrix only the upper half is used. At each iteration of the procedure, we select the shortest distance in the matrix. Then we fuse the two data points that gave rise to it. To iterate, we assign a distance from the newly formed cluster to all other points. For example, if point A and point B are fused to

https://doi.org/10.1017/S0954394501133041 Published online by Cambridge University Press

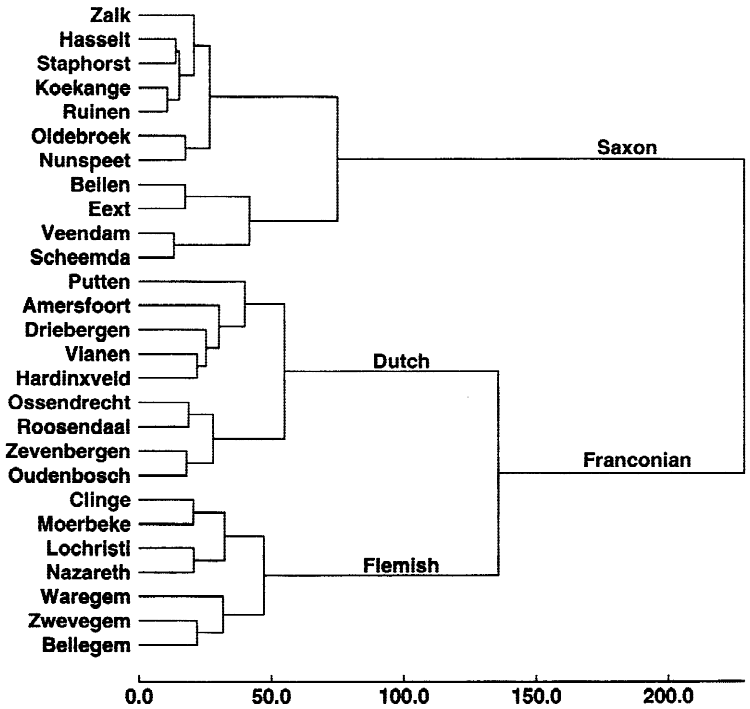

FIGURE 9. A dendrogram derived from the distance matrix based on Levenshtein distances as measured on 125 words. The feature system of Vieregge et al. (1984) is used; diphthongs are represented as one phone; Manhattan distance between feature bundles is calculated. The two main groups are the Saxon (top) and Franconian dialects (bottom). Within the Franconian dialects a Dutch (upper) and a Flemish (lower) subgroup can be distinguished.

one cluster AB, the distance between a point S and cluster AB could be defined as the average of the distance between S and A and the distance between S and B. Besides the average, there were several other alternatives of which Ward's method turned out to be most suitable for our research. Ward's method is very similar to using the average, but it minimizes squared error (Jain & Dubes, 1988). Note that the method is always forced to find groups.

The result may be seen in Figure 9. As the two most significant groups, a Saxon group and a Franconian group emerged, with a border between Nunspeet (Saxon) and Putten (Franconian). Within the Franconian group, a Dutch subgroup and a Flemish subgroup could be found. A border could be drawn between Ossendrecht (Dutch Franconian) and Clinge (Flemish Franconian). The dendrogram accords with the geography in the sense that, for each pair that fall within a single dialect group, all intermediate points fall within the group as well. Here we see that classification yields geographic information from the phonological dis-

https://doi.org/10.1017/S0954394501133041 Published online by Cambridge University Press

tances, while it is a characteristic of a continuum and a fundamental dialectological assumption that dialect distances are related to geographic distances fairly directly.

DIALECT CONTINUUM

In this section we undertake a closer investigation of the dialect continuum. We show that there may be parallel isoglosses, which suggests the relativity of the term "border" given in traditional methodology. Next we show that distances are not completely cumulative. We also try to represent the shape of a one-dimensional continuum in a two-dimensional plot. Finally, we apply multidimensional scaling (Kruskal & Wish, 1984) to the Levenshtein distances. The result is a map where the distance between kindred dialects is small and that between different dialects is great.

In this article we do not specifically research lexical diffusion. Lexical diffusion is the hypothesis that sound change proceeds word by word, where each change spreads in a wave, leaving residues of non-overlapping differences. In particular, non-overlapping residues of waves of changes could easily result in a continuum of varieties of the sort we explore here.

