Lingua style submission version 2008/07/28-19:02, 40 pages, 8774 words, 50171 characters.

Associations among Linguistic Levels

Marco René Spruit[a], Wilbert Heeringa[b], John Nerbonne[b]

[a] *Meertens Instituut*
Royal Netherlands Academy of Arts and Sciences
Joan Muyskenweg 25
1090 GG Amsterdam
The Netherlands
tel: (+31) 20 4628500,
fax: (+31) 20 4628555
marco.rene.spruit@meertens.knaw.nl
*(*corresponding author*)*

[b] *Center for Language and Cognition Groningen*
Rijksuniversiteit Groningen
9700 AS Groningen
The Netherlands

Abstract

In this paper we measure the degrees of association among aggregate pronunciational, lexical and syntactic differences in 70 Dutch dialect varieties. First, we show that pronunciation is marginally more strongly associated with syntax than it is with lexis and that syntax and lexis are only weakly associated. Then, we check for the influence of geography as an underlying factor because geography is known to strongly correlate with each of the linguistic levels under investigation. We find that pronunciation and syntax are more strongly associated with geography than lexis is. Finally, we refine the results by accounting for the influence of geography as an underlying factor and show that the association between pronunciation and syntax turns out to be largely based on geography. Some influence between pronunciation and syntax remains but the association between pronunciation and lexis is stronger. There is virtually no association between syntax and lexis.

Keywords: dialectometry, microvariation, pronunciation, lexicon, syntax, correlation

Associations among Linguistic Levels*

1. Introduction

The goal of this paper is to contribute to the understanding of the associations among linguistic levels by examining geographical distributions of linguistic microvariation. Investigations of linguistic variation in geographical space can not only illustrate patterns of variation at a certain point in time, but may also reflect residues of linguistic and cultural changes over historical time. This argument effectively interprets synchronic distributions as evidence of diachronic patterns of diffusion (Nerbonne and Heeringa t.a. 2007). We study distributions of linguistic variation in the aggregate to compensate for the noisiness of individual distributions and to examine the data from more general perspectives in which we aggregate over many variables. We conduct this investigation at an aggregate level in order to avoid the choice of a single individual variable, such as the pronunciation of /r/, which risks biasing its results based on the selection. The current study necessarily examines the linguistic levels under investigation on the basis of large collections of numerically interpreted data, because a robust, empirical foundation is required to analyse data from a more general perspective. The adopted quantitative methodology focuses on more general characteristics within and among linguistic levels because individual variables are only taken into account through their relationships with other variables. Metaphorically speaking, the current approach quantifies associations among "the linguistic forests behind the variable trees".

We are now in a position to assess the dialectometrical distances among fairly many sites at three different linguistic levels: pronunciation, lexicon (or vocabulary) and syntax. Pronunciational differences mainly arise from linguistic variation at the phonetic level, but may also include variation at the phonological and morphological levels. We quantify lexical and syntactic differences at a nominal level using a frequency-weighted similarity measure introduced by Goebl (1982) and we measure pronunciational differences numerically using Levenshtein distance (Nerbonne et al. 1999; Heeringa 2004). The novelty of this paper consists first in the opportunity to include syntax among the linguistic levels we analyse, and second, in its attention to potential, mutually structuring elements among the linguistic levels.

We suggest that the associations we attempt to detect are interesting first from a typological point of view, and second, from the point of view of identifying what influences linguistic variation. Addressing the second point first, we note that, although there are many candidate influences which might be affecting how languages vary, including e.g. settlement size, social class, sex, and educational level, only geography has proven its value in large-scale, quantitative studies (Nerbonne and Heeringa t.a. 2007). We proceed here from the common assumption that there are no structural ties between lexical and nonlexical variables. In the present context, this means that we assume there is no linguistic reason to suspect correlation either between the pronunciational and lexical levels or between the syntactic and lexical levels. If we were to demonstrate significant correlations between lexical and nonlexical levels beyond those geography can explain, we would conclude that extralinguistic, non-geographical influences were at work. This should encourage the search for extralinguistic variables, but also suggest how important it might be.

The relation between phonology and syntax is more complicated, since it is easily conceivable that there might be *structural* constraints linking variation at these levels (see

---

* This paper was presented in the special session Comparing Aggregate Syntaxes at the Digital Humanities conference in Paris on July 6, 2006. It is based on joint research by the University of Groningen and the Meertens Instituut in Amsterdam. The Meertens Instituut is the national institute for research and documentation of Dutch language and culture. The Computational Linguistics department at the University of Groningen is known for its attention to quantitative linguistics and dialectometry. For three years now, these two research groups have been collaborating in the Determinants of Dialectal Variation project, NWO number 360-70-120, P.I. J. Nerbonne. More information is available on our project's website at http://dialectometry.net.

below). If phonology and syntax turn out to co-vary beyond the level explained by geography, this might reflect the influence of such structural, typological constraints. Of course, it might just as well reflect the influence of the same variables which account for the correlations between lexical and non-lexical variables, so we shall need to interpret any correlation between phonology and syntax in light of the investigation between the lexical and non-lexical levels.

This paper is structured as follows. Section 2 formulates the two research questions addressed in this work. Section 3 describes the two data sources. Section 4 explains the two measurement procedures used to quantify linguistic differences. Section 5 presents colour maps of the Dutch dialect areas based on pronunciational, lexical and syntactic differences to provide a visual indication of the degrees of association. Section 6 analyses our distance measurements with respect to consistency to ensure that the results are reliable. Section 7 lists the exact degrees of association between pronunciation, lexis and syntax. Section 8 provides the degrees of association between geography and the linguistic levels under investigation. Section 9 refines the results in Section 7 by accounting for the influence of geography as an underlying, third factor. Section 10 recapitulates the main results. The paper concludes with a discussion and directions for future research in Section 11.

2. Research questions

While most linguists would predict that vocabulary is more volatile than pronunciation and syntax and might predict that lexical choice should show little association with other linguistic levels, there have been predictions linking pronunciational with syntactic properties (Donegan and Stampe 1983). Both pronunciation and syntax are highly structured systems, within which a single linguistic parameter might lead to a multitude of concrete and measurable effects.

We address two research questions in the present paper, the first of which is fairly straightforward:

1. To what degree are aggregate pronunciational, lexical and syntactic distances associated with one another when measured among varieties of a single language? Particularly, are syntax and pronunciation more strongly associated with one another than either (taken separately) is associated with lexical distance?

To answer the questions above, it is sufficient to calculate correlation coefficients among the distance measurements for the three linguistic levels. This is a reasonable measure of the degree to which the three linguistic levels are associated.

However, it would be a mistake to interpret any such correlation as influence without checking for the influence of a third factor, especially since geography has already independently been shown to strongly correlate with each of the linguistic levels under investigation (Heeringa and Nerbonne 2001; Cavalli-Sforza and Wang 1986; Spruit 2006). Therefore, it is quite plausible that geography could influence each of the levels separately, leading to the impression of structural influence between them. We suggest that this should be regarded as a null hypothesis, i.e. that there is no influence among the various linguistic levels. This leads to the second research question we address in this paper:

2. Is there evidence for influence among the linguistic levels, even once we control for the effect of geography? Particularly, do syntax and pronunciation more strongly influence one another than either—taken separately—influences or is influenced by lexical distance?

We attack these latter questions in multiple regression designs, checking for the effects of linguistic levels on one another once geography is included as an independent variable.

3. Data sources

This research is based on two Dutch dialectal data sources: the *Reeks Nederlandse Dialectatlassen* (RND; 'Series of Dutch Dialect atlasses'; Blancquaert and Peé, 1925-1982) and the first volume of the *Syntactische Atlas van de Nederlandse Dialecten* (SAND1; 'Syntactic Atlas of the Dutch Dialects'; Barbiers et al., 2005). Both atlasses describe Dutch dialects in the Netherlands, the Northern part of Belgium and a small northwestern part of France. The RND data also include the north-eastern area of the Belgian province Luik and the German county Bentheim.

The RND is a 16-volume series of Dutch dialect atlasses which were edited by Blancquaert and Peé. The first volume was compiled by Blancquaert and appeared in 1925. The final volume was published in 1982 and was edited by Peé. The RND contains translations and phonetic transcriptions of 139 sentences in 1 956 Dutch dialects. The data were recorded between 1922 and 1975. We use a digitised selection of 125 words from 360 dialects. Figure 1 shows the geographical distribution of the RND locations. The selected words represent all vowels and consonants and are used to measure both pronunciational and lexical distances. Heeringa (2001) discusses the selection of words and dialect locations from the RND in detail.[1] The next section provides several examples of words and transcriptions in the RND.

