Dialect Areas and Dialect Continua

Short title:
Dialect Areas and Dialect Continua

Author:
Wilbert Heeringa and John Nerbonne

Affiliations:
Humanities Computing, Faculty of Arts, University of Groningen

Mailing address:
P.O.Box 716, NL 9700 AS Groningen, The Netherlands

Phone:
+31 50 363 5970, +31 50 363 5815

Email:
heeringa@let.rug.nl, nerbonne@let.rug.n

Dialect Areas and Dialect Continua

Abstract

The organising concept behind dialect variation is still seen predominantly as realized by the areas within which similar varieties are spoken. The opposing view, that dialects are organised in a continuum without sharp boundaries is likewise popular. This paper introducing a new element into this traditional discussion, the opportunity to view dialectal differences in the aggregate. We employ a dialectometric technique which provides an additive measure of pronunciation difference the (aggregate) pronunciation distance. This allows us to determine how much of the linguistic variation we find is accounted for by geography – between 65% and 81% in our sample of 27 Dutch towns and villages, a fact 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 about the subjective perception of continua introduced by Chambers and Trudgill, who consider a traveller walking in a straight line and noticing successive small changes as he walks from village to village, but seldom, if ever large differences. This sounds like a justification of a the continuum view, but there is an added twist: might the traveller be misled by the perspective of most recent memory? We shall use the Chambers-Trudgill puzzle to organise this paper at several points.

Acknowledgements

We thank Peter Kleiweg for his graphic programs seen in all the maps in this paper. We thank Frits Steenhuisen for providing the coordinates for these maps. The paper is written as a consequence of a LOT Summerschool course at the Tilburg University given by Jack Chambers. We thank him for his valuable suggestions and remarks. The paper is part of a larger dialect project under the supervision of John Nerbonne.

1 Introduction

> 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 organising concept behind dialect variation is still seen predominantly as realized by the areas within which similar varieties are spoken. The opposing view, that dialects are organised in a continuum without sharp boundaries is often alluded to, not only by Bloomfield, but also by frustrated researchers who attempt to determine the boundaries predicted by the areal view.[1] This paper aims at introducing a new element into this traditional discussion, the opportunity to view dialectal differences in the aggregate.

Throughout this essay we shall focus only on pronunciation differences in a small sample (27) of Dutch towns, expressing the hope that other levels of linguistic structure might yield insight to similar analyses. We introduce a dialectometric technique which provides an additive measure of pronunciation difference when applied to varying dialectal pronunciations. We apply this over 125 words at a series of towns and villages, and call the result the (aggregate) pronunciation distance and also phonological distance. This allows us not only to rise above the difficulties of identifying particular isoglosses as more significant (Bloomfield 1933:344); it also allows us 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 the study below lends credibility to the continuum view.

Our conclusion, in brief, 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 aggregate pronunciation distance.

Chambers & Trudgill (1998) introduce an interesting puzzle that is related to the issue of whether dialects should be viewed as organised by areas or via a geographic continuum. They notice that a traveller walking in a straight line will notice successive small changes as he walks from village to village, but seldom, if ever will he notice 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

[1] See, e.g., Tait (1994).

memory? We shall use the Chambers-Trudgill puzzle to organise this paper at several points.

1.1 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, publishing six atlas volumes, containing maps with exquisite detail (Chambers et al. 1998:137). However, Séguy looked for a way to analyse the maps in a more objective way than was possible with traditional analytic methods. Therefore he introduced a new concept, keeping track of points at which dialectal varieties differ, and recording this in what amounts to a dissimilarity matrix. The number of disagreements between two neighbours was expressed as a percentage, and the percentage was treated as a measure indicating the linguistic distance between any two places (Chambers et al. 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, in section 3 below.

1.2 Areas and Continua

The dialectal landscape is also often described as a continuum. Chambers et al. (1998) suggest the perspective of a traveller going from village to village, in a particular direction. He would notice linguistic differences which distinguish one village from another. As Chambers et al. (1998) note, it is essential to the continuum view that these differences are 'cumulative', which means the further we get from our starting point, the larger the differences will become. Mostly the villagers of two successive villages will understand each other's dialects very well, but the longer the chain, the greater the chance that the dialects on the outer edges of the geographical area may not mutually intelligible. At no point is there a complete break such that geographically adjacent dialects are not mutually intelligible, but the extent to which dialects are intelligible seems to depend on their geographic distance in the continuum perspective.

When the traveller walks in the dialect landscape, would he notice only gradual changes? Or would he notice abrupt changes, i.e. borders? From the view of the 'Chambers & Trudgill traveller', we will 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, he 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, e.g., the Saxon area and enter the Franconian area.

If the dialect landscape is a perfect continuum, the traveller will never notice that dialects are the same, nor that there are abrupt changes, but the extent to which dialects change could be predicted by geographic distance. The more remote that the traveller is from a starting point, the more differences accumulate. So Chambers et al. (1998:5-7) describe the distances as being ‘cumulative’. This could also be seen in the Rhenish Fan (Bloomfield 1933). It falls along German-Romance language border from the ‘Schelde’ (North West) to the ’Elzas’ (Southeast), in which 30 parallel isoglosses can be found. When the traveller travels from the first to the thirtieth isogloss, he will find that differences are cumulative.

1.3 This paper

In this paper we study the concept dialect area and dialect continuum, using Levenshtein distance. The metric is explained in section 3. In order to focus on Chambers & Trudgill’s puzzle we use 27 dialects which lies on a straight line. Using Levenshtein distances we calculate linguistic distances between all pairs of these dialects.

In section 4 we research the relation between phonological distances and geographic distances. Using regression we can find how much variation can be explained by geographic distance.

In section 5 we show how dialect areas can nonetheless be identified. Like the arrow method, a distance greater than a certain threshold indicates a border. Clustering show the dialect areas which are implied by the linguistic distances. Also using clustering it will appear 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, namely that 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 dialect continuum. In section 6 we show the relativity of the term ’border’. Further we show that distances are not completely cumulative and visualise the 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.

2 Dialect data

The data used for comparing dialects comes from the 'Reeks Nederlands(ch)e Dialectatlassen' (RND), which was compiled by Blancquaert & Peé (1925–1982). From these atlases we chose 27 sites. The 27 sites form roughly a straight line from the Northeast to the Southwest in the Dutch language area (see Fig. 1). In the RND for each dialect the same 141 sentences are recorded and transcribed in phonetic script. From these sentences we chose 125 words, which we think are representative of the range of sounds in the varieties. For this word list, usually one form is given. Sometimes more than one form is given. It is not clear to what extent this may be due to different social status. It seems that the RND interviewers did not consciously distinguish social status. In Table 1 more information about the informants can be found. In the table the periods of recording are taken from Wijngaard & Belemans (1997). Nunspeet, Putten, Amersfoort and Driebergen are located in the transition zone between the Saxon and Franconian area. Nunspeet and Putten are recorded in the period 1950–1970, while Amersfoort and Driebergen are recorded in 1950–1962. The recordings of Nunspeet and Putten are not made by the same person but the recordings of Amersfoort and Driebergen were.

— Fig. 1 —
— Table 1 —

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

— Fig. 2 —
— Fig. 3 —

— Fig. 4 —
— Fig. 5 —

3 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 & van de Vis (1996) who extended the application of this technique to Dutch dialects, similarly 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 costs half that of substitutions. The principle can be illustrated by a small example. In Standard American 'saw a girl' is pronounced as [sɔ:əgIrl]. In the dialect of Boston it is pronounced as [sɔ:rəgøl]. Now we can change the the first pronunciation into the second as follows:

sɔəgIrl	delete r	1
sɔəgIl	replace I/ø	2
sɔəgøl	insert r	1
sɔrəgøl		
		4

This example has been simplified in order to clarify the fundamental idea. It is crude to treat segments as alike or different simpliciter. We shall present refinements below.

In fact many sequence operations map [sɔ:əgIrl] into [sɔ:rəgøl]. In the worst case, first we delete all sounds of the first pronunciation (7 deletions), and next we insert all sounds of the second pronunciation (7 insertions). Then we get a total cost of 14. However, there is an algorithm which 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: non-identical phones contribute to phonological distance, 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 the pair [a,p] counts as much more different than [b,p]. The measurements we employ below are sensitive to segmental similarity in exactly this way.

We experimented with two systems to guard against special dependency. The one is developed by Hoppenbrouwers & Hoppenbrouwers (1988) and further described in Hoppenbrouwers & Hoppenbrouwers (1993) and Hoppenbrouwers (1994). The other is constructed by Vieregge, Rietveld & Jansen (1984). Hoppenbrouwers' system is based on Chomsky and Halle's *Sound Pattern of English* and consists of 21 binary features which apply to all phones (vowels and consonants). Vieregge's system consist of 4 multi-valued features only for vowels, and 10 multi-valued 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'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 in this paper are made on the basis of the system of Vieregge, and details are given in Nerbonne & Heeringa (1998).