Parallel isoglosses

As we have seen, the Saxon area and the Franconian area are separated by three borders. This may correspond with the fact that not all isoglosses coincide, but may be parallel to each other. If we consider Figure 2, we see that in most Saxon dialects [døːr] is followed by [ə], whereas for the Franconian dialects this is never the case. So we see an isogloss between Nunspeet and Putten. If we look at Figure 4, we see that in most Saxon dialects a variant of [bɪn] is used, whereas in the Franconian dialects a variant of [zɛˑⁱn] is used. So there is an isogloss between Putten and Amersfoort. In Figure 5 we see that in all Saxon dialects the vowel in [vɛˑⁱn] is [i], whereas in the Franconian dialects it is always another vowel. So there is an isogloss between Amersfoort and Driebergen. Here we find three parallel isoglosses. This is a little bit like the Rhenish Fan, where no less than thirty parallel isoglosses are found. The presence of parallel isoglosses makes it clear that no sharp borders can be found by looking for coinciding isoglosses. Rather one should speak of transition zones. In the continuum view, this fact is taken into account in a suitable way.

Cumulative distances

A property of geographic distances is that they are simply cumulative. Assume three points A, B, and C, which lie on a straight line. It is certain that distance(A,C) = distance(A,B) + distance(B,C). For each site the distances can be calculated in two ways: indirectly and directly. If calculating indirectly, we can measure the distance via the intermediate points:

https://doi.org/10.1017/S0954394501133041 Published online by Cambridge University Press

$$d(x_n, x_1) = d(x_n, x_{n-1}) + d(x_{n-1}, x_1)$$

Alternatively, if calculating directly, we take the direct distance as it is given:

$$d(x_n, x_1)$$

We could illustrate this perfect, simple cumulativity by illustrating the relation between direct and indirect measures for the 27 dialect points. For each location on the line the geographic distance would be compared to a starting point. As starting point we would take the site at the outer end of the line in the northeast and southwest. Next we could draw a scatterplot, where the *x* axis represents the direct and the *y* axis the indirect distances. The dots in the plot would lie on a straight line, representing a linear relation. The relation between indirect and direct geographic distances is linear, which is in accordance with the fact that geographic distances are cumulative.

The distinction between direct and indirect distances can be applied not only to geographic distances but also to phonological distances. Calculating phonological distances indirectly models the assumption that Chambers and Trudgill's traveller remembers only the last variety and the total accumulation until then. The memoryless traveller cannot compare the current variety to varieties much earlier on the path. We explored the model by drawing scatterplots in which indirect phonological distances were plotted as a function of direct phonological distances. The plot is shown in Figure 10. In contrast to the geographic distances, the plots do not show a linear curve; thus phonological distances are not simply cumulative.

The shape of continua

We showed earlier that phonological and geographic distances are related. We now try to understand the relation more deeply. Earlier we cited Chambers and Trudgill (1998:5): "If we travel from village to village, in a particular direction, we notice linguistic differences which distinguish one village from another." This suggests a novel perspective on linguistic variation. Rather than view the phonological distance from Scheemda to Bellegem directly, we adopt the traveller's perspective: one who notices the incremental differences between Scheemda and Veendam, between Veendam and Eext, and so on. Chambers and Trudgill's traveller develops a notion of indirect phonological distance, which is the sum of distances from pairwise neighboring points on a connected line. The question then is, what would this traveller's view of the dialectal landscape be? Figure 11 shows that the view is a linear relationship between geographic distance and the traveller's sum of incremental distances. This is the indirect phonological distance.

Of course the view of Chambers and Trudgill's traveller is misleading! Phonological distances do not sum along (geographic) paths. To examine the real relationship, we also needed to draw a scatterplot with direct geographic distance versus phonological distances, which would reflect the fact that phonological distances are not simply cumulative. The result can be found in Figure 12. The

https://doi.org/10.1017/S0954394501133041 Published online by Cambridge University Press