SAND1 contains 145 geographical distribution maps of individual syntactic variables in 267 Dutch dialects. Figure 2 shows the geographical distribution of the SAND locations. It covers syntactic variation related to the left periphery of the clause and pronominal reference. This includes variation with respect to complementisers, subject pronouns and expletives, subject doubling and subject clitisation following yes/no, reflexive and reciprocal pronouns, and fronting phenomena. SAND1 contains 106 syntactic contexts.[2] Tables 1 and 2 provide two examples of variation in syntactic contexts as described in SAND1.[3] The second and final volume of the SAND is due to appear in 2007 and will describe syntactic variation in Dutch dialects with respect to verbal clusters, negation and quantification.[4]

As stated above, the RND data are used to measure both pronunciational and lexical distances. The SAND1 data are used to measure syntactic distances. However, we can only relate the measurements obtained from these two data sources if the results are based on exactly the same set of dialect locations. We cannot assume that two geographically close locations are also closely related on all three linguistic levels. Therefore, we only use the intersection of the 360 RND dialects and the 267 SAND1 dialects.[5] As shown in Figure 3, the resulting 70 common dialect varieties in the Netherlands and the Northern part of Belgium are not perfectly geographically distributed.[6] The north-eastern and southern areas are overrepresented and the western and central areas are somewhat underrepresented. However, these underrepresented

---

[1] The RND data are publicly available at http://www.let.rug.nl/~heeringa/dialectology/atlas/rnd.

[2] The number of available syntactic contexts is lower than the number of geographical maps because SAND1 contains numerous correlation maps which show syntactic variables from different perspectives. Also, some syntactic contexts are presented using multiple maps.

[3] Spruit (2006) provides more examples of syntactic contexts in SAND1.

[4] The SAND data are accessible from the Dynamic Syntactic Atlas of the Dutch dialects (DynaSAND) at http://www.meertens.knaw.nl/sand.

[5] The Dutch language area under investigation, as shown in Figure 3, borders on the North Sea in the North and in the West. Germany lies along the Eastern border. The south-western border of the province West-Vlaanderen lies adjacent to France. The remaining southern border follows the Dutch-French language border in Belgium.

[6] These are the 70 common dialect varieties in alphabetical order, as shown in Figure 3: Aalst, Aalten, Almelo, Anjum, Appelscha, Arendonk, Bakkeveen, Bellingwolde, Bergum, Beveren, Boutersem, Bree, Brugge, Coevorden, Druten, Eibergen, Emmen, Ferwerd, Fijnaart, Gemert, Gent, Geraardsbergen, Gistel, Goes, Gramsbergen, Groenlo, Groesbeek, Groningen, Haaksbergen, Heerenveen, Hindeloopen, Hollum, Houthalen, Huizen, Humbeek, Kamperhout, Kerkrade, Kollum, Kortrijk, Lauw, Lemmer, Mechelen, Midsland, Oldemarkt, Onstwedde, Oostende, Ootmarsum, Opperdoes, Ossendrecht, Overijse, Roeselare, Ronse, Roswinkel, Schiermonnikoog, Spakenburg, Staphorst, Steenbergen, Steenwijk, Tegelen, Tienen, Urk, Utrecht, Vaals, Veurne, Vriezenveen, Waregem, Warffum, West-Terschelling, Zierikzee, Zundert.

areas are known to have relatively fewer differentiating characteristics than the overrepresented areas. Therefore, we expect the intersection of the RND and SAND1 dialects to adequately represent the language variation spectrum in the Dutch dialect area for our purposes.

4. Distance measures

The dialect differences *within* each linguistic level need to be measured before the associations *between* the linguistic levels can be quantified. We use the *Levenshtein* distance and the *gewichteter Identitätswert* method to measure the dialect differences within each linguistic level.

The Levenshtein distance is used to measure pronunciational differences. It was first described in Levenshtein (1966). Generally speaking, it is a string edit distance measure which calculates the minimally required steps to change one sequence of symbols to another sequence of symbols. Sankoff and Kruskal (1999) discuss a broad range of applications of the Levenshtein distance. Contrary to other well-known distance measures such as the Hamming, Manhattan and Euclidean distance measures, the Levenshtein distance measure is able to quantify the differences between sequences of different lengths. The algorithm is based on the optimal alignment between two sequences of symbols and uses one of the operations *insert*, *delete* or *substitute* at each symbol comparison. Kessler (1995) first applied the Levenshtein distance to measure differences between phonetic transcriptions of word pronunciations in Irish Gaelic dialects. Heeringa (2004) refines the Levenshtein algorithm in several ways to more accurately measure pronunciational differences in Dutch dialects. It describes the enhanced version of the algorithm we use in this work in great detail on pages 79-119. The refinement uses comparisons of spectograms of the component sounds to differentiate between dissimilar sounds acoustically.

Table 3 illustrates the string alignment principle and the Levenshtein distance calculation between two pronunciations of the Dutch word *hart* ‘heart’. The example does not take into account the refinements mentioned above as to more clearly illustrate the general principle. The word *hart* is pronounced as [hart] in Haarlem, whereas in Brugge people say [ærtə]. First, the Levenshtein algorithm aligns the two pronunciations optimally. Then, the number of edit operations are counted which are required to change the first pronunciation into the second one. Finally, the number of operations is divided by the string alignment length to obtain the normalised Levenshtein distance between these two pronunciations of the word *hart*, which, in this case, is (1+1+1 / 5 =) 0.6. The aggregate pronunciational distance between Haarlem and Brugge is calculated by accumulating all pronunciational distances between the two dialects and dividing the aggregate distance by the total number of pronunciational comparisons.

Lexical and syntactic distances are measured at a nominal level using the *gewichteter Identitätswert* (GIW).[7] This is a frequency-weighted similarity value which was introduced in dialectometry by Goebl (1984). The GIW method counts infrequent words more heavily than frequent ones. This opposes the tendency in several areas of quantitative linguistics that very infrequent words should be treated as noise, unreliable evidence of linguistic structure (Nerbonne and Kleiweg 2007). We use the inverse of Goebl’s original similarity measure to obtain GIW distance values by subtracting the similarity value from 1.

The RND data require additional preparation before they can be used to measure lexical distances. This step arises because the RND does not contain lexical identity information between word pronunciations. Therefore, we manually determined the represented lexemes for each of the 125 word pronunciations from a layman’s perspective. We did not analyse the word pronunciations from an etymological point of view. The following example of the lexical concept *zijn* ‘to be’ illustrates the lexeme identification procedure. The word pronunciations [bɪn̩], [bɪnt] and [bɛnə] are considered to be forms of a single lexeme, which however differs

[7] The GIW method measures differences between variable pairs at a nominal level. This means that two variables are either equal or unequal. The Levenshtein distance is a numerical measure which allows differentiation between variable pairs in terms of degrees of similarity.

from the single lexeme instantiated in [zɛɪn], [zɪnt] and [zɑn], even though the pronunciations [bɪnt] and [zɪnt] seem very similar. Also, inflectional variants do not play a role in the context of this procedure. For example, the words [boːmkə] and [bo̝ˑᵘmpjə] are identified as two pronunciations of the lexeme *boompje* 'little tree'. Both words are morphologically derived from the root *boom* 'tree'. Heeringa et al. (2007) contains more information regarding the lexeme identification procedure. In contrast to the above, SAND1 does not require additional annotation of the data. It already presents each syntactic variable within its syntactic context.[8]

Table 4 illustrates the GIW method as a measure of lexical similarity between the dialects of Middelstum and Ommen. As already noted, we employ the inverse of the original similarity measure as illustrated in Table 4 to obtain GIW distance values by subtracting each similarity value from 1. The distance calculations are omitted from Table 4 to enhance readability. The described procedure is identical when syntactic differences are measured. The example shows that the two dialects use the same lexemes for the concepts *vriend* 'friend' and *schip* 'ship': *kameraad* 'comrade' and *schip* 'ship', respectively. However, the two dialects use a different lexeme to reference the concept *duwen* 'to push'. In Middelstum people say *stoten* 'to thrust', whereas in Ommen people use *drukken* 'to press'.