Assume 125 words are transcribed from two different dialects. Then for 125 word pairs the Levenshtein distance can be calculated. Now the total distance between the dialects is equal to the sum of the 125 Levenshtein distances. In this paper, we call this total distance *linguistic distance* or *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 that interesting, perhaps related techniques could be developed for other linguistic levels.

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

4 Linguistic vs. geographic distances

If a dialect landscape is a perfect continuum in which borders exist (more or less sharp), linguistic distances completely depend on geographic distances. To find the extent to which linguistic and geographic distances are related to each other, we correlate them (4.1) and perform regression analyses (4.2).

As the basis for the following subsections we calculated the geographic distances on the basis of coordinates given by Map Blast, a mapping program which can be found on the World Wide Web via 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 will give some distortion, because it ignores the Earth's curvature, but since the area is small, the distortion will be minimal.

4.1 Correlation

When calculating the correlation coefficient between the phonological distances and the geographic distances, this turned out to be equal to r=0.8054, which is highly significant.[2] This means that ($r^2 \times 100 =$) 65% 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. So for 125 words we get 125 distance matrices. We correlated each of them to the geographic distances. The word groen (English: 'green') has the highest correlation: 0.6842 which is significant. The word zoon (English: 'son') has the lowest correlation: -0.0885. The mean of all separate word correlation values is 0.3372 with a standard deviation of 0.1784.

Note that the highest correlation with distances on the basis of one word separately (0.6842) is lower than the correlation with distances which are equal to the mean of 125 word distances (0.8054). It is normally the case that averages show higher correlations than component scores, although theoretically this need not be the case. We choose 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 a combination of 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 normally have a higher correlation.

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

[2] In a larger study of 350 Dutch varieties, we obtain a pronunciation-geography correlation of r=0.6555.

4.2 Explaining Linguistic Distances using Geography

We use regression to fit the relation between phonology and geography into a formula. The formula may represent a linear relation, but also several more complex relations are possible. Using SPSS we found that the 'logarithmic regression line' represents the relation between the phonological and geographic distances fairly well. This is visualised in Fig. 6. The 'logarithmic' correlation coefficient is equal to 0.899, which means that 80.8 % 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 only matters more <u>that</u> they are far away, and less <u>how</u> far away they are.

Once the relation is fixed in a formula, the expected phonological distances could be calculated on the basis of the geographic distances. The dialects of our dataset lie on the straight line. Now we are especially interested in distances between two geographically successive dialects. In Fig. 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, we get residues, differences between actual and predicted values. If a residue is positive, the distance between two successive points is greater then we should 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 then we should expect on the basis of their geographic distance. In order to focus on significant values, we transform 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 whether they are adjacent). Next we calculated z values (differences expressed as numbers of standard deviations) and found 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 and Bellegem was significantly higher than one should expect on the basis of their geographic distance. This is visualised in Fig. 8.

Looking at the residues, we see that phonological distances mostly can be explained by geographic distance. This justifies the continuum perspective, which is investigated further in section 6. 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. We investigate this further in section 5.

— Fig. 6 —
— Fig. 7 —
— Fig. 8 —

5 Dialect Areas

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

5.1 Arrow method

In traditional dialectology, researchers sought dialect areas, trying to find borders which separate one area from another. As an example we mention the dialect map which can be found in Daan & Blok (1969). For the Netherlandic part of the Dutch language area Daan used the arrow method to find dialect borders. Dialects which 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 (nearly) the same, and when not. When the 'Chambers & Trudgill traveller' travels from Northeast to Southwest, when will he 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 reifying what speakers do by the arrow method. When the Levenshtein distance between two dialects exceeds some threshold, we might hypothesise that these dialects are separated by a dialect border.

How do we fix the threshold? In section 4.2 we described how to split up phonological distances into a geographic component and a residual part. Converting the residues to z values, the chance could be calculated that the residual distance between two dialects is equal to or greater than the distance which was found. When the chance is lower than a reasonable value α, than the residue represents a significant deviation from the distance which would be expected on the basis of the geographic distance. The α value is the threshold. We use α=0.05.

Looking at Fig. 8 we see that the Saxon dialects (numbers 1-12) and the Franconian dialects (numbers 14-27) are separated by two borders, namely between 12 and 13 (Putten and Amersfoort), and 13 and 14 (Amersfoort and Driebergen). The p values were respectively 0.0294 and 0.0262. So the distinction between both areas is very clear. Between 18 and 19 (Oudenbosch and Roosendaal) we also found a border. The p value is 0.0409. Furthermore there are 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 0.0102, 0.0049 and 0.0057. Possibly this can be explained by the fact that Nazareth, Waregem and Zwevegem draw near the Flemish-French border, so they belong to the French-Flemish transition zone.

5.2 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.

After calculating the distances between the 27 dialects we get a 27 × 27 matrix. On the basis of that matrix we 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 which gave rise to it. To iterate, we have to assign a distance from the newly formed cluster to all other points. For example if point A and point B are fused to one cluster AB, the distance between a point S and cluster AB could be the average of the distance between S and A and the distance between S and B. Besides the average, there are several more alternatives (Jain et al. 1988). From them, the Ward's method turned out to be most suitable for our research. Ward's method is very similar to using the average, but it minimises squared error (Jain et al. 1988). Note that the method is always forced to find groups.

The result may be seen in Fig. 9. As the two most significant groups, a Saxon group and a Franconian group emerge. A border could be drawn 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 distances, while it is a characteristic of a continuum and a fundamental dialectological assumption that dialect distances are related to geographic distances fairly directly.

— Fig. 9 —

6 Dialect Continuum

In this section we undertake closer investigation of the dialect continuum. We will show that there may be parallel isoglosses, which shows the relativity of the term 'border' given traditionally methodology. Next we show that distances are not completely cumulative. We will also try to visualise 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 great.

In this paper 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.

6.1 Parallel Isoglosses

As we found in section 5.1, 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 look at Fig. 2, we see in most Saxon dialects [døːr] is followed by the [ə], while for the

Franconian dialects this is never the case. So we see an isogloss between Nunspeet and Putten. If we look at Fig. 4, we see in most Saxon dialects a variant of [bIn] is used, while in the Franconian dialects a variant of [zɛˑⁱn] is used. So there is an isogloss between Putten and Amersfoort. In Fig. 5 we see that in all Saxon dialects the vowel in [ʋɛˑⁱn] is a [i], while in the Franconian dialects always another vowel is used. 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 simply sharp borders could be found by looking for coinciding isoglosses. Rather one should speak of transition zones. In the continuum view, this fact could be taken into account in a suitable way.

6.2 Cumulative distances

A property of geographic distances is that they are simply cumulative. Assume three points A, B and C which lay on a straight line. Then 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:

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

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 take the site at the outer end of the line in the Northeast or Southwest. Next could draw a scatter-plot, 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 is like an assumption that the 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 explore the comparison by drawing scatter-plots in which indirect phonological distances are plotted as a function of direct phonological distances. The plot is shown in Fig. 10. In contrast to the geographic distances, the plots do not show a linear curve. So phonological distances are not simply cumulative.

— Fig. 10 —

6.3 The Shape of Continua

In section 4 we showed that phonological and geographic distances are related. In this section we will try to understand the relation more deeply.

In section 1 we cited Chambers et al. (1998): "If we travel from village to village, in a particular direction, we notice linguistic differences which distinguish one village from another" (p. 5). This suggests a novel perspective on linguistic variation. Rather then viewing the phonological distance from Scheemda to Bellegem directly, we adopt the traveller's perspective – one who notices incrementally the differences from Scheemda to Veendam, and from Veendam to Eext, et cetera. The Chambers-Trudgill traveller develops a notion of indirect phonological distance – the sum of distances from pairwise neighbouring points on a connected line. The question is then: what would this traveller's view of the dialectal landscape be? Fig. 11 shows that the view is a linear relationship between geographic distance and this traveller's sum of incremental distances. This is the indirect phonological distance.

Of course the Chambers-Trudgill traveller's view is misleading! Phonological distances do not sum along (geographic) paths. To examine the real relationship, we also draw a scatter-plot with direct geographic distance versus direct phonological distances, which reckon with the fact that phonological distances are not simply cumulative. The result can be found in Fig. 12. The relation is obviously not linear. The graph's slope clearly decreases as a function of distance. The distances of the dialects about the middle of the line varies at least with respect to the starting point. The relatively flat sections of the curve correspond to relatively homogeneous linguistic areas.

Why does the discrepancy arise between the indirect, traveller's view and the true pronunciation difference? We suggest that this arises because the traveller is reacting to his global (aggregate) impression of the (pairwise) differences. As Fig. 11 demonstrates, these accumulate in a linear fashion, giving the traveller the impression that the continuum is simple and dialectologically real. But a brief thought experiment demonstrates how fallacious this is. We can easily imagine a line in which two dialects alternate, first A, then B, then A again, etc. 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 which may be lost again over a longer distance.

— Fig. 11 —
— Fig. 12 —

The contrast between Fig. 11 and Fig. 12 is our analysis of the Chambers-Trudgill puzzle. The perception of the traveller is that he 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 the traveller's path accumulate linearly, as Fig. 11 shows. But as Fig. 12 shows, the traveller is deceived. The true pronunciation difference simply is not the sum of the pairwise differences along the path. In the following section we develop this contrast further.