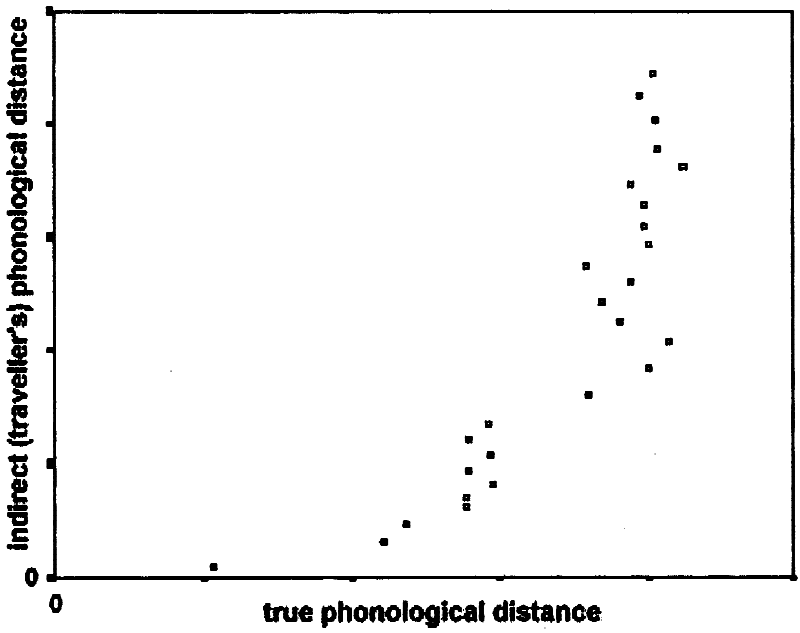

FIGURE 10. True phonological distance versus indirect (traveller's) phonological distance with Scheemda as starting point. The form suggests an exponential relation: the mirror of the logarithmic relation that we hypothesize exists between geographic and phonological distance.

relation is obviously not linear. The graph's slope clearly decreases as a function of distance. The relatively flat sections of the curve correspond to relatively homogeneous linguistic areas.

Why does the discrepancy arise between the traveller's indirect view and the true pronunciation difference? We suggest that it arises because the traveller is reacting to a global (aggregate) impression of the (pairwise) differences. As Figure 11 demonstrates, these accumulate in a linear fashion, giving the traveller the impression that the continuum is simple and dialectologically real. But a simple thought experiment demonstrates how fallacious this is. We can easily imagine a line in which two dialects alternate: first A, then B, then A, and so on. In this case the indirect accumulation would still grow linearly, while the true distance would be alternatively zero ($d(A,A) = 0$) or the distance between A and B ($d(A,B)$). The cumulative view loses track of local differences that may be lost again over a longer distance.

The contrast between Figures 11 and 12 is our analysis of the Chambers–Trudgill puzzle. The perception of the traveller is that one keeps hearing small differences, so that pronunciation difference is a simple, linear function of geography (distance). The pronunciation differences of all the towns and villages along

https://doi.org/10.1017/S0954394501133041 Published online by Cambridge University Press

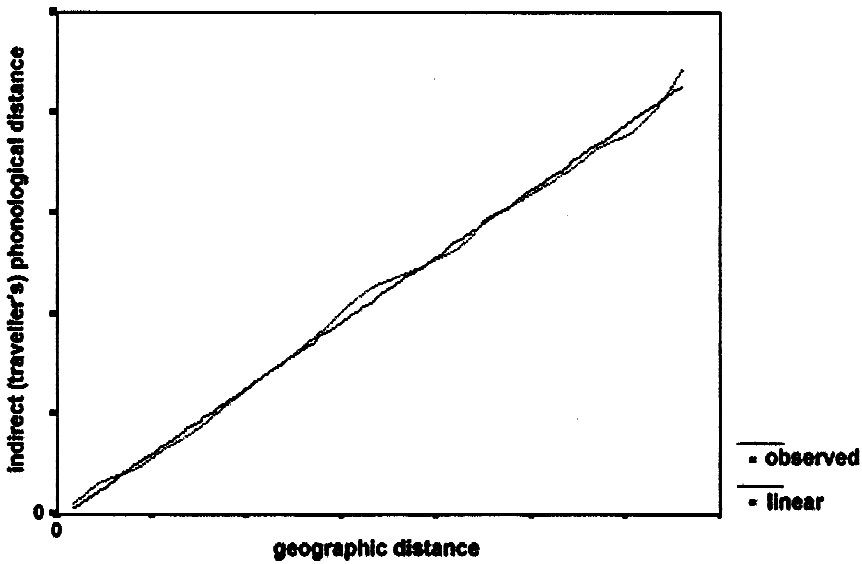

FIGURE 11. Geographic distance versus mean indirect (traveller's) phonological distance with Scheemda (northeast) as the starting point. Essentially the same graph results if one begins in Bellegem (southwest). This graph explains the perception of Chambers and Trudgill's traveller that the dialect landscape is a simple accumulation of differences.

the traveller's path accumulate linearly, as Figure 11 shows. But as Figure 12 shows, the traveller is deceived. The true pronunciation difference simply is not the sum of the pairwise differences along the path. We develop this contrast further later.