This information is used to calculate the lexical distance between the two dialects. First, the lexeme *kameraad* 'comrade' references the concept *vriend* 'friend' in 140 dialects. In 114 dialects a different lexeme is used instead. This results in a weighted similarity of (1 – 140 / 354 =) 0.6 and a complementary GIW distance value of (1 - 0.6 =) 0.4. Unfortunately, in six dialects no data was available for this concept. We ignore missing concepts because there is obviously nothing to measure.[9] Next, the concept *schip* 'ship' is nearly always referenced by the same lexeme. Therefore, the GIW method considers this information to be of little help in quantifying the linguistic variation between the two dialects. The weighted similarity of (1 – 353 / 360 =) 0.02 and the corresponding GIW distance of 0.98 reflect this consideration appropriately. The different similarity weights (0.6 versus 0.02) assigned to the first two concepts in Table 4 demonstrate that similarity weighting in the GIW method *emphasizes* rather than ignores infrequently occurring words. Finally, the third concept *duwen* 'to push' in Table 4 is realised with different lexemes in the two dialects. The GIW method always assigns different lexemes a similarity value of 0.0 to designate the dissimilarity between the lexemes. This is equivalent to a maximum GIW distance of 1.0. The lexical GIW distance measurements between the dialects of Middelstum and Ommen based on the three concepts shown in Table 4 result in a weighted similarity of (0.62 / 3 =) 0.21, which this work translates into the corresponding lexical GIW distance value of (1 – 0.21 =) 0.79.

5. Dutch dialect area perspectives

We present colour maps of the Dutch dialect areas based on pronunciational, lexical and syntactic differences in pairwise comparisons to provide a general impression of the associations between the pronunciational, lexical and syntactic levels before we calculate the exact degrees of association in Section 7. We first present the De Schutter (1994) map and the Daan and Blok (1969) map of the Dutch dialect areas in Figures 4 and 5 as external points of reference. These two non-computational dialect maps are based on perception and expert opinion, respectively.

Our dialect colour maps employ Multidimensional Scaling (MDS) to visualise the pronunciational, lexical and syntactic variation in the Dutch language area. This statistical technique was first described in Torgerson (1952) . We apply the MDS procedure to display the general dialect relationships as faithfully as possible in one three-dimensional, full-colour

[8] Roughly speaking, in SAND1 each geographical distribution map represents a syntactic context and each map symbol represents a syntactic variable. A map symbol, by definition, can only be shown on a map. Therefore, SAND1 variables are always presented within context.

[9] The lexical distance measurements are based on GIW comparisons between 103 and 125 concepts, with 121 concepts on average.

picture. The procedure to visualise the distance measurements consists of the following three steps. First, each dialect's distance relationships to all other dialects are reduced to coordinates in a three-dimensional space using the three most important dimensions arising from the MDS analysis. These coordinates optimally represent the original dialect distance relationships. They do not directly correspond to actual dialect distances anymore. Second, the three-dimensional coordinates are used as values between light and dark of the three colour components red, green and blue. This effectively means that a dialect's unique set of characteristics is translated into one unique composite colour. Neighbouring dialects have corresponding colours if they are also linguistically close to each other. They are progressively assigned less related colours as they are less related linguistically. Third, the dialect points on the maps are blown up to small areas until they border each other and there is no uncoloured space left. This space partitioning technique uses the well-known Delaunay triangulation to obtain a pattern of Voronoi polygons.[10] The final result is that a colour continuum arises if there is a perfect relation between geographical distance and linguistic distance, whereas a mosaic-like map results when this relation is not strong. Heeringa (2004:156-163) discusses the technical details of the MDS technique and the Delaunay triangulation in detail from a linguistic perspective.

We need to ensure that the MDS maps visualise the linguistic variation accurately. The MDS procedure calculates a correlation coefficient to indicate the amount of linguistic variance which is represented in the first three dimensions of the MDS solution and, therefore, in the MDS map colours. The correlation coefficients between the dialect distance relationships and the MDS coordinate distance relationships are 0.94, 0.74 and 0.89 at the pronunciational, lexical and syntactic levels, respectively. The coefficients are also shown as *r*-values in Figures 6-8. In most applications correlations below 0.8 tend to be too inaccurate to be interpreted meaningfully, whereas coefficients between 0.9 and 1 are generally considered to be high. Norušis (1997) notably defines $r^2 = 0.6$ (i.e. $r = 0.77$) as the minimum acceptable correlation in the context of the MDS procedure. However, the exact correlation threshold values likely vary within each specific research context. All in all, we conclude that the dialect colour maps in Figures 6 and 8 accurately represent the original distance measurements at the pronunciational and syntactic levels. The lexical MDS map in Figure 7 also represents the original lexical variance to an acceptable extent for our exploratory purposes, but it should be interpreted more cautiously.

The Daan and Blok dialect map in Figure 5 shows the classification of the Dutch dialect varieties based on subjective judgements from local speakers, local experts and the map designers themselves. The Netherlandic dialect area borders were derived from a written survey from 1939 among 1 500 local dialect speakers. The Belgian part was mostly based on the perception of local dialect experts. The methodology and results of the map creation procedure are discussed in Heeringa (2004:12-13), among others. Our dialect colour maps follow Daan and Blok (1969) in the assignment of the colour blue to the Frisian area in the central north and the colour green to the north-eastern Lower Saxon region to simplify comparisons between the dialect maps.

Spruit (2005) provides a visual comparison between a syntactic MDS map based on Hamming distances and the perceptual Daan and Blok map in Figure 5. The syntactic MDS map in Spruit (2005) is very similar to the syntactic MDS map based on GIW distances shown in Figure 8. Therefore, the dialect maps in Figures 5 and 8 are also remarkably similar. Interestingly, "[…] the Belgian dialect classification on the Daan and Blok map based on more objective expert judgements corresponds to a higher degree with the classification based on the objective syntactic measure than with the Netherlandic dialect classification based on intuitive judgements" (Spruit 2005:189). Among other suggestions, the work mentions the role of prejudice in perception and sensitivity to pronunciational differences as possible explanations. Heeringa (2004:230-233) discusses the similarities and differences between the perceptual Daan and Blok map in Figure 5 and the pronunciational MDS map shown in Figure 6.

[10] Goebl (1982) introduces the use of Voronoi tiling, sketched here, to illustrate the results of dialectometric analyses. Alternatively, an interpolation procedure could be applied to colour the space between dialect locations.

The De Schutter map in Figure 4 is a simplified expert consensus map of the Dutch dialect areas. It is heavily based on the Daan and Blok map but also relies on several other dialect maps.[11] The author considers this map to reflect the general opinion of traditional dialectologists at the end of the 20th century. It shows the six main Dutch dialect areas: the central-northern Frisian, north-eastern, central-western, south-western, central-southern and south-eastern dialects.

The two maps in Figures 6 and 7 visualise the variation in the Dutch language area with respect to pronunciational differences and lexical differences, respectively. We can expect a substantial correlation between the two linguistic levels based on the visual correspondences between the two maps. For example, the central-northern Frisian area in blue stands out very prominently on both maps. The most prominent difference is arguably the clear-cut northern border on the lexical map of the central-south area in pink. This border can not be made out on the pronunciational map.

Figure 8 shows the variation in Dutch dialects with respect to syntactic variation. When we visually compare this map with the lexical map shown in Figure 7, we can already be quite certain that the degree of association between the syntactic and lexical levels will be lower than the correlation between the pronunciational and lexical levels, as discussed in the previous paragraph. For example, the appearance of the Frisian area in blue on the syntactic map is not nearly as prominent as on the lexical map. Also, the south-east area on the syntactic map in light-blue is quite prominently present, whereas this area can hardly be made out on the lexical map.

The final pair of maps compares pronunciational (Figure 6) and syntactic (Figure 8) differences in Dutch dialects. After a single glance at these two maps we can already speculate that the correlation between these two linguistic levels will be higher than the correlation between the lexical and syntactic levels in Figures 7 and 8, respectively. However, it is uncertain whether the correlation between the pronunciational and syntactic levels is also stronger than the correlation between the pronunciational and lexical levels in Figures 6 and 7, respectively. For example, these two maps differ in the degree of separation with respect to the blue Frisian area, but they correspond to a higher degree in the southern areas than the pronunciational versus lexical maps correspond with each other. Therefore, we need to *calculate* the degree of association among these three linguistic levels to answer this question satisfactorily.