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

— Fig. 13 —
— Fig. 14 —

These figures reinforce the earlier point made about the Chambers-Trudgill traveller. The cumulative view (Fig. 13) is simplistic, ignoring the fact that local changes may be undone. If Fig. 14 appears chaotic, perhaps that is an admonition that we ought to be focused on aggregates, not individual words, as we study variational linguistics.

6.4 Multidimensional scaling

On the basis of geographic coordinates the distances between locations can be determined. The reverse is also possible: on the basis of the mutual phonological distances, an optimal coordinate system can be determined with the coordinates of the locations in it. The latter is realised by a technique known as 'multidimensional scaling' (MDS). In a multidimensional scaling plot, strongly related dialects are close to each other, while strongly different dialects are located far away from each other (Kruskal et al. 1984).

As input each dialect is defined as a range of distances, namely the distance to itself and the distances to other dialects. The distances correspond to dimensions. If we have 27 variants, we get 27 dimensions. With multidimensional scaling, the dimensions can be reduced to one, two, or more dimensions, so we get 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 scale the 27 dimensions to two dimensions (see Fig. 15). Although similarities with the geographic map can be identified, the plot does not show a straight line as on the geographic map. In the plot clearly three groups could be distinguished, namely a Saxon group, a Dutch Franconian group and a Flemish Franconian group. Comparing 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 points again to the necessity of the dialect continuum perspective.

In the MDS plot Saxon and Flemish Franconian are more closely related than the geographic line suggests. This can be explained (maybe among other things) 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]. Fig. 3 illustrates this.

We try to determine whether the data is uni- or multidimensional. Therefore we scaled the data not only to 2 dimensions, but also to 1 and 3 to 8 dimensions. For each number of dimensions we calculate a the squared correlation (r-squared, abbreviated as RSQ) between the given dialect distances and the corresponding distances on the multidimensional scaling plot. RSQ can be interpreted as the proportion of variance of the distance that is accounted for the distances between the points on the multidimensional scaling plot. In Fig. 16 for each number of dimensions the RSQ value is given. Here we see a clear difference between the RSQ value for one dimension on the one side, and the RSQ values for two or more dimensions on the other side. This suggests that there are at least two dimensions, perhaps three. The fourth and further dimensions explain very little of the variance in the data.

— Fig. 15 —
— Fig. 16 —

7 Conclusions

There is a strong correlation between phonological distance and the logarithm of geographic distance (0.9), accounting for 81% of the variation in pronunciation. The correlation per word in our sample varies greatly (from -0.0885 to 0.6842 using a linear model). Using regression, we could see how phonological distances depends on geographic distances. We also argued that the logarithmic model was reasonable, proposing that distances further away are less significant than local distances. In the linear model, the correlation is also strong (0.8).

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, then we conclude 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 & Kleiweg (2001) show that the dialectometric method used here is validated by the expert opinion on Dutch dialect areas.

In the Rhenish Fan isoglosses are parallel. This is also the case between the Saxon and Franconian area. When regarding the dialect landscape as a continuum, we can deal with 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 all intermediate distances, each distance corresponding with 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, so they are not completely cumulative. This constitutes our perspective on the puzzle of the Chambers-Trudgill traveller: the traveller perceives phonological distance indirectly, and is therefore inclined to overestimate the real degree of change.

The relation between geographic distances and direct phonological distances can be visualised as a continuum in a two-dimensional plot. Since phonological distances are not simply cumulative, we obtain a relation which looks like 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 could explain a great deal of variation clearly suggests that a view of the dialectal landscape as a continuum should assume multidimensional determinants of phonological distance. Geographic distance explains a great deal, but not everything. In the plot clearly three groups could be distinguished, namely a Saxon group, a Dutch Franconian group and a Flemish Franconian group. Putten lays exactly between Saxon and Dutch Franconian, and Clinge is lays exactly between Dutch Franconian and Flemish Franconian. This shows the need of the continuum view. In the plot Saxon and Flemish Franconian are more related than the geographic line suggest. This can be explained (perhaps among other things) 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 syllable is reduced to [ə]. When searching for significant dimensions, we noted that there are at least two dimensions, and that the fourth and further dimensions explain very little.

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

8 Further work

In this study only four varieties lay in the transition zone between Saxon and Franconian. It may 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 from Saxon to Franconian. A line from Frisia to the Franconian area is not possible since the line would pass trough a great deal of water for which no dialect data is available.

We are aware of the fact that the continuum we studied is a flat area. It may 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 lay on a line. It would also be interesting to research the continuum as it is, namely in two dimensions. If visualising a two-dimensional continuum, the graph should be three dimensional, like a mountain landscape, where the height represent the phonological distance with respect to e.g. standard Dutch. In this larger set we plan to examine the degree to which areas can explain linguistic distances.

It would also be interesting to explore our dialectal areas, particular those with well-known divergent factors such as national borders.

References

Blancquaert, E. & W. Peé (1925–1982). *Reeks Nederlands(ch)e Dialectatlassen.* De Sikkel. Antwerpen.

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

Chambers, J. K. & Peter Trudgill (1998). *Dialectology.* 2nd edn. Cambridge University Press. Cambridge.

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

Heeringa, W., J. Nerbonne & P. Kleiweg (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'. University of Passau, Springer. Heidelberg. Accepted.

Hoppenbrouwers, C. (1994). 'De indeling van de zuidoostelijke steektalen'. *TABU: Bulletin voor taalwetenschap* **24**(2): 37–63.

Hoppenbrouwers, C. & G. Hoppenbrouwers (1988). 'De featurefrequentiemethode en de classificatie van nederlandse dialecten'. *TABU: Bulletin voor taalwetenschap* **18**(2): 51–92.

Hoppenbrouwers, C. & G. Hoppenbrouwers (1993). 'De indeling van noordoostelijke dialecten'. *TABU: Bulletin voor taalwetenschap* **23**(4): 193–217.

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

Kessler, B. (1995). Computational Dialectology in Irish Gaelic. In 'Proceedings of the European Association for Computational Linguistics'. EACL. Dublin. pp. 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 edn. CSLI. Stanford. pp. 1–44. 1st edition appeared in 1983.

Kruskal, J. B. & M. Wish (1984). *Multidimensional Scaling.* Sage Publications. Beverly Hills and London.

Nerbonne, J. & W. Heeringa (1998). 'Computationele Classificatie van Nederlandse Dialecten'. *Taal en Tongval* **50**(2): 164–193. Available as: http://www.let.rug.nl/˜nerbonne/papers/tetv99.ps.

Nerbonne, J., W. Heeringa, E. van den Hout, P. van der Kooi, S. Otten & W. van de Vis (1996). Phonetic Distance between Dutch Dialects. In G.Durieux, W.Daelemans & S.Gillis, eds, 'CLIN VI, Papers from the sixth CLIN meeting'. University of Antwerp, Center for Dutch Language and Speech. Antwerp. pp. 185–202. Available as: http://www.let.rug.nl/˜nerbonne/papers/clin96.ps.

Nerbonne, J., W. Heeringa & P. Kleiweg (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 edn. CSLI. Stanford. pp. v–xv. Available as: http://www.let.rug.nl/˜nerbonne/papers/timewarp.ps.

Tait, Mary (1994). North America. In C.Moseley & R.Asher, eds, 'Atlas of the World's Languages'. Routledge. London and New York. pp. 3–30.

Vieregge, W. H., A. C. M. Rietveld & C. I. E. Jansen (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. pp. 654–659.