Naturally one can ask whether the individual words would give a different, perhaps clearer, picture of role of areas versus continua. To this end we plotted pronunciation distance versus geography by individual word in a manner parallel to the way that we examined the aggregate pronunciation difference: that is, first deceptively, as if the differences accumulated, and second directly, as they were measured. The results are shown in Figures 13 and 14.

These figures reinforce the earlier point made about Chambers and Trudgill's traveller. The cumulative view (Figure 13) is simplistic, ignoring the fact that local changes may be undone. If Figure 14 appears chaotic, perhaps that is an admonition that we ought to focus on aggregates, not individual words, as we study linguistic variation.

Multidimensional scaling

On the basis of geographic coordinates, it is possible to determine the distances between locations. The reverse is also possible: on the basis of the mutual pho-

https://doi.org/10.1017/S0954394501133041 Published online by Cambridge University Press

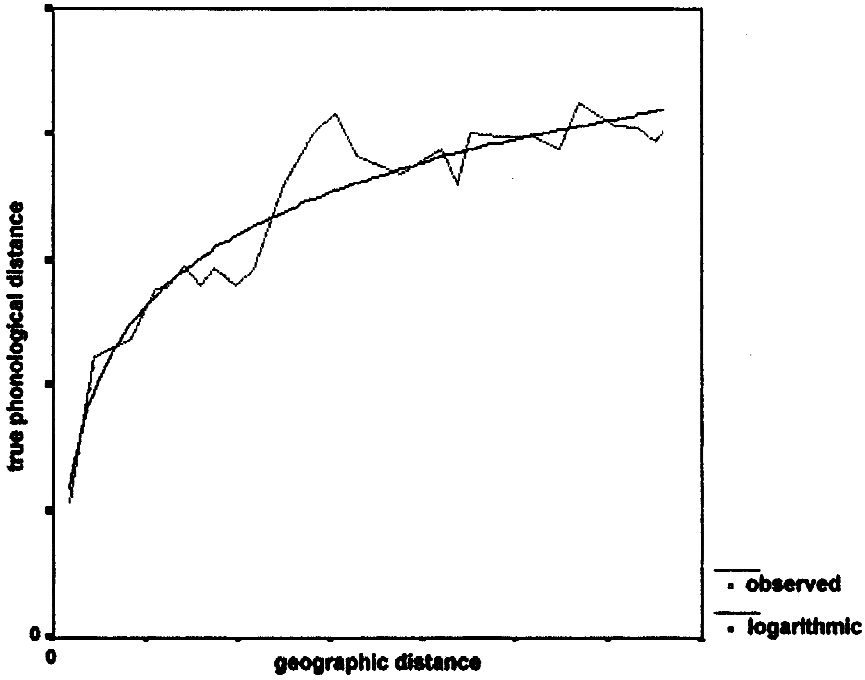

FIGURE 12. Geographic distance versus mean true phonological distance with Scheemda (northwest) as starting point. In SPSS the logarithmic regression line was drawn. A similar graph results if one begins at Bellegem (southwest). This graph illustrates the fallacy in the memoryless traveller's view of the dialect landscape. In fact, pronunciation differences accumulate slowly with respect to remote areas, although there are significant differences. Furthermore, the slope is not entirely smooth.

nological distances, an optimal coordinate system can be determined with the coordinates of the locations in it. The latter task is implemented by a technique known as multidimensional scaling. In a multidimensional scaling plot, closely related dialects are near each other, while very different dialects are located far away from each other (Kruskal et al., 1984).

As input each dialect is defined as a range of distances: that is, the distance to itself and the distances to other dialects. The distances correspond to dimensions. If we have 27 variants, we obtain 27 dimensions. With multidimensional scaling, the dimensions can be reduced to one, two, or more dimensions, and so we can obtain coordinates in, respectively, one, two, or more dimensional space. Here the one, two, or three dimensions still represent the information of all 27 dimensions as best as possible.