6. Consistency

We want to ensure that our distance measurements are consistent—we want to know that our results are reliable. Therefore, we use Cronbach's alpha to measure the minimum reliability of our distance measurements when applied to our data sources. Cronbach's alpha was first described in Cronbach (1951). It is a coefficient of consistency and can be described as a function of the number of linguistic variables (n_{var}) and the average inter-correlation value among the variables (r), i.e. the mean of all the familiar Pearson product-moment correlation coefficients, given in Equation 2.[12] Its values range between zero and one. Higher values indicate more reliability. As a rule of thumb, values higher than 0.7 are considered sufficient to obtain consistent results in social sciences (Nunnally 1978). The Cronbach's alpha formula is shown in Equation 1. The formula to obtain the average inter-correlation value among the variables (r) is listed in Equation 2.

Table 5 presents the Cronbach's alpha values which indicate the reliability of our measurement results at the pronunciational, lexical and syntactic levels. Based on the Cronbach's alpha value of 0.97 we can conclude that the Levenshtein analysis of the

[11] The De Schutter map based on expert consensus also takes the Dutch dialect area classifications in Weijnen (1958) and Goossens (1977) into account.

[12] The Pearson product-moment correlation coefficient (PMCC) is the most commonly used method of computing a correlation coefficient between variables that are linearly related.

pronunciational data is very reliable. The GIW analysis of the syntactic data results in a value of 0.94 which also indicates very consistent results. The GIW analysis of the lexical data brings about a coefficient of consistency of 0.75, which is acceptable. It does indicate, however, that the analysis of the lexical data may be less reliable than the analyses at the pronunciational and syntactic levels.

Finally, it may be helpful to explicitly point out the interpretational difference between Cronbach's alpha coefficients and MDS correlation coefficients which were described in the previous section. Cronbach's alpha coefficients estimate how well the measurement results of an analysed dataset, representing the variation within the linguistic level, can be expected to capture the variation within the entire linguistic domain. Simplified, it measures the level of reliability of the results. In our research context the MDS correlation coefficients indicate the amount of linguistic variance which is represented in the first three dimensions of the MDS solution. Simplified, it measures the level of accuracy of the scaling procedure.

7. Correlations between linguistic levels

Table 6 answers the first of the two research questions central to this paper, as stated in Section 2: to what degree are aggregate pronunciational, lexical and syntactic distances associated with one another, when measured among varieties of a single language? We have calculated the Pearson product-moment correlation coefficients among the distance measurements for the three linguistic levels as a measure of the degree to which the three linguistic levels are associated. The results show that pronunciation is marginally more strongly associated with syntax (42%) than with lexis (38%) and that syntax is much more strongly associated with pronunciation (42%) than with lexis (25%).

The results below are based on the 70 common varieties as described in Section 3. The pronunciational differences were measured using the Levenshtein distance and the GIW method was applied to measure the variation at the lexical and syntactic levels. The percentages in Table 6 indicate the amount of variation at the first linguistic level which can be explained with the amount of variation at the second linguistic level. All correlation coefficients are significant at the 0.001 level. In order to not confound significance calculations between distance tables, the significance levels of the correlation coefficients were calculated using the Mantel test (Mantel 1967).

The Mantel test calculates the significance levels of correlation coefficients between distance tables while taking into account the structured, interdependent nature of distance matrices. The null hypothesis in this asymptotic test states that there is no correlation between the symmetrical dialect distances in the two matrices. In other words, the test assumes that changes in dialect distances at the first linguistic level do not influence the dialect relationships at the second linguistic level. The hypothesis is evaluated by randomly reallocating the order of elements in the first matrix many times, and recalculating the correlation between the permuted first matrix and the original second matrix after each permutation. The significance of the observed correlation results from the proportion of the permutations that lead to a higher correlation than the actual coefficient. With a significance level of $\alpha = 0.05$ the number of repetitions should be equal to about 1 000 (Manly 1997). This means that less than five percent of thousand permuted matrix correlations may yield higher coefficients than the correlation coefficient between the original matrices. The reasoning is that if the null hypothesis—there is no correlation between the two matrices—is correct, then the permuted matrix should be equally likely to produce a larger or a smaller correlation coefficient.

With this information we can also answer the subquestion of our first research question: are syntax and pronunciation more strongly associated with one another than either is associated with lexical distance? To our surprise, the expectations we laid out in Section 2 were not decisively met. Although syntax is clearly more strongly associated with pronunciation (r = 0.684) than with lexis (r = 0.496), the syntax-pronunciation association (r = 0.648) is not much stronger than the lexis-pronunciation connection (r = 0.617). At this point we can only speculate about these outcomes. We already pointed out that the Cronbach's alpha value for the lexical

analysis is relatively low. This leaves room for less reliable results. Also, we already acknowledged that the pronunciational data includes subphonological variation. It might be the case that variation at the phonetic and morphological sublevels is distributed in different patterns than purely phonological variation. This could reduce the expected correlation with the syntactic data. However, we should not draw any conclusions before having checked the correlations for the influence of a third, underlying factor: geography.

8. Linguistic levels correlated with geography

Geography has independently been shown to correlate strongly with each of the three linguistic levels under investigation. Heeringa and Nerbonne (2001) examined the degrees of association between geographical and pronunciational distances in Dutch dialects. Cavalli-Sforza and Wang (1986) related geographical distances with lexical similarities in a chain of Micronesian islands. The correlation between geographical and syntactic distances in Dutch dialects was analysed in Spruit (2006). In this study we present the scatterplots and correlation values of pronuncational Levenshtein distances versus geographical distances in Figure 9, lexical GIW distances versus geographical distances in Figure 10 and syntactic GIW distances versus geographical distances in Figure 11. All results are based on the 70 common varieties as described in Section 3. The scatterplots show the associations between each of the three linguistic levels as dependent variables on the Y-axes and geography as the independent variable on the X-axis.

The geographical distances have been calculated using the *ll2dst* programme, which is part of the freely available dialectometry software package RuG/L04. The programme takes longitude-latitude coordinates to calculate the corresponding geographical distances in kilometers ‘as the crow flies’. The algorithm assumes that the earth is a perfect sphere and that it has a circumference of 40 000 kilometers. Although neither of these two assumptions is entirely correct, it should not noticably affect the accuracy of our distance calculations. The Dutch language area only covers a very small surface of the earth’s sphere. Therefore, the Dutch area surface remains relatively flat and the distance calculations remain accurate.[13]

The current operationalisation of the factor geography as Euclidean distances between longitude-latitude coordinates is an acceptable approximation of geographical distance in the case of the Dutch language area under investigation. However, a more refined measure of geographical distance may be required in situations where geographical barriers may influence the chance of social contact considerably. Gooskens (2004) notably illustrates the effect of geography on dialect variation in Norway, where the central mountain range prevented direct travel until recently. In Norway travel time turns out to be a much better predictor of linguistic distance than distance ‘as the crow flies’. Of course, there are no mountain ranges, dry deserts, tropical forests or other types of inhospitable geographical barriers within the Dutch language area. Van Gemert (2002) examines the influence of water barriers such as lakes and rivers in the Netherlands on pronunciational distances between dialects. Contrary to its expectations, however, it concludes that traveling costs between dialects never correlate to a higher degree with pronunciational variation than geographical distances ‘as the crow flies’. The remainder of this work, therefore, feels confident in the application of distances ‘as the crow flies’ as an adequate operationalisation of geography.

Table 7 shows the degrees of association between each linguistic level versus geography. The results clearly demonstrate that linguistic differences at the pronunciational and syntactic levels are more strongly associated with geographical distances (47% and 45%, respectively) than with variation at the lexical level (33%). All Pearson correlation coefficients are significant at the 0.001 level. The percentages in the right column are based on r^2 values, which indicate the amount of variation at the specified linguistic level which can be explained with

[13] The *ll2dst* manual at http://www.let.rug.nl/~kleiweg/L04/Manuals/ll2dst.html contains more information on this software programme.

geographical distance. The results confirm the fundamental postulate in dialectology that language varieties are structured geographically (Nerbonne and Kleiweg 2007).

9. Linguistic correlations without the influence of geography

Section 7 presented the degrees of association among aggregate pronunciational, lexical and syntactic distances. However, in Section 8 we found that geography influences each of the three linguistic levels separately. Therefore, we need to refine the results in Section 7 by accounting for the structural influence of geography as an underlying, third factor. Based on the strong correlations between geography and each linguistic level separately, as shown in Section 8, we cannot assume that there is influence among the various linguistic levels. However, we can test for this.