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

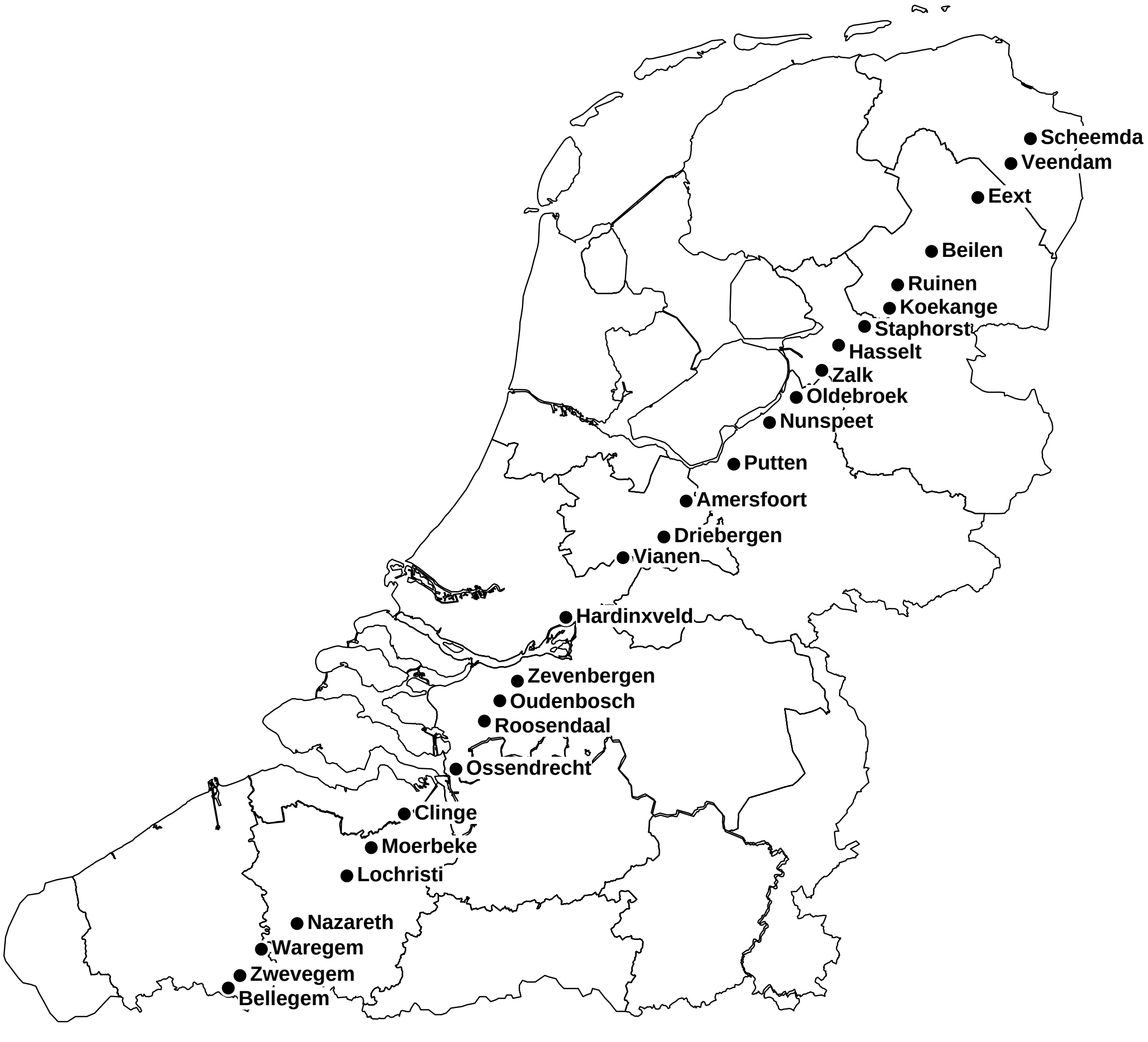

Figure 1: The locations of the 27 Dutch dialects studied.

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

Table 1: Overview of the places we used. The column 'NOW' gives the number of words for which more than one variant was used. The column 'prof.' gives the professions of the informants. Here we distinguished the following categories: a=agricultural, n=non agricultural, s=student, ?=unknown. For housewives the profession of their husband is given. The column 'ages' gives the 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 last year of the recording period. In the column 'period' the recording period is given. The column 'volume' gives the part of the RND in which the dialect could be found.

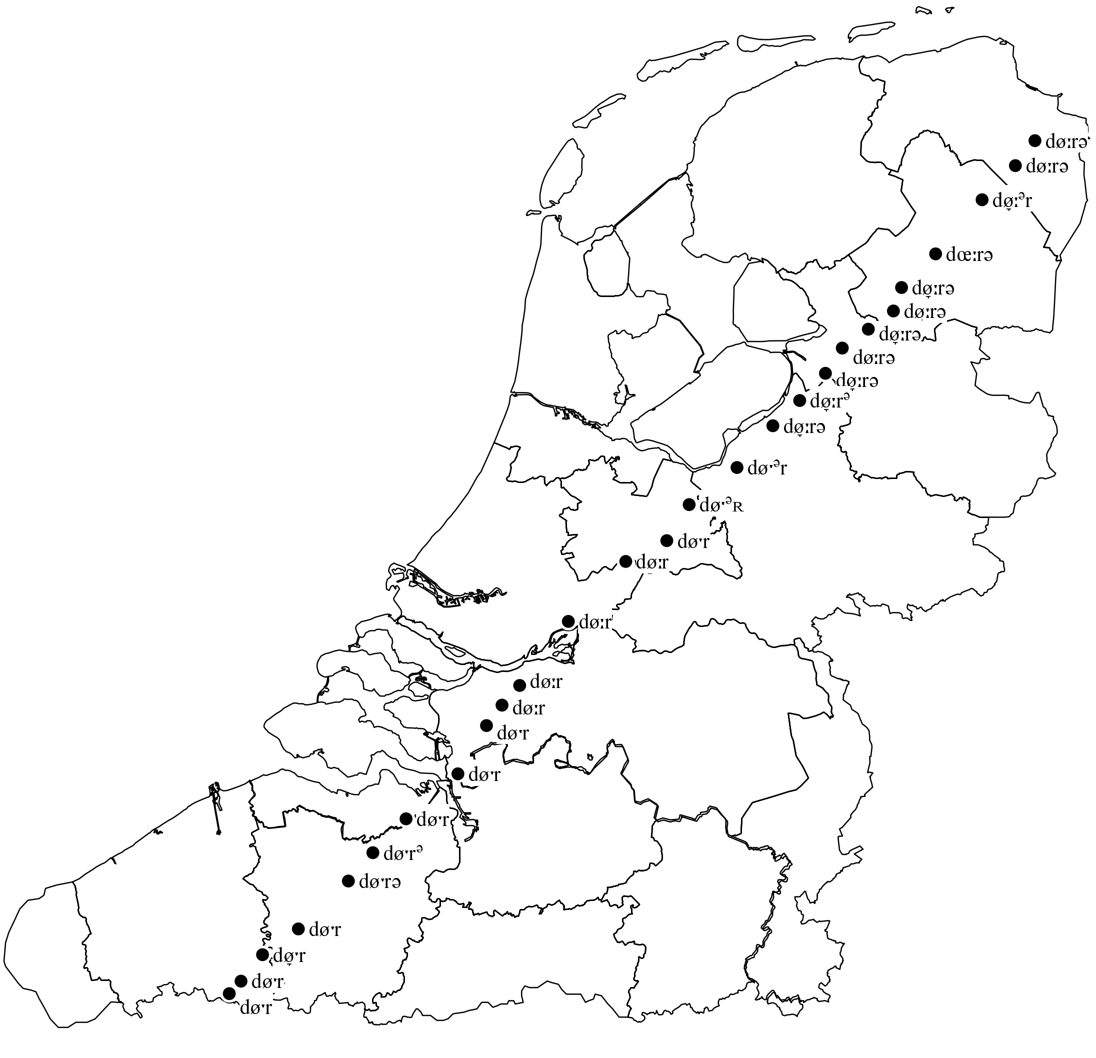

Figure 2: Variation of <u>deur</u> (English ‘door’) in IPA.