We scaled the 27 dimensions to two dimensions (see Figure 15). Although similarities with the geographic map can be identified, the plot does not show a

https://doi.org/10.1017/S0954394501133041 Published online by Cambridge University Press

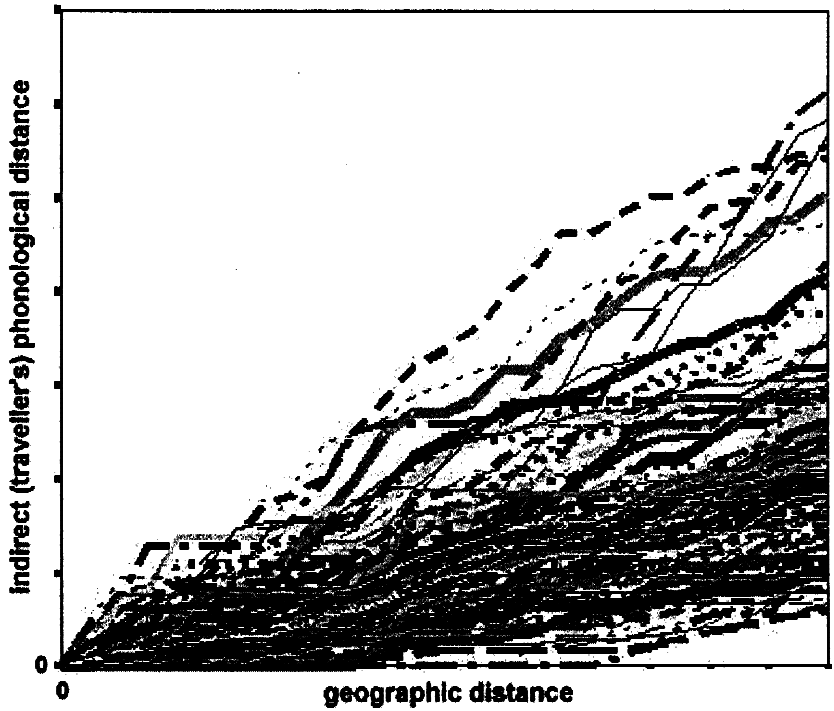

FIGURE 13. Geographic distances versus indirect (traveller's) phonological distances for each of the 125 words with Scheemda (northeast) as starting point.

straight line as on the geographic map. In the plot three groups could be clearly distinguished: a Saxon group, a Dutch Franconian group, and a Flemish Franconian group. By comparison to the geographic map, there is a border between Nunspeet (Saxon) and Amersfoort (Dutch Franconian) and between Ossendrecht (Dutch Franconian) and Moerbeke (Flemish Franconian). However, Putten lies exactly between the Saxon and Dutch Franconian dialects, and Clinge lies exactly between the Dutch Franconian and Flemish Franconian dialects. This again points to the necessity of the dialect continuum perspective.

In the multidimensional scaling plot, Saxon and Flemish Franconian are more closely related than the geographic line suggests. This can be partially explained by the fact that in both groups the final syllable [ən] is often reduced to a syllabic nasal [m̩], [n̩], or [ŋ̩], while in the Dutch Franconian group that syllable is reduced to [ə] (see Figure 3).

We tried to determine whether the data was uni- or multidimensional. Therefore, we scaled the data not only to two dimensions, but also to one dimension and to three and more dimensions. For each number of dimensions we calculated the squared correlation (*r*-squared, abbreviated as RSQ) between the given dialect

https://doi.org/10.1017/S0954394501133041 Published online by Cambridge University Press