The following three steps describe the procedure to calculate the correlation between two linguistic levels without geography as an influencing factor. This example takes pronunciational variation as the first linguistic level and lexical variation as the second linguistic level. First, we perform a regression analysis between the pronunciational distances and the geographical distances. This results in the pronunciational residuals. Residuals are those parts of the data which the regression model does not explain. Second, we likewise perform a regression analysis between the lexical distances and the geographical distances, which results in the lexical residuals. Third, we run a regression analysis between the pronunciational residuals and the lexical residuals which we obtained in steps one and two. This provides the correlation coefficient between pronunciational distances and lexical distances without the influence of geographical distances.

We repeat this procedure to calculate the correlation between the lexical and syntactic levels and the correlation between the syntactic and pronunciational levels. The results are presented in Table 8. Again, the results are based on the 70 common varieties as described in Section 3. The pronunciational differences were measured using the Levenshtein distance and the GIW method was applied to measure the variation at the lexical and syntactic levels. All correlation coefficients are significant at the 0.001 level using the Mantel test. Section 7 already explained why the significance levels of the calculated correlations between the linguistic levels are reliable, even when applied to the structured, interdependent data of distance matrices. The percentages in Table 8 indicate the amount of variation at the first linguistic level which can be explained with the amount of variation at the second linguistic level.

With the degrees of association in Table 8 we can answer the second of our two research questions: is there evidence for influence among the linguistic levels, even once we control for the effect of geography? The answer is that some influence between pronunciation and syntax (12%) remains, although the association between pronunciation and lexis is stronger (14%). There is virtually no association between syntax and lexis (merely 3%).

Table 9 presents the influence of geography as a factor of influence underlying the associations between aggregate pronunciational, lexical and syntactic distances. Equation 3 shows the formula to calculate the influence of geography underlying the associations between the linguistic levels:

The formula in Equation 3 takes the correlation *(r)* values from Tables 6 and 8, respectively. Table 9 evidently shows the substantial influence of geography as a factor of influence underlying the associations between the linguistic levels. The degree of association between pronunciational and lexical distances turns out to be based on geography as an underlying factor for no less than 39%.[14] The association between syntactic and pronunciational distances is even more heavily based on geography as a third factor (46%). The apparant association between syntactic and lexical distances turns out to be principally due to geography as a third factor (63%).

[14] The geographical influence underlying the association between pronunciational and lexical distances is calculated as follows: (1 - (0.374 / 0.617)) * 100 = 39%.

10. Conclusions

Without controlling for the effect of geography, pronunciation is marginally more strongly associated with syntax (42%) than with lexis (38%) and syntax is much more strongly associated with pronunciation (42%) than with lexis (25%). Pronunciation and syntax are more strongly associated with geography (47% and 45%, respectively) than lexis is (33%).

However, once the influence of geography is filtered away as a factor of influence underlying the associations among the linguistic levels under investigation, the association between pronunciation and syntax turns out to be largely based on geography as an underlying factor (46%). Some influence between pronunciation and syntax remains (12%), although the association between pronunciation and lexis is stronger (14%). There is virtually no association between syntax and lexis (3%).

11. Discussion and future research

We wish to point to two consequences beyond the raw correlations of the distances among the linguistic levels, as interesting as these are on their own. First, the modest correlation (r = 0.35) between syntactic and pronunciational variables in Table 8 indicates that 12% of the proportion of variance in common between the two variables cannot be explained by geography. It might be explained by typological constraints—i.e. by constraints obtaining between syntactic and phonological structure—which would be very interesting. If we had found no interesting level of correlation between these levels on the one hand and the lexical level on the other, one might postulate immediately that typological constraints are responsible for this modest correlation. But we, in fact, did find a comparable level of correlation between pronunciation and lexical choice, for which structural, typological constraints seem unlikely. We therefore must allow that extralinguistic, but clearly non-geographic explanations are equally plausible as candidates to explain the correlation.

Second, we turn to the modest correlation (r = 0.37) between pronunciational and lexical variation on the one hand and the low, but significant correlation (r = 0.18) between lexical and syntactic variation on the other. These coefficients in Table 8 indicate that 14% of the proportion of variance in common between lexical and pronunciational distances on the one hand, and 3% of the proportion of variance in common between lexical and syntactic distances on the other hand, cannot be explained by geography. As we have argued above, it is unlikely that these correlations may be explained by linguistic constraints, and since the correlations were obtained from the residues of a regression analysis in which geography was the independent variable, they are not explained by geography. This suggests that there must be further extralinguistic conditioning of variation that we as dialectologists should set in our sites. The literature on language variation suggests many candidates for such conditioning variables, but there have been too few data collection efforts aimed at cataloguing linguistic variation and candidate explanatory variables, including e.g. sex, education, class, social network, etc. This would indeed be a daunting task, but the present paper has sketched the sorts of analysis one could perform on the data, once it is available.

To summarise, the degrees of association among the linguistic levels presented in Section 9 are substantial but not overwhelming. There is influence between the various linguistic levels, even once we control for the dominant effect of geography. We assume that a more evenly geographically distributed set of dialect varieties may result in stronger degrees of association, since the current set of common varieties overrepresents the average variation spectrum in the Dutch language area. Regardless, the results further strengthen the fundamental postulate in dialectology that language varieties are structured geographically.

We note, however, that the results at the lexical level are consistently less strong in comparison to the results at the pronunciational and syntactic levels. We speculate that the unfavorable lexical results reflect the lower quality of the lexical data set. The consistency analysis of the lexical data in Section 6 hints at this direction. Future work will further examine

the lexical data using a bootstrapping technique to analyse the influence of the selection of words on the results.

Once geography is controlled for as an underlying factor of influence, the lack of association between lexis and syntax accords with our expectations as stated in Section 2. However, we are surprised that the association between lexis and pronunciation is somewhat stronger than the correlation between syntax and pronunciation. We would have expected the highly structured syntactic and pronunciational systems to share more distributional patterns, in contrast to the volatile lexicon. We suspect that this outcome is another reflection of the somewhat lower quality of the lexical data. Also, the unbalanced nature of the syntactic data may be a factor of influence. SAND1 only describes variation in the left periphery of the clause and pronominal reference. However, the second volume of the SAND (SAND2, Barbiers et al., t.a. 2007) will concentrate on syntactic variation with respect to verbal clusters, negation and quantification. We will integrate the variation in these right peripherial domains in our syntactic measurements to further enhance the accuracy of our results.

Finally, pronunciational differences can arise from variation at the phonetic, phonological and morphological levels. Future research will attempt to dissect the complex interplay of these linguistic levels underlying pronunciational differences as follows. First, we are currently processing the purely morphological data in the first volume of the Morphological Atlas of the Dutch Dialects (MAND, De Schutter et al., 2005).[15] Second, we are also investigating the purely phonological data in the Phonological Atlas of the Dutch Dialects (FAND, Goossens et al., 1998-2005). We expect these extensive sources of purely morphological data and purely phonological data to provide new insights in the roles of the various linguistic levels underlying pronunciational differences, and to enrich our understanding of the associations among linguistic levels.

References

Barbiers, S., Bennis, H., Devos, M., Vogelaer, G. de, Ham, M. van der (Eds.), 2005. *Syntactic Atlas of the Dutch Dialects*, Vol. 1. Amsterdam University Press, Amsterdam.

Barbiers, S., Bennis, H., Devos, M., Vogelaer, G. de, Ham, M. van der (Eds.), t.a. 2007. *Syntactic Atlas of the Dutch Dialects*, Vol. 2. Amsterdam University Press, Amsterdam.

Blancquaert, E., Peé, W. (Eds.), 1925-1982. *Reeks Nederlands(ch)e dialectatlassen*. De Sikkel, Antwerpen.

Cavalli-Sforza, L., Wang, W., 1986. *Spatial distance and lexical replacement*. In: Language, Vol. 62, nr. 1, pp. 38–55.

Cronbach, L., 1951. *Coefficient alpha and the internal structure of tests*. In: Psychometrika, Vol. 16, pp. 297–334.

Daan, J., Blok, D., 1969. *Van Randstad tot Landrand; toelichting bij de kaart: Dialecten en Naamkunde*, Vol. XXXVII, Bijdragen en mededelingen der Dialectencommissie van de Koninklijke Nederlandse Akademie van Wetenschappen te Amsterdam. Noord-Hollandsche Uitgevers Maatschappij, Amsterdam.