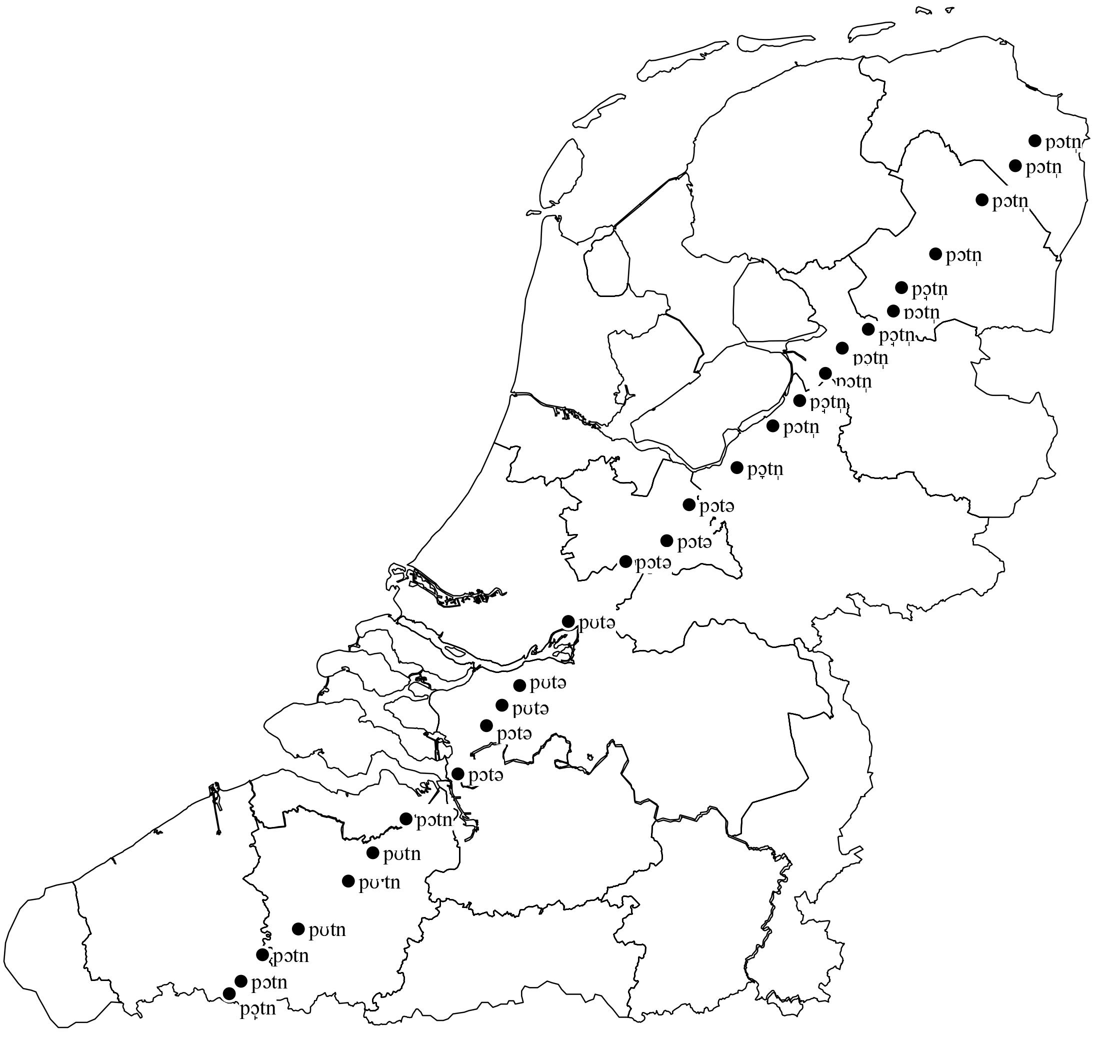

Figure 3: Variation of potten (English ‘pots’) in IPA.

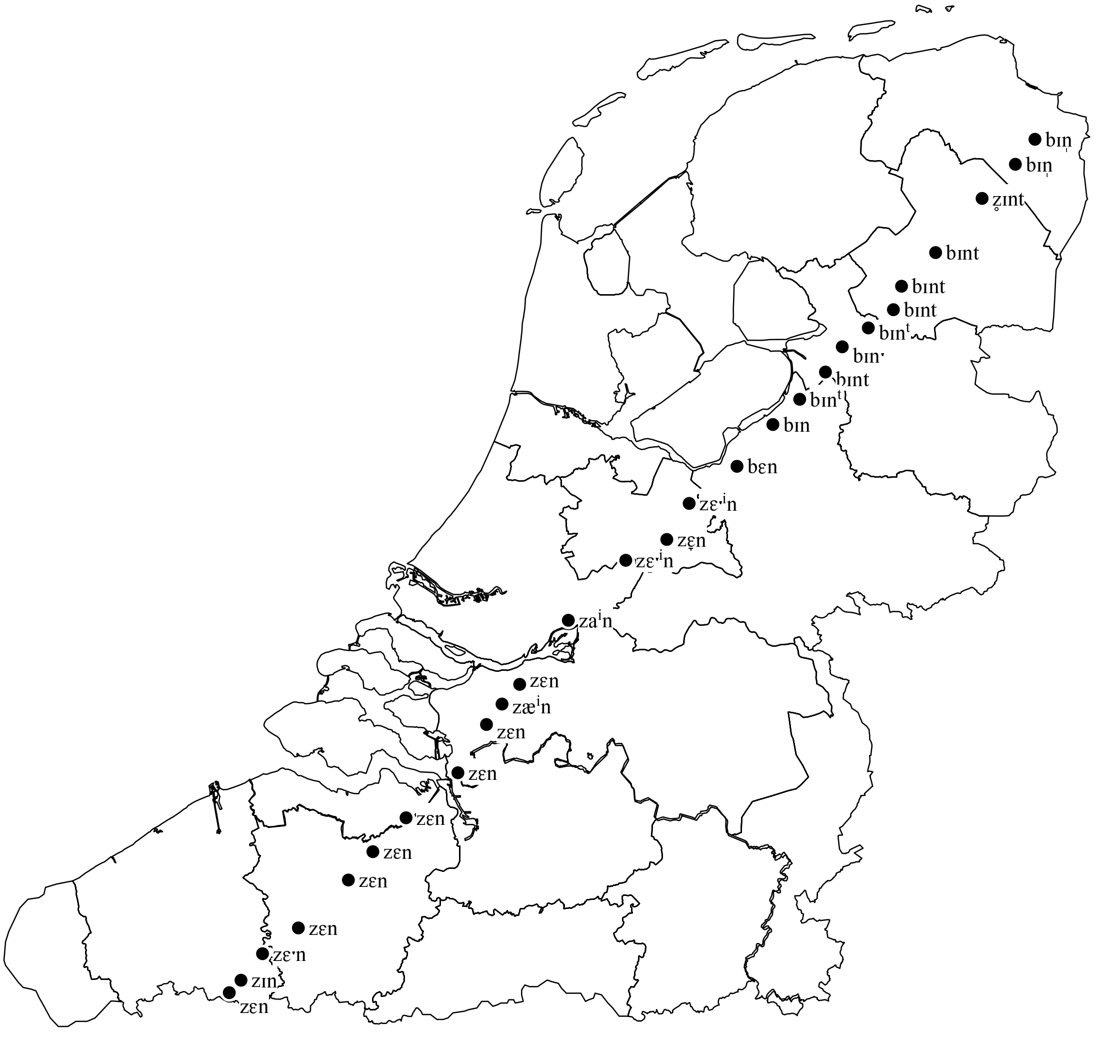

Figure 4: Variation of zijn (English ‘to be’) in IPA.

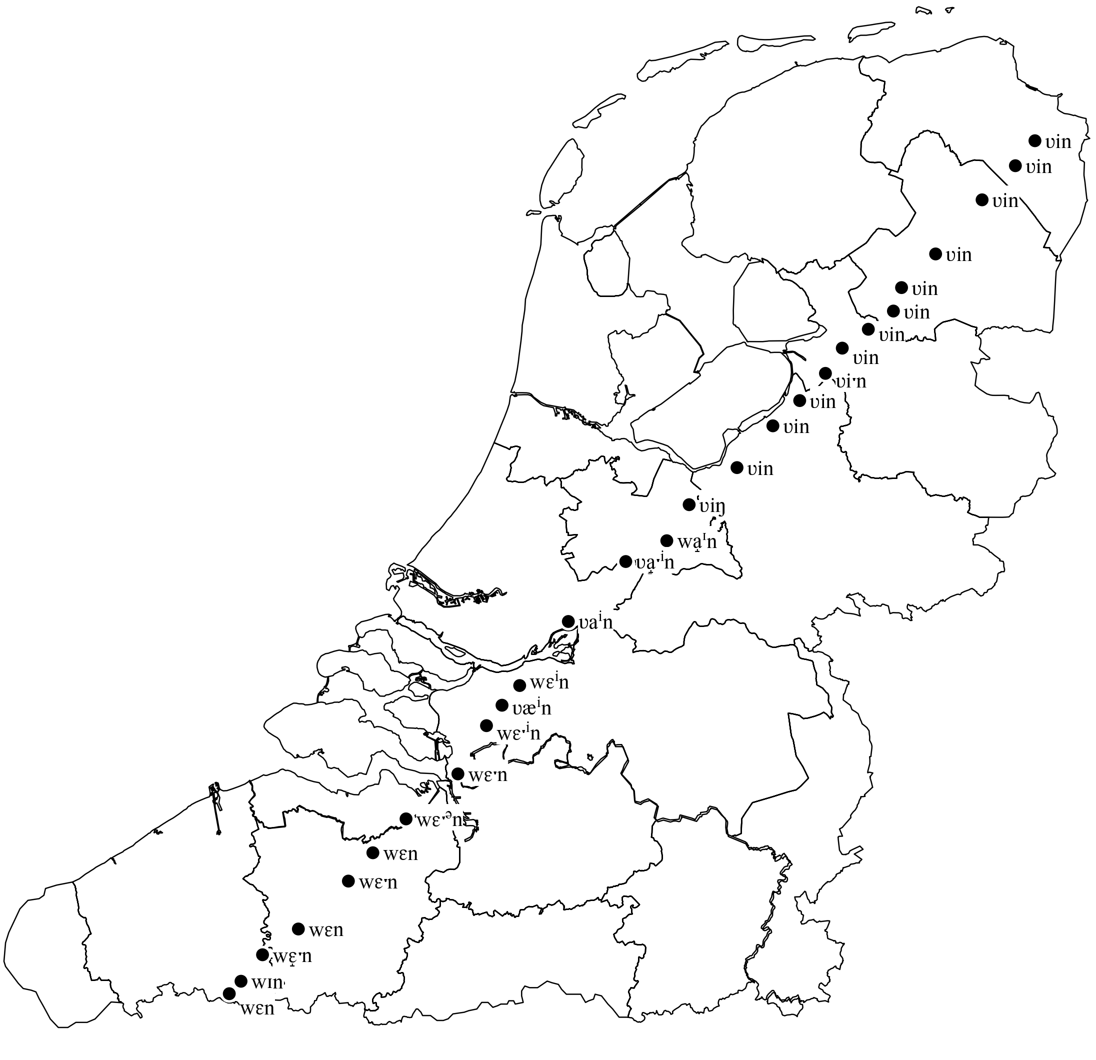

Figure 5: Variation of <u>wijn</u> (English ‘wine’) in IPA.