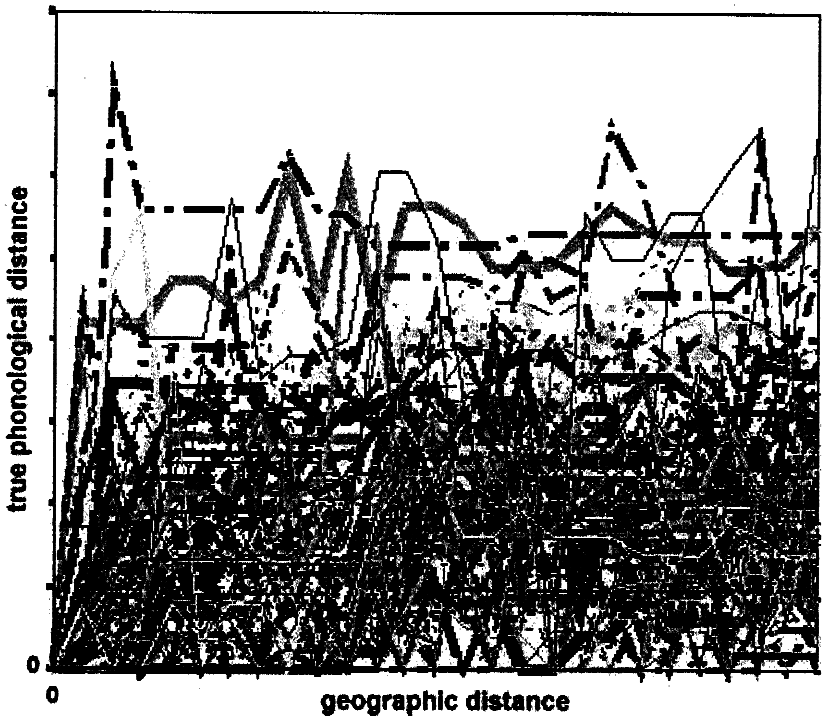

FIGURE 14. Geographic distances versus mean true phonological distances for each of the 125 words with Scheemda (northwest) as starting point. This is a metric perspective for bundles of (word) isoglosses. If such bundles existed, we would see points at which several lines rose together.

distances and the corresponding distances on the multidimensional scaling plot. RSQ can be interpreted as the proportion of variance in the distance that is accounted for by the distances between the points on the multidimensional scaling plot. Figure 16 gives the RSQ value for each number of dimensions. Here we see a clear difference between the RSQ value for one dimension on the one hand and the RSQ values for two or more dimensions on the other. This suggests that there are at least two dimensions, perhaps three. The fourth and additional dimensions explain very little of the variance in the data.

CONCLUSIONS

There is a strong correlation between the phonological distance and the logarithm of geographic distance (.9), accounting for 81% of the variation in pronunciation. The correlation per word in our sample varies greatly (from −.0885 to .6842 using a linear model). Using regressions, we could see how phonological dis-

https://doi.org/10.1017/S0954394501133041 Published online by Cambridge University Press

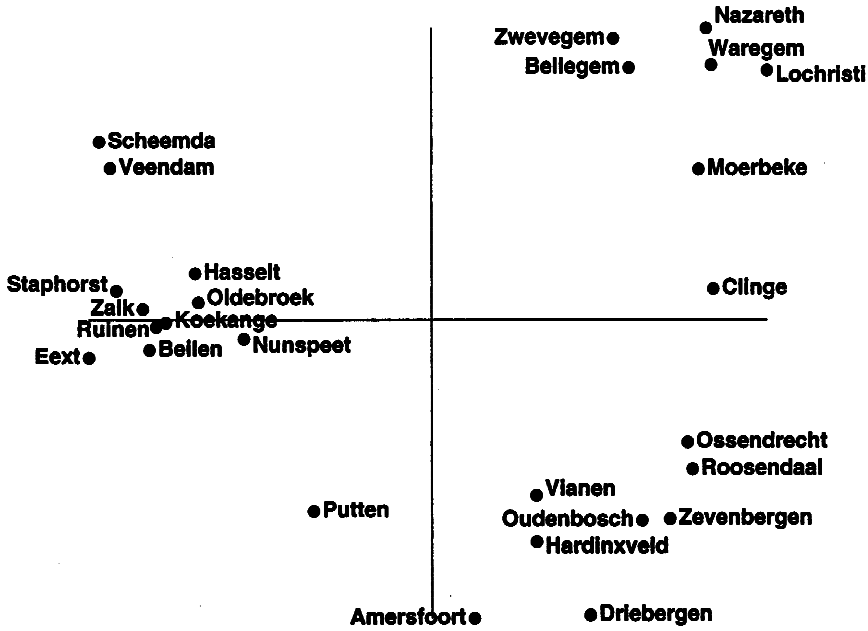

FIGURE 15. The two most significant dimensions in multidimensional scaling. The *x* dimension is more significant than the *y* dimension. The three main groups are the Saxon (left), Dutch Franconian (lower right), and Flemish Franconian (upper right) dialects. The dialects do not lie on a line as on the geographic map, but the three groups are clearly distinct.

tances depends on geographic distances. We argue that the logarithmic model is reasonable, proposing that distances further away are less significant than local distances. In the linear model, the correlation is also strong (.8).

The regression analysis suggests a novel perspective on dialect areas. When the distance between two dialects is significantly higher than would be expected on the basis of their geographic distance, we conclude that they are separated by a linguistic border between adjacent areas.