Donegan, P., Stamp D., 1983. *Rhythm and the holistic organization of language structure*. In: Richardson, J.F. et al. (Eds.), Papers from the Parasession on the interplay of phonology, morphology and syntax. Chicago Linguistic Society, Chicago, pp. 337-353.

Gemert, I. van, 2002. *Het geografisch verklaren van dialectafstanden met een geografisch informatiesysteem (GIS).* Master's thesis, Rijksuniversiteit Groningen, Groningen.

Goebl, H., 1982. *Dialektometrie. Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie.* Austrian Academy of Science, Wien.

Goebl, H., 1984. *Dialektometrische Studien: Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF*, Vol. 3, Max Niemeyer, Tübingen.

[15] More information regarding De Schutter et al. (2005) is available on the official MAND website at http://www.meertens.knaw.nl/projecten/mand.

Gooskens, C., 2004. *Norwegian Dialect Distances Geographically Explained*. In: Gunnarson, B., Bergström, L., Eklund, G., Fridella, S., Hansen, L., Karstadt, A., Nordberg, B., Sundgren, E., Thelander, M. (Eds.), Language Variation in Europe. Papers from the Second International Conference on Language Variation in Europe ICLAVE 2, Uppsala, Sweden: Uppsala University, pp. 195–206.

Goossens, J. 1977. *Inleiding tot de Nederlandse dialectologie*. Wolters-Noordhoff, Groningen.

Goossens, J., Taeldeman J., Verleyen, G., 1998. *Fonologische atlas van de Nederlandse dialecten*, Vol. I, Het korte vocalisme. Koninklijke Academie voor Nederlandse Taal- en Letterkunde, Gent.

Heeringa, W., 2001. *De selectie en digitalisatie van dialecten en woorden uit de Reeks Nederlandse Dialectatlassen*. In: Hoeksema, J. et al. (Eds.), TABU: Bulletin voor Taalwetenschap, Vol. 31, nr. 1/2. Rijksuniversiteit Groningen, Groningen, pp. 61-103.

Heeringa, W., Nerbonne, J., 2001. *Dialect Areas and Dialect Continua*. In: Sankoff, D. (Eds.), Language Variation and Change, Vol. 13, Cambridge University Press, New York, pp. 375-400.

Heeringa, W., 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance*, PhD thesis, Rijksuniversiteit Groningen, Groningen.

Heeringa, W., Nerbonne, J., Bezooijen, R. van, Spruit, M., 2007. *Geografie en inwoneraantallen als verklarende factoren voor variatie in het Nederlandse dialectgebied*. In: Nederlandse Taal- en Letterenkunde, Vol. 123, nr. 1, Uitgeverij Verloren, Hilversum, pp. 70-82.

Kessler, B., 1995. *Computational dialectology in Irish Gaelic*. In: Proceedings of the 7th Conference of the European Chapter of the Association for Computational Linguistics, EACL, Dublin, pp. 60-67.

Levenshtein, V., 1966. *Binary codes capable of correcting deletions, insertions, and reversals*. In: Cybernetics and Control Theory, Vol. 10, nr. 8, pp. 707-710.

Manly, B., 1997. *Randomization, Bootstrap and Monte Carlo Methods in Biology*. Chapman and Hall, London, Second edition.

Mantel, N., 1967. *The detection of disease clustering and a generalized regression approach*. In: Cancer Research, Vol. 27, pp. 209-220.

Nerbonne, J., Heeringa, W., Kleiweg, P., 1999. *Edit Distance and Dialect Proximity*. In: Sankoff, D., Kruskal, J. (Eds.), Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison. CSLI Press, Stanford, pp. v-xv.

Nerbonne, J., Kleiweg, P., 2007. *Toward a Dialectological Yardstick*. Journal of Quantitative Linguistics Vol. 14, nr .2, Routledge, New York, pp. 148-167.

Nerbonne, J., Heeringa, W., t.a. 2007. *Geographic Distributions of Linguistic Variation Reflect Dynamics of Differentiation*. In: Featherston, S., Sternefeld, W. (Eds.), Roots: Linguistics in Search of its Evidential Base, Mouton De Gruyter, Berlin.

Norušis, M., 1997. *SPSS Professional Statistics 7.5*. SPSS Inc, Chicago.

Nunnally, J., 1978. *Psychometric Theory*. McGraw-Hill, New York.

Sankoff, D., Kruskal, J. (Eds.), 1999. *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*. CSLI Press, Stanford.

Schutter, G. de, 1994. *Dutch*. In: König, E., Auwera, J. van der (Eds.), The Germanic languages. Routledge Language Family Descriptions, London, Routledge, New York.

Schutter, G. de, Berg, B. van den, Goeman, T., Jong, T. de (Eds.), 2005. *Morphological Atlas of the Dutch Dialects*. Vol. 1, Amsterdam University Press, Amsterdam.

Spruit, M., 2005. *Classifying Dutch dialects using a syntactic measure. The perceptual Daan and Blok dialect map revisited*. In: Doetjes, J., Weijer, J. van de (Eds.), Linguistics in the Netherlands 2005, John Benjamins, Amsterdam, pp. 179-190.

Spruit, M., 2006. *Measuring syntactic variation in Dutch dialects*. In: Nerbonne, J., Kretzschmar, W. Jr. (Eds.), Literary and Linguistic Computing, special issue on Progress in Dialectometry: Toward Explanation, Vol. 21, nr. 4, pp. 493-506.

Weijnen, A., 1958. *Nederlandse dialectkunde*. Van Gorcum & Comp. N.V. - G.A. Hak & Dr. J. Prakke, Assen.

Figures

Figure 1. *Distribution of the 360 Dutch dialects in the RND atlas.*

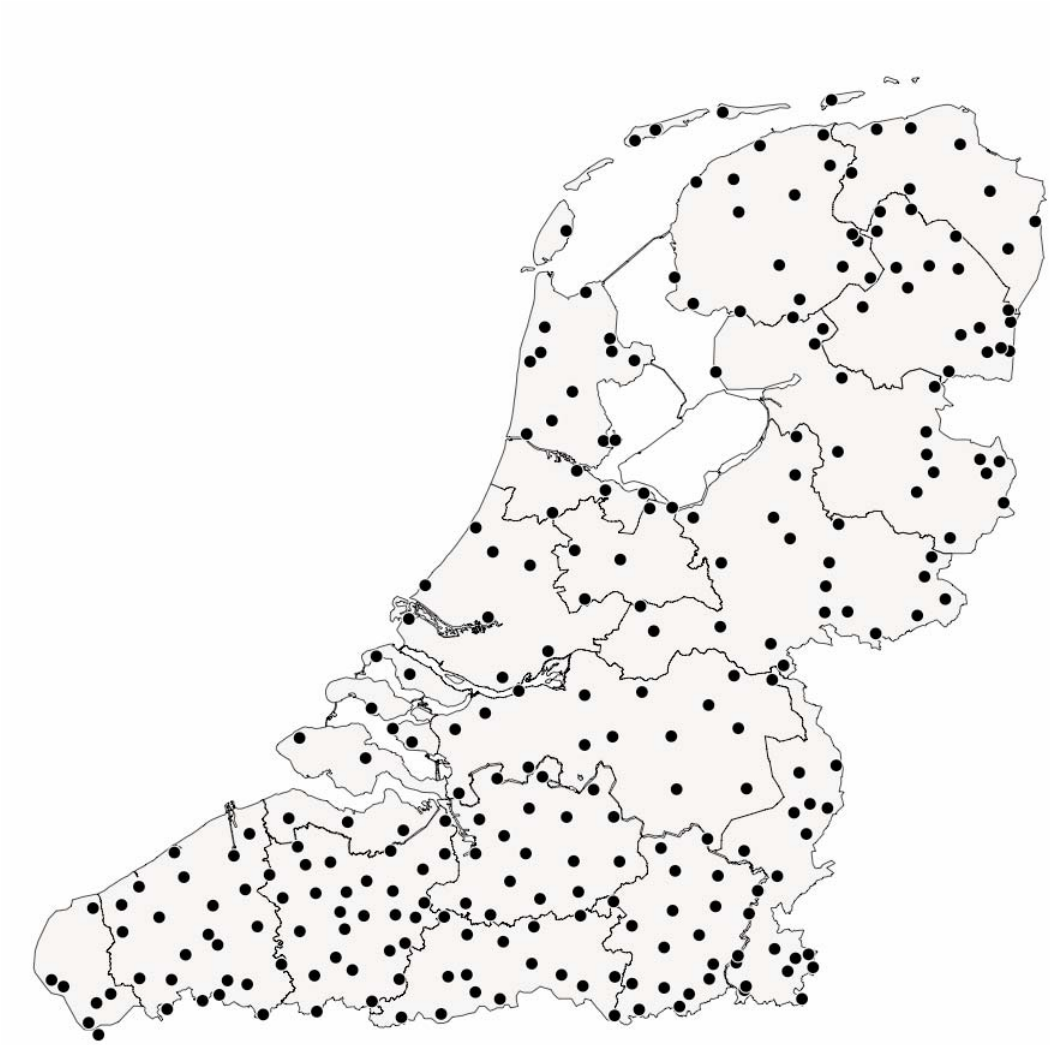

Figure 2. *Distribution of the 267 Dutch dialects in the SAND atlas.*

Figure 3. *Distribution of the 70 common Dutch dialects in the RND and SAND atlases with the relevant province names.*