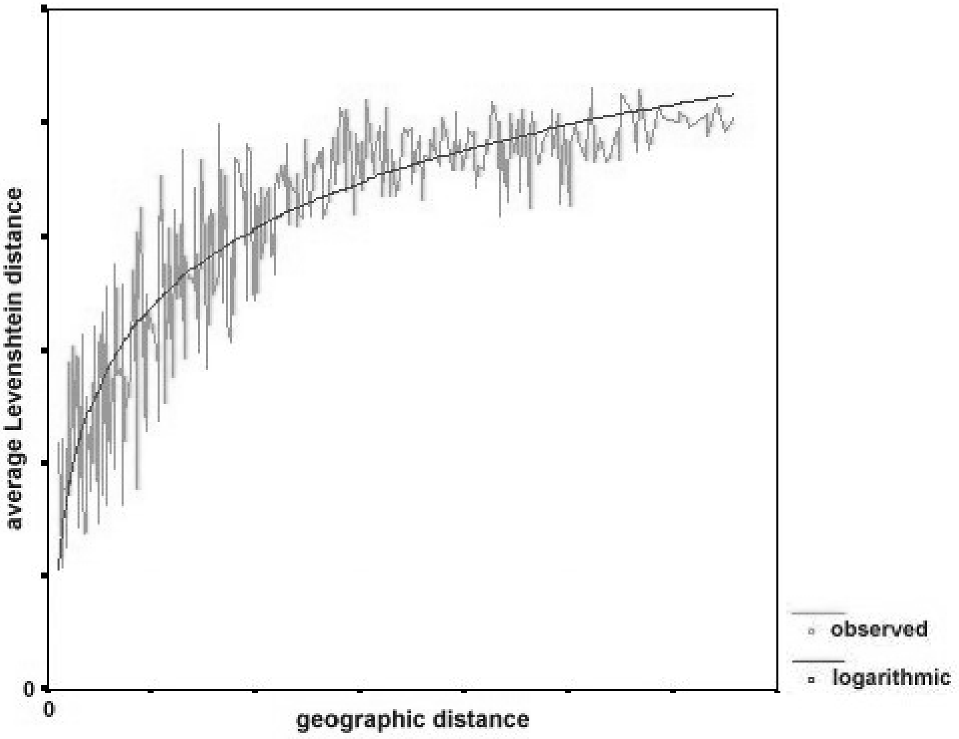

Figure 6: Geographic distances vs. average Levenshtein distances. Two successive points are connected by a straight line, illustrating the range of variation for average Levenshtein (pronunciation) distance. In SPSS the logarithmic regression line was drawn. Note that the logarithmic line seems to overestimate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

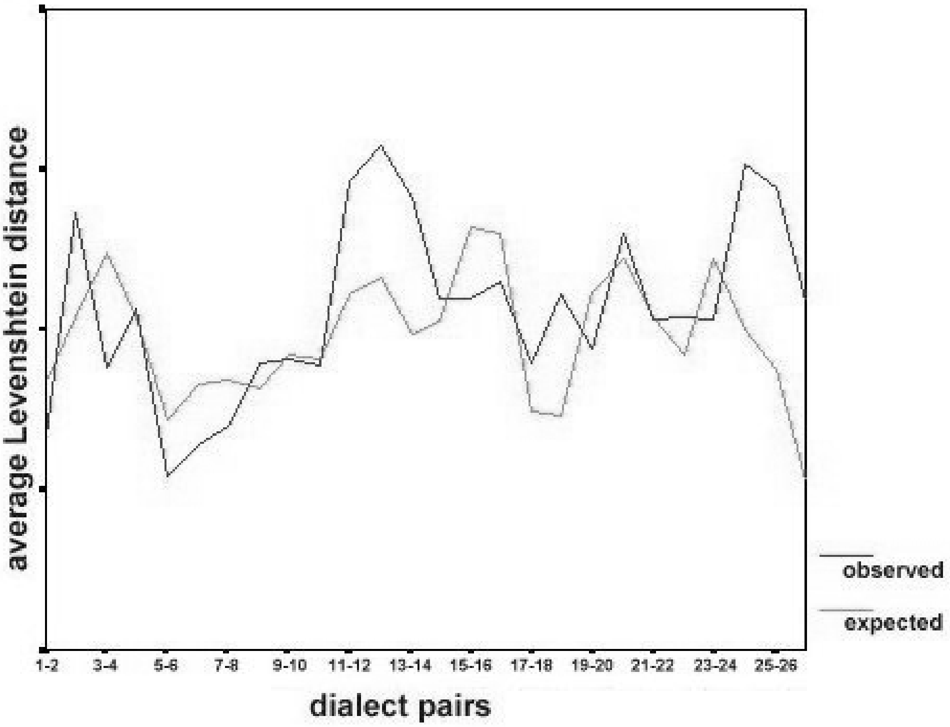

Figure 7: The observed and expected Levenshtein distances between successive dialects on the path of the “Chambers traveller”. The dialects are numbered from Northwest to Southeast in the same order as on the geographic map (Fig. 1). The points at which observed and expected distances differ, greatly suggest themselves as candidates as borders of distinct areas.

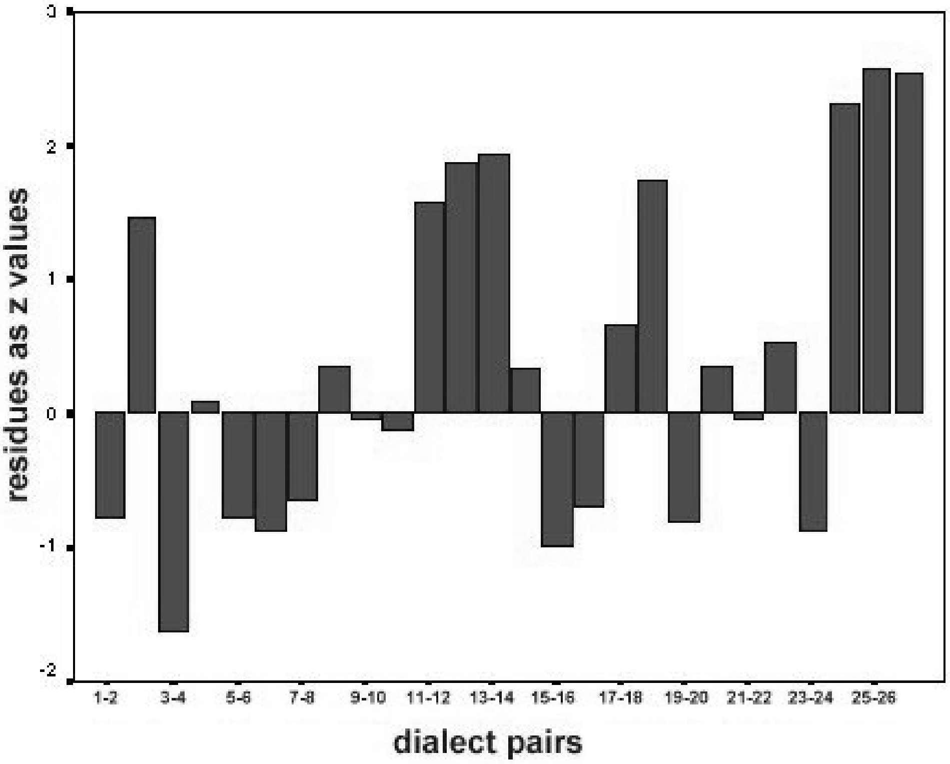

Figure 8: Differences between observed and expected average Levenshtein distances for all pairs of two successive dialects, given as z values (standard deviations). The dialects are numbered from Northwest to Southeast in the same order as on the geographic map (Fig. 1). The large positive residues mark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