After clustering the dialects, the dendrogram accords with the geography in the sense that, for each pair that falls within a single dialect group, all intermediate points fall within the group as well. Heeringa, Nerbonne, and Kleiweg (2001) showed that the dialectometric method used here is validated by expert opinion on Dutch dialect areas.

In the Rhenish Fan isoglosses are parallel. This is also the case between the Saxon and Franconian area. Regarding the dialect landscape as a continuum accommodates this fact.

For each dialect on the line the distance could be calculated in relation to a starting point indirectly and directly. As indirect distance we take the sum of

https://doi.org/10.1017/S0954394501133041 Published online by Cambridge University Press

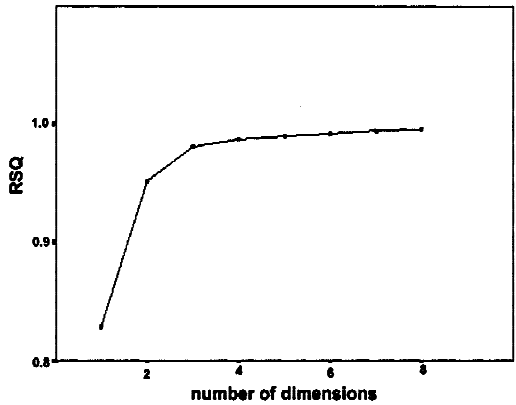

FIGURE 16. The dialect distances are scaled to 1, 2, 3, 4, 5, 6, 7, and 8 dimensions. For each number of dimensions the r^2 (RSQ) value is given as fit measure, ranging from 0 (perfect fit) to 1 (worst possible fit). This is the correlation of the phonological distances, with the distances in the proposed low-dimensional space. The plot suggests that there are at least two dimensions, and that the third is also informative.

all intermediate distances, each distance corresponding to two successive points. Because geographic distances are cumulative, the relation between indirect and direct distances is linear. For phonological distances the relation is not linear, and so they are not completely cumulative. This constitutes our perspective on the Chambers–Trudgill puzzle: the traveller perceives phonological distance indirectly and therefore is inclined to overestimate the real degree of change.

The relation between geographic distances and direct phonological distances can be represented as a continuum in a two-dimensional plot. Since phonological distances are not simply cumulative, we obtain a relation resembling a flattened logarithmic (or logistic) curve.

Although the two-dimensional plot has similarities with the geographic map, it does not show a straight line as on the geographic map. The fact that a second dimension explains a great deal of variation clearly suggests that a view of the dialectal landscape as a continuum should assume the multidimensional determinants of phonological distance. Geographic distance explains a great deal, but not everything. In the plot three groups could be clearly distinguished: a Saxon group, a Dutch Franconian group, and a Flemish Franconian group. Putten lies exactly between Saxon and Dutch Franconian, and Clinge lies exactly between Dutch Franconian and Flemish Franconian. This shows the need for the continuum view. In the plot, Saxon and Flemish Franconian are more closely related than the geographic line suggests. This can be partially explained by the fact that in the Saxon group as well as in the Flemish Franconian group the end syllable [ən] is often reduced to a syllabic nasal, while in the Dutch Franconian group that

https://doi.org/10.1017/S0954394501133041 Published online by Cambridge University Press

syllable is reduced to [ə]. When searching for significant dimensions, we find that there are at least two dimensions, and that the fourth and additional dimensions explain very little.

We conclude that both the area view and the continuum view are useful for gaining insight in the nature of the dialect landscape, which may be described as a continuum with varying slope or, alternatively, as a continuum with unsharp borders between dialect areas.

FURTHER WORK

In this study, only four varieties lie in the transition zone between Saxon and Franconian. It would be interesting to study this zone in a more detailed way, including many more varieties.

The continuum line we studied starts in the Saxon area and ends in the Franconian area. So the Frisian area is not involved. It would also be interesting to research a continuum line from Frisian to Saxon. The transition from Frisian to Saxon may be sharper than that from Saxon to Franconian. A line from Frisia to the Franconian area is not possible since the line would pass through a great deal of water for which no dialect data are available.

We are aware that the continuum we studied is a flat area. It would be interesting to research the role of mountains, rivers, or traffic in a continuum. We are collaborating with Charlotte Gooskens, who is applying similar techniques to Norwegian dialects.