Figure 4. *Expert consensus map of the Dutch dialects (translated from De Schutter 1994).*

Figure 5. *Perceptual map of the Dutch dialects based on subjective judgements (reprinted from Daan and Blok 1969).*

NOTE: Intended for colour reproduction in print and on the web

Figure 6. *Pronunciational MDS map of the Dutch dialects based on Levenshtein distances (r = 0.94).*

NOTE: Intended for colour reproduction in print and on the web

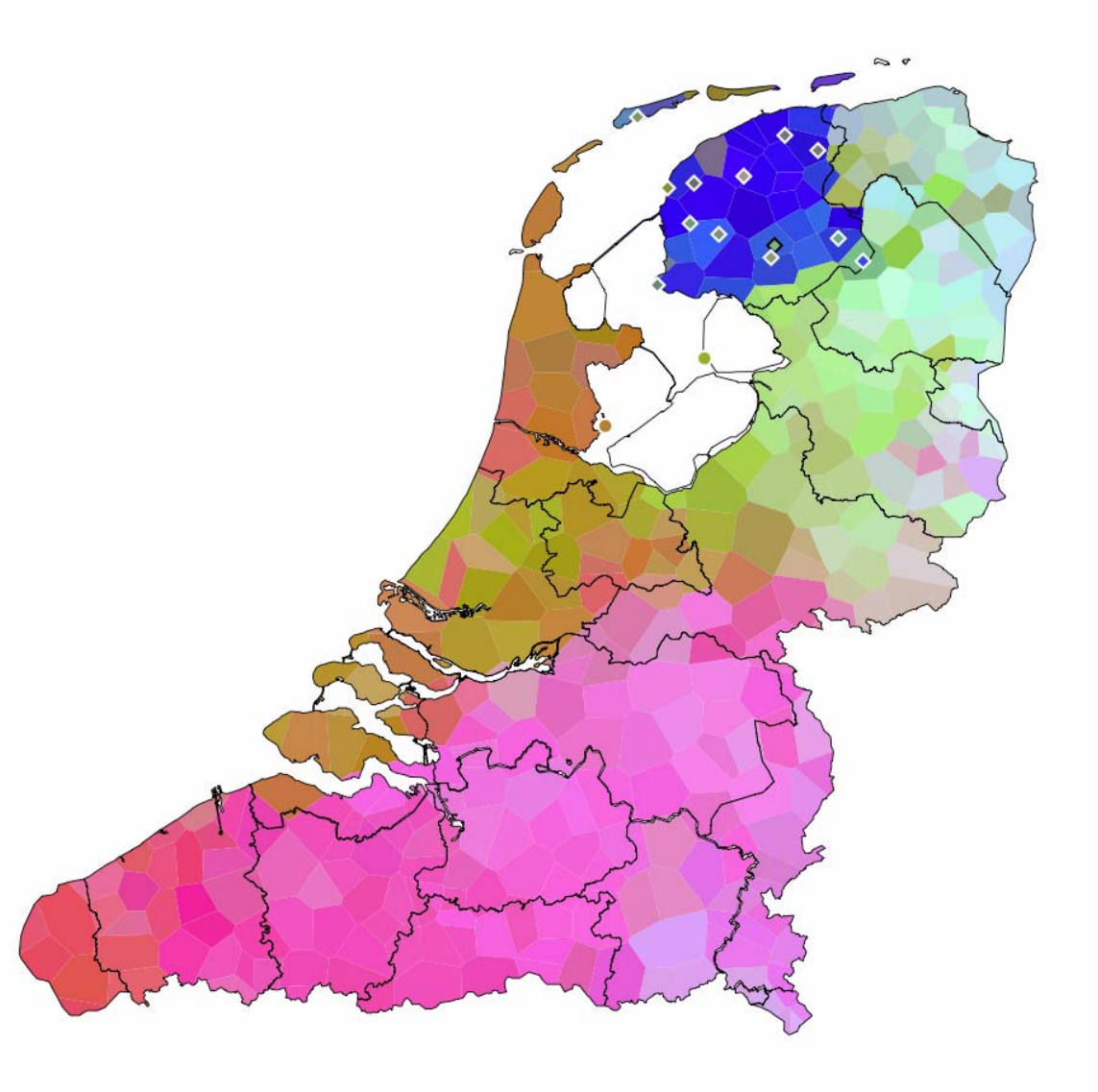

Figure 7. *Lexical MDS map of the Dutch dialects based on GIW distances (r = 0.74).*

NOTE: Intended for colour reproduction in print and on the web

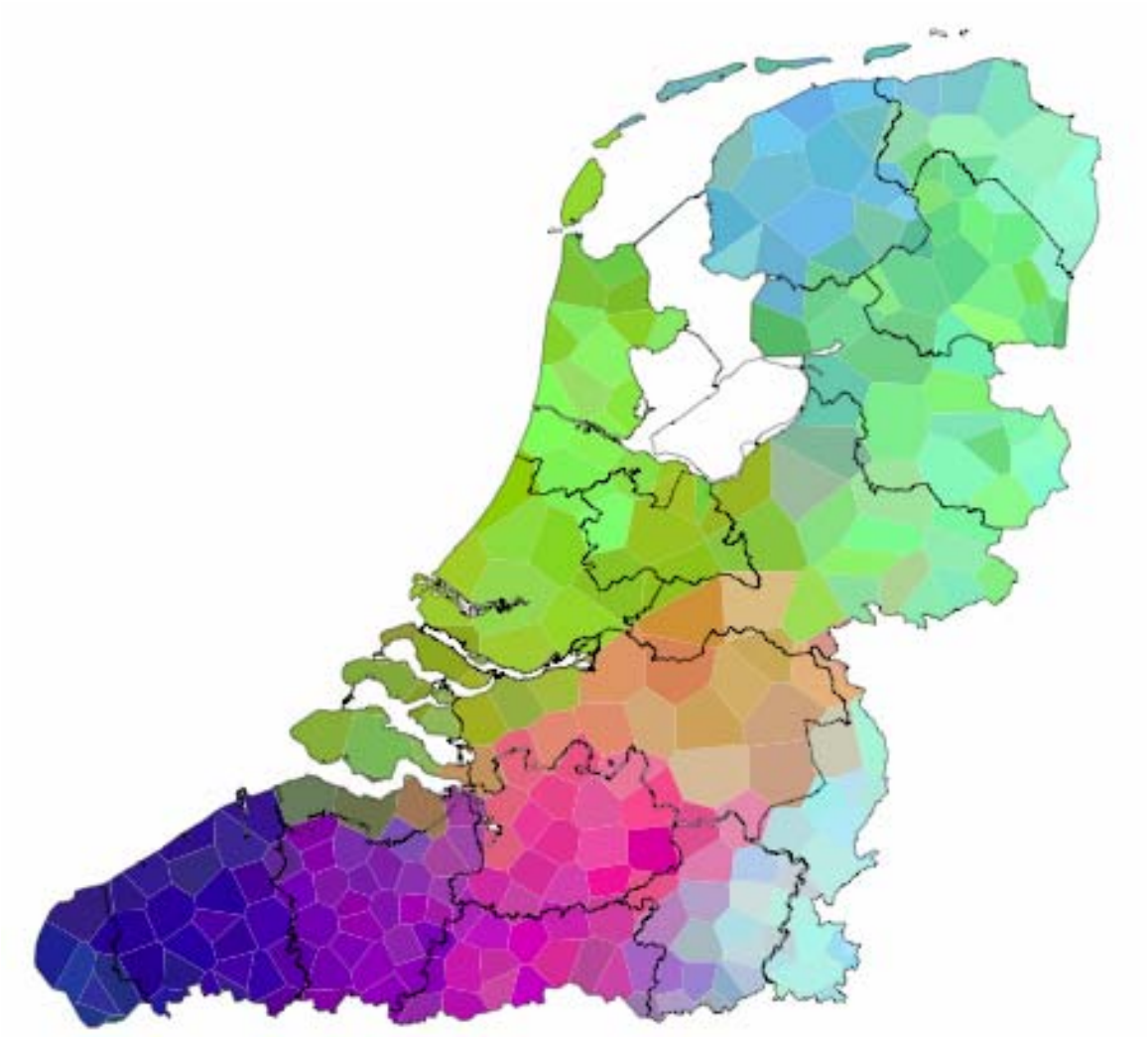

Figure 8. *Syntactic MDS map of the Dutch dialects based on GIW distances (r = 0.89).*