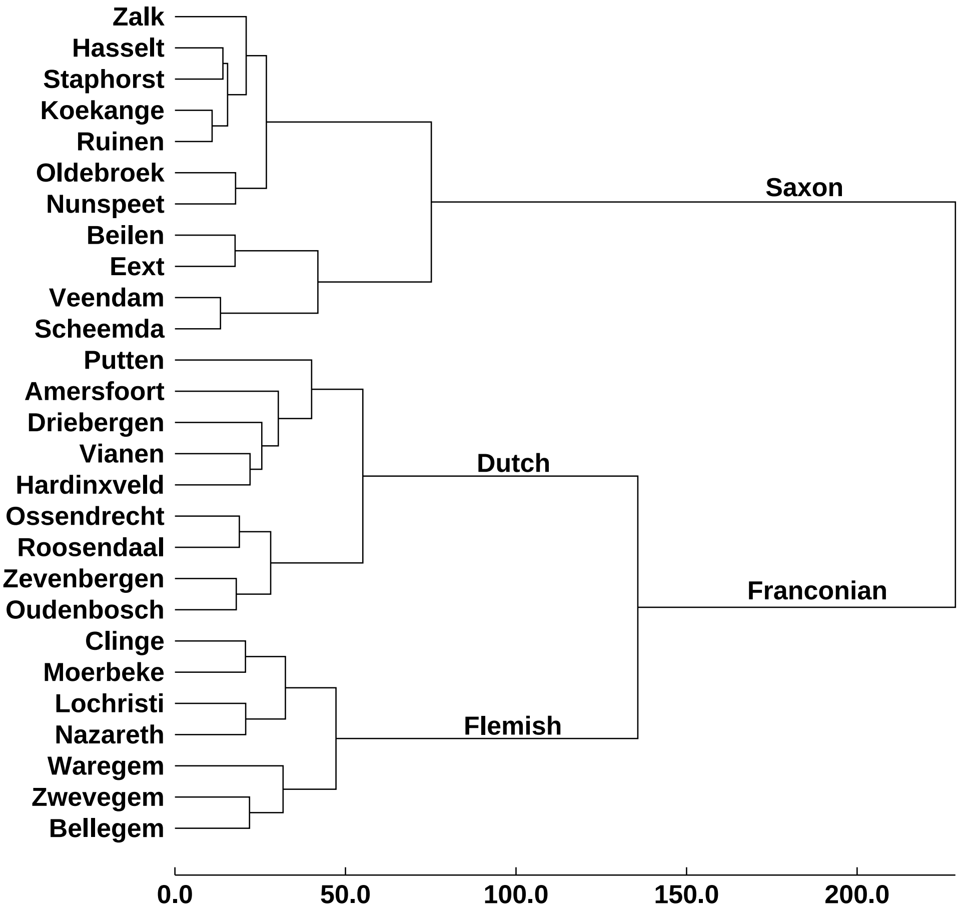

Figure 9: A dendrogram derived from the distance matrix based on Levenshtein distances as measured on 125 words. The feature system of Vieregge is used; diphthongs are represented as one phone; Manhattan distance between feature bundles is calculated. The two main groups are the Saxon (top) and Franconian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

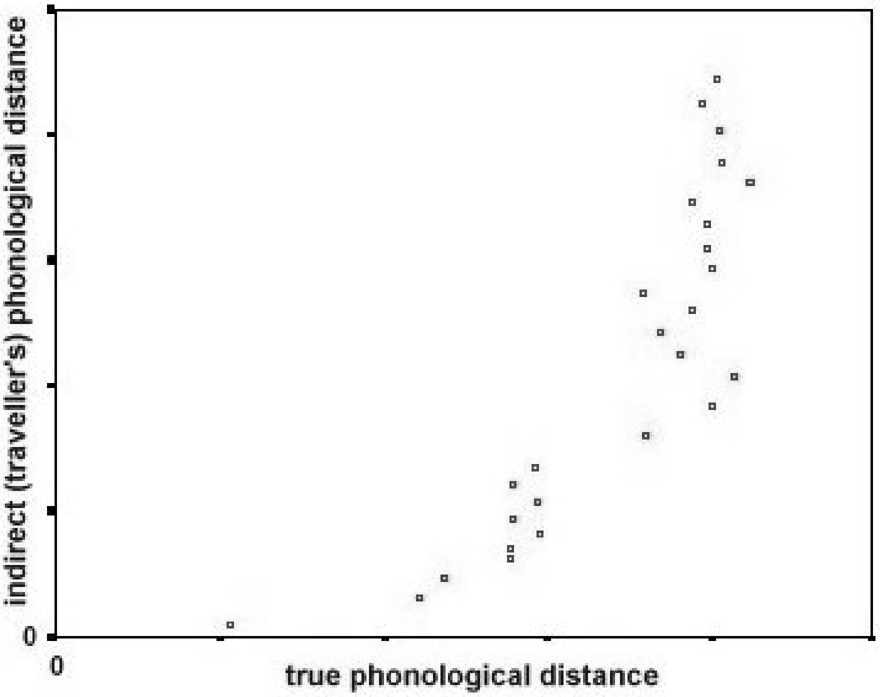

Figure 10: True phonological distance versus indirect ("traveller's") phonological distance with Scheemda as starting point. The form suggests an exponential relation, the mirror of the logarithmic relation we hypothesise exists between geographic and phonological distance.

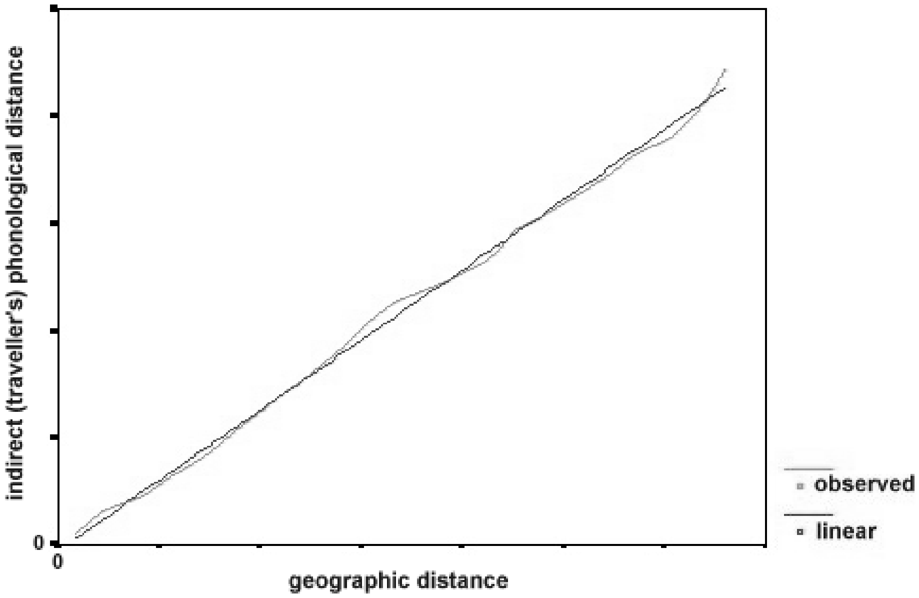

Figure 11: Geographic distance versus mean indirect ("traveller's") phonological distance with Scheemda (Northeast) as starting point. Essentially the same graph results if one begins in Bellegem (Southwest). This graph explains the Chambers-Trudgill's traveller's perception that the dialect landscape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

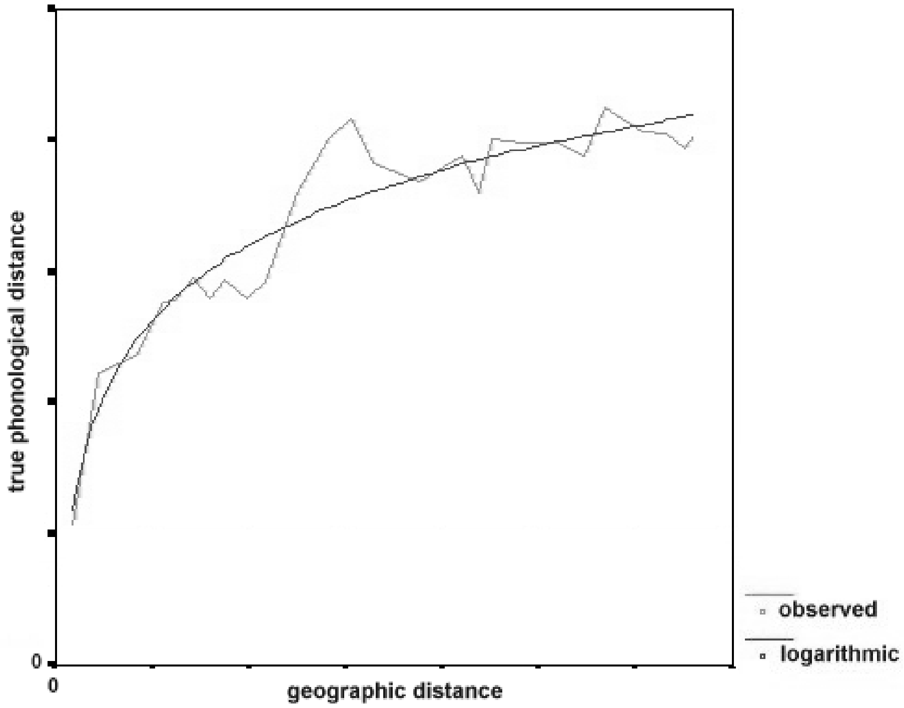

Figure 12: Geographic distance versus mean true phonological distance with Scheemda (Northwest) as starting point. In SPSS the logarithmic regression line was drawn. A similar graph results if one begins at Bellegem (Southwest). This graph illustrates the fallacy in the memory-less traveller's view of 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.