For analytic purposes, we restricted the continuum to one dimension (i.e., the points lie on a line). It would also be interesting to research the continuum as it is: that is, in two dimensions. In order to represent a two-dimensional continuum, the graph should be three-dimensional, like a mountain landscape, where height represents the phonological distance with respect to, for instance, standard Dutch. In this larger set we plan to examine the degree to which areas can explain linguistic distances.

Finally, it would be interesting to explore other dialectal areas, particular those with well-known divergent factors such as national borders.

NOTE

1. In a study of 104 Dutch varieties, we obtained a pronunciation-geography correlation of $r = .68$ (Nerbonne et al., 1999:x).

REFERENCES

Blancquaert, E., & Peé, W. (1925–1982). *Reeks nederlands(ch)e dialectatlassen*. Antwerpen: De Sikkel.

Bloomfield, Leonard. (1933). *Language*. New York: Holt, Rinehart and Winston.

Chambers, J. K., & Trudgill, P. (1998). *Dialectology* (2nd ed.). Cambridge: Cambridge University Press.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. New York: Harper & Row.

Daan, J., & Blok, D. P. (1969). *Van randstad tot landrand; toelichting bij de kaart: Dialecten en naamkunde*. Amsterdam: Noord-Hollandsche Uitgevers Maatschappij.

Heeringa, W., Nerbonne, J., & Kleiweg, P. (2001). Validating dialect comparison methods. In W. Gaul & G. Ritter (eds.), *Classification, Automation, and New Media: Proceedings of the 24th Annual Conference of the Gesellschaft für Klassifikation*. Springer: Heidelberg. 445–452.

Hoppenbrouwers, C. (1994). De indeling van de zuidoostelijke steektalen. *TABU: Bulletin voor taalwetenschap* 24:37–63.

Hoppenbrouwers, C., & Hoppenbrouwers, G. (1988). De featurefrequentiemethode en de classificatie van nederlandse dialecten. *TABU: Bulletin voor taalwetenschap* 18:51–92.

_____ (1993). De indeling van noordoostelijke dialecten. *TABU: Bulletin voor taalwetenschap* 23:193–217.

Jain, A. K., & Dubes, R. C. (1988). *Algorithms for clustering data*. Englewood Cliffs, NJ: Prentice Hall.

Kessler, B. (1995). Computational dialectology in Irish Gaelic. In *Proceedings of the European Association for Computational Linguistics*. Dublin: EACL. 60–67.

Kruskal, J. B. (1999). An overview of sequence comparison. In D. Sankoff & J. Kruskal (eds.), *Time warps, string edits, and macro molecules: The theory and practice of sequence comparison* (2nd ed.). Stanford: CSLI. 1–44.

Kruskal, J. B., & Wish, M. (1984). *Multidimensional scaling*. Beverly Hills: Sage Publications.

Nerbonne, J., & Heeringa, W. (1998). Computationele classification van nederlandse dialecten. *Taal en Tongval* 50:164–193. http://www.let.rug.nl/~nerbonne/papers/tetv99.ps.

Nerbonne, J., Heeringa, W., Hout, E. van den, Kooi, P. van der, Otten, S., & W. van de Vis (1996). Phonetic distance between Dutch dialects. In G. Durieux, W. Daelemans, & S. Gillis (eds.), *Papers from the Sixth CLIN Meeting*. Antwerp: University of Antwerp, Center for Dutch Language and Speech. 185–202. http://www.let.rug.nl/~nerbonne/papers/clin96.ps.

Nerbonne, J., Heeringa, W., & Kleiweg, P. (1999). Edit distance and dialect proximity. In D. Sankoff & J. Kruskal (eds.), *Time warps, string edits, and macro molecules: The theory and practice of sequence comparison* (2nd ed.). Stanford: CSLI. x–xv. http://www.let.rug.nl/~nerbonne/papers/timewarp.ps.

Tait, M. (1994). North America. In C. Moseley & R. Asher (eds.), *Atlas of the world's languages*. London: Routledge. 3–30.

Vieregge, W. H., Rietveld, A. C. M., & Jansen, C. I. E. (1984). A distinctive feature based system for the evaluation of segmental transcription in Dutch. In *Proceedings of the 10th International Congress of Phonetic Sciences*. Dordrecht: Foris. 564–659.

Wijngaard, H. H. A. van de, & Belemans, R. (1997). *Nooit verloren werk: Terugblik op de Reeks nederlandse dialectatlassen (1925–1982)*. Groesbeek: Stichting Nederlandse Dialecten.