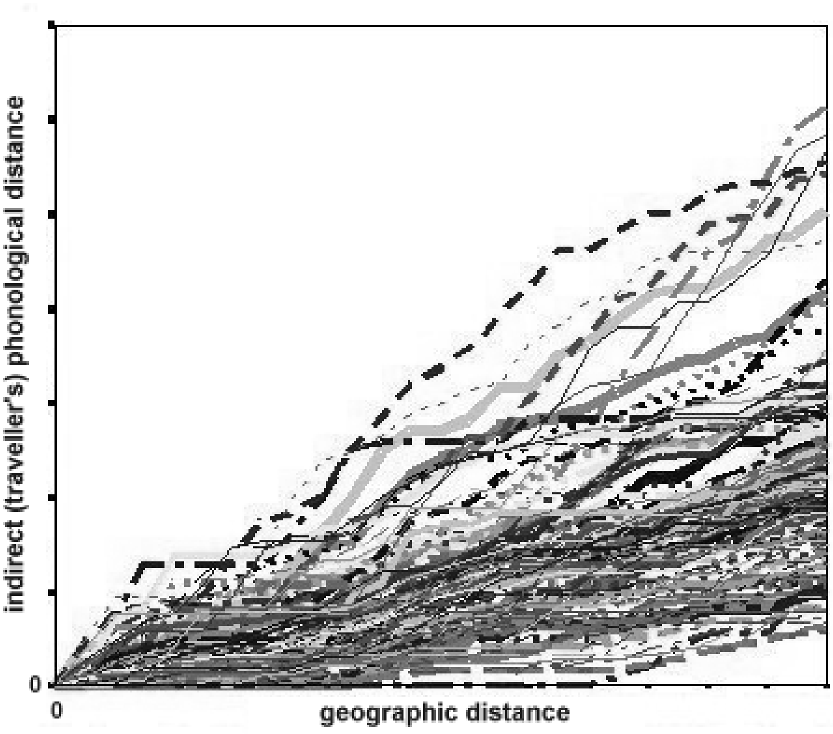

Figure 13: Geographic distances versus indirect ("traveller's") phonological distances for each of the 125 words with Scheemda (Northeast) as starting point.

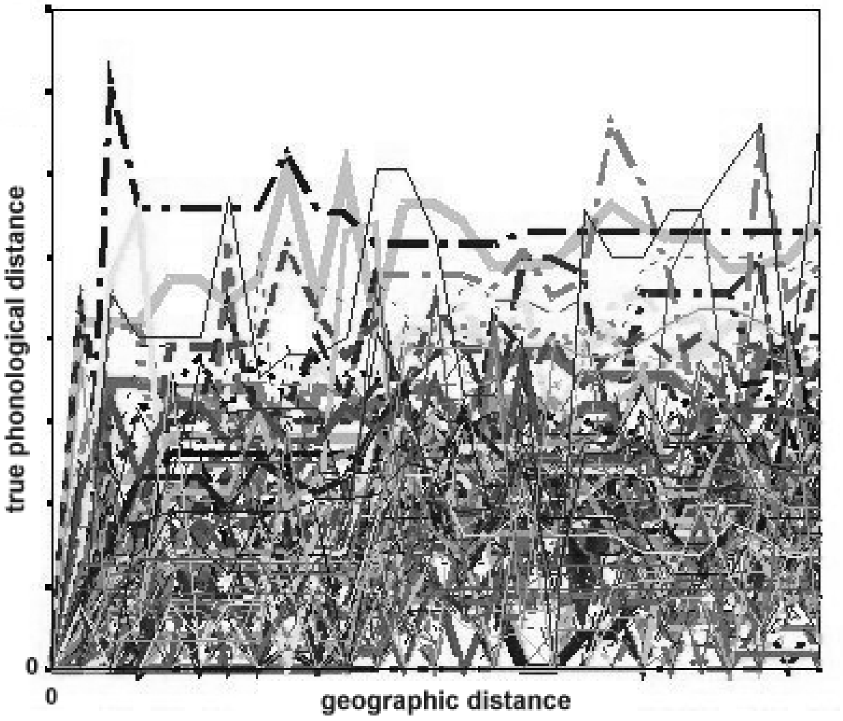

Figure 14: Geographic distances versus mean true phonological distances for each of the 125 words with Scheemda (Northwest) as starting point. This is a metric perspective for bundles of (word) isoglosses. If such bundles existed, we should see a lighter 'V' shape in the graph.

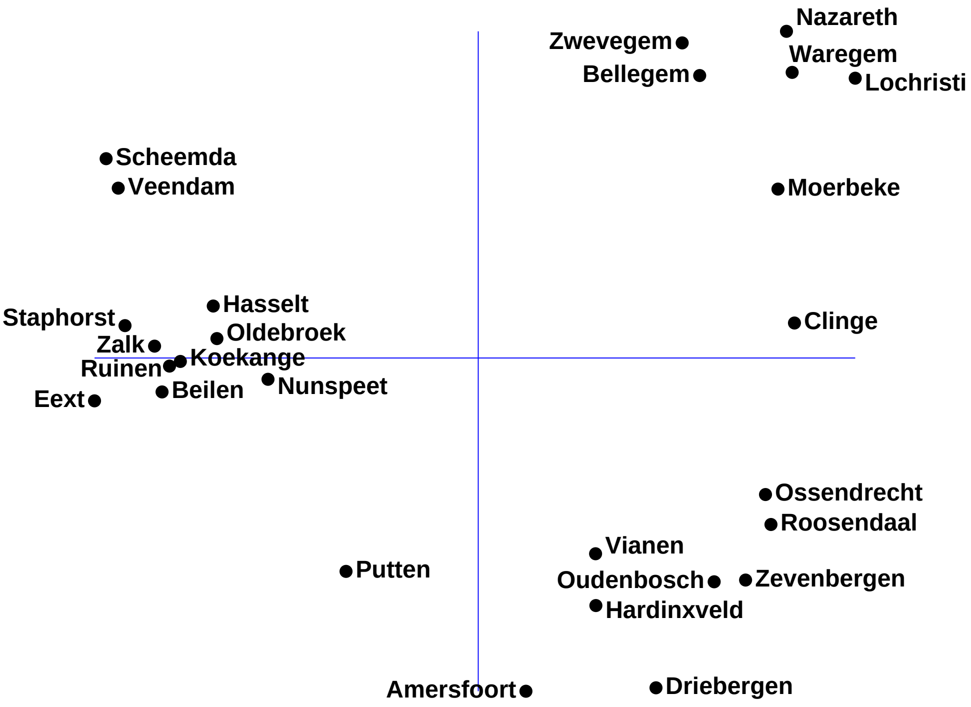

Figure 15: The two most significant dimensions in multidimensional scaling. The x-dimension is more significant than the y-dimension. The three main groups are the Saxon (left), Dutch Franconian (lower right) and Flemish Franconian dialects (upper right). Dialects do not lie on a line as on the geographic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

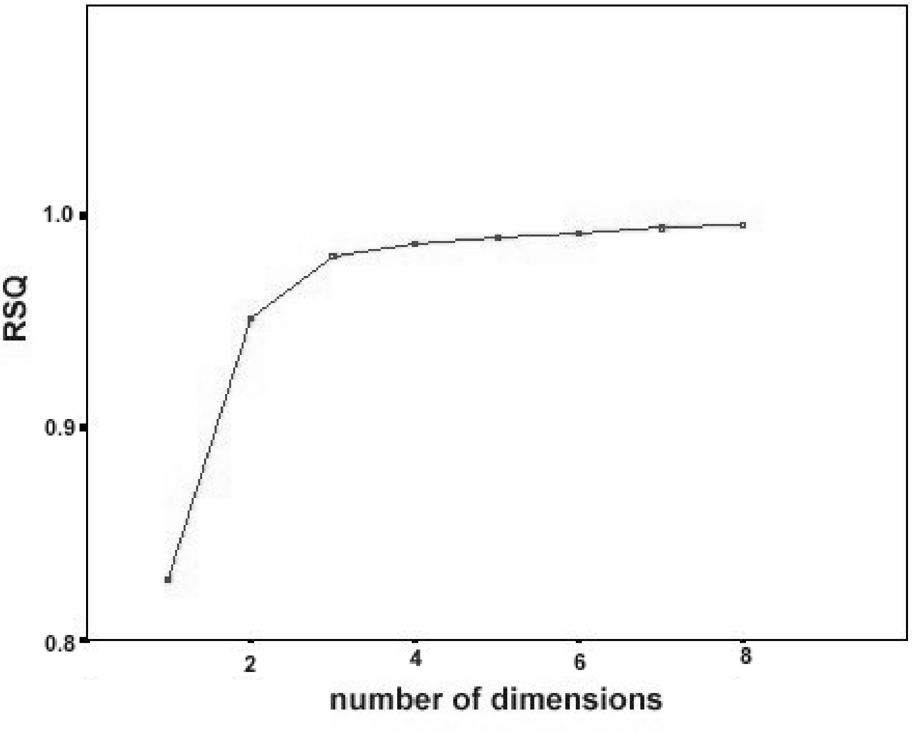

Figure 16: The dialect distances are scaled to 1, 2, 3, 4, 5, 6, 7 and 8 dimensions. For each number of dimensions the r^2 (RSQ) value is given as fit measure, ranging from 0 (perfect fit) to 1 (worst possible fit). This is the correlation of the phonological distances with the distances in the proposed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