NOTE: Intended for colour reproduction in print and on the web

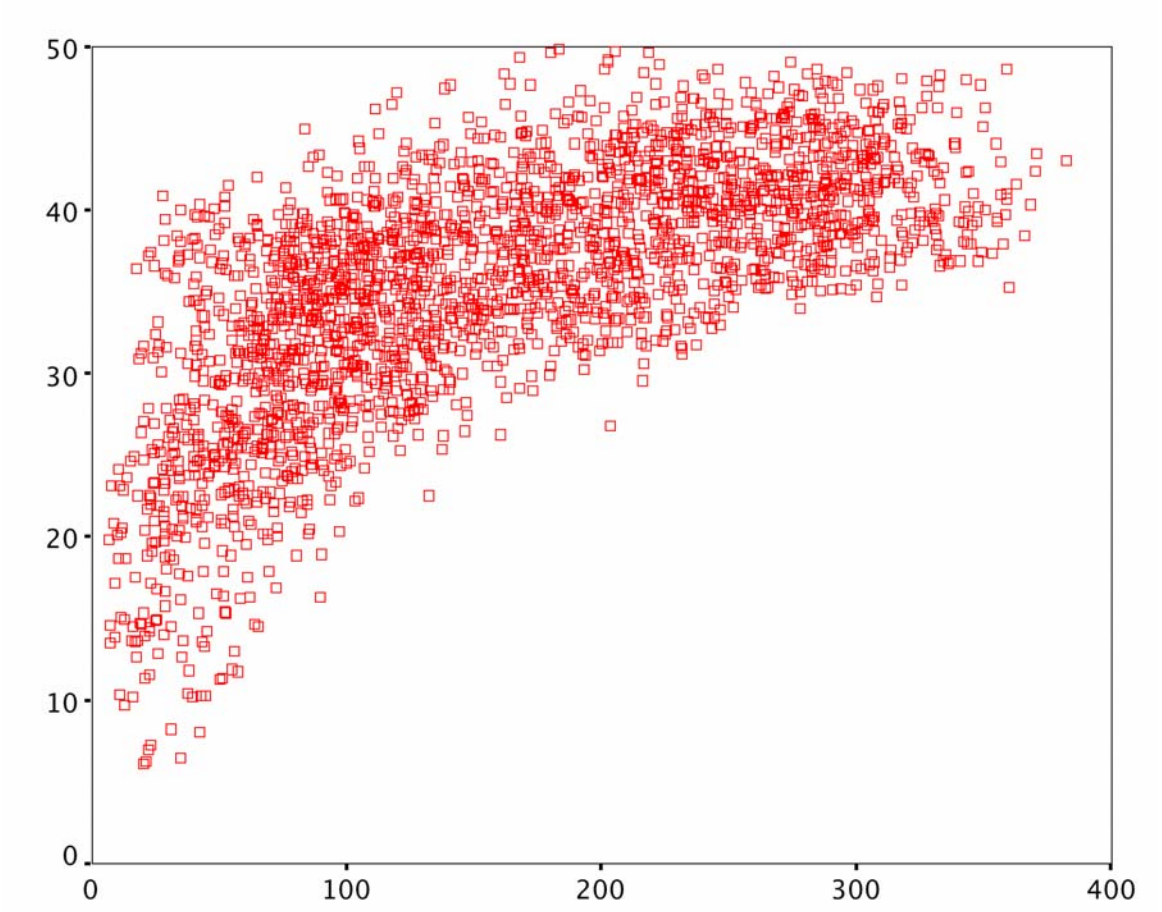

Figure 9. *This scatterplot shows the relation between pronunciational Levenshtein distances on the Y-axis and geographical distances on the X-axis.*

NOTE: Intended for colour reproduction on the web

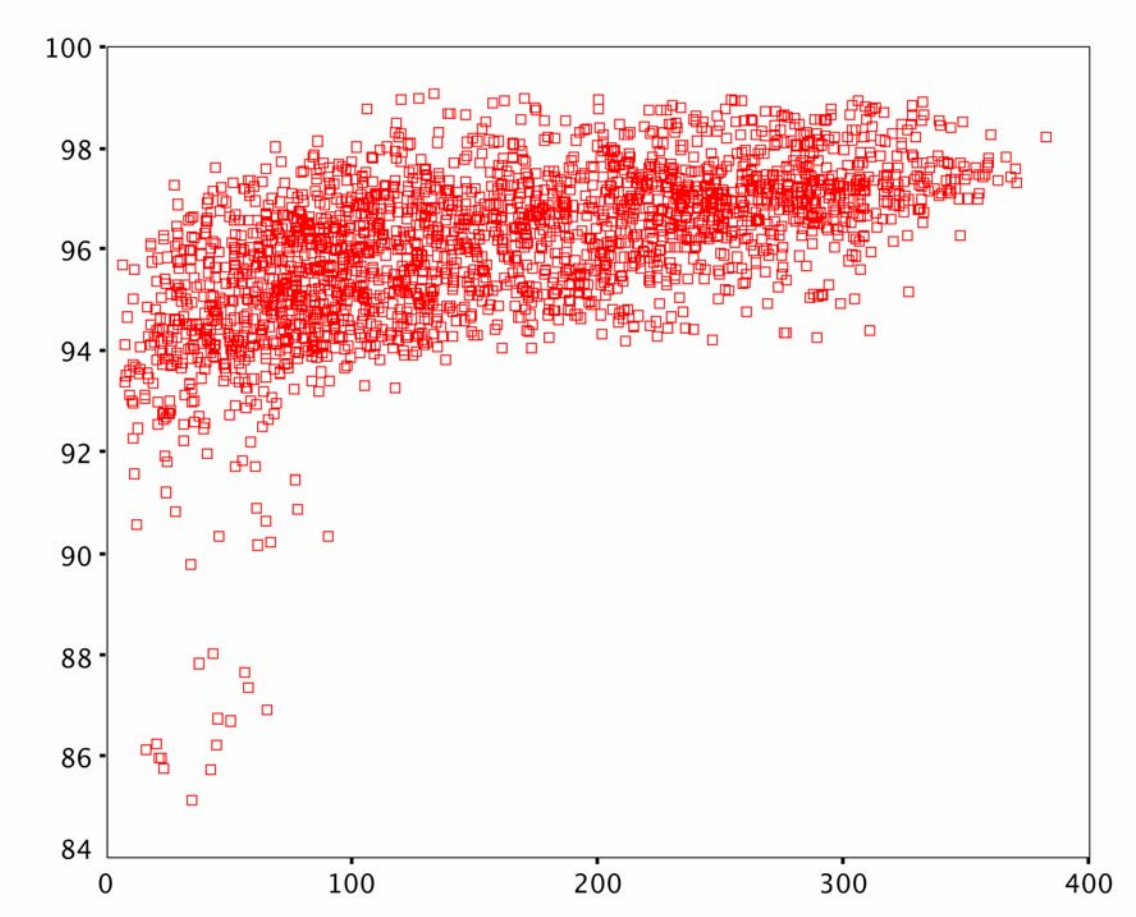

Figure 10. *This scatterplot shows the relation between lexical GIW distances on the Y-axis and geographical distances on the X-axis.*

NOTE: Intended for colour reproduction on the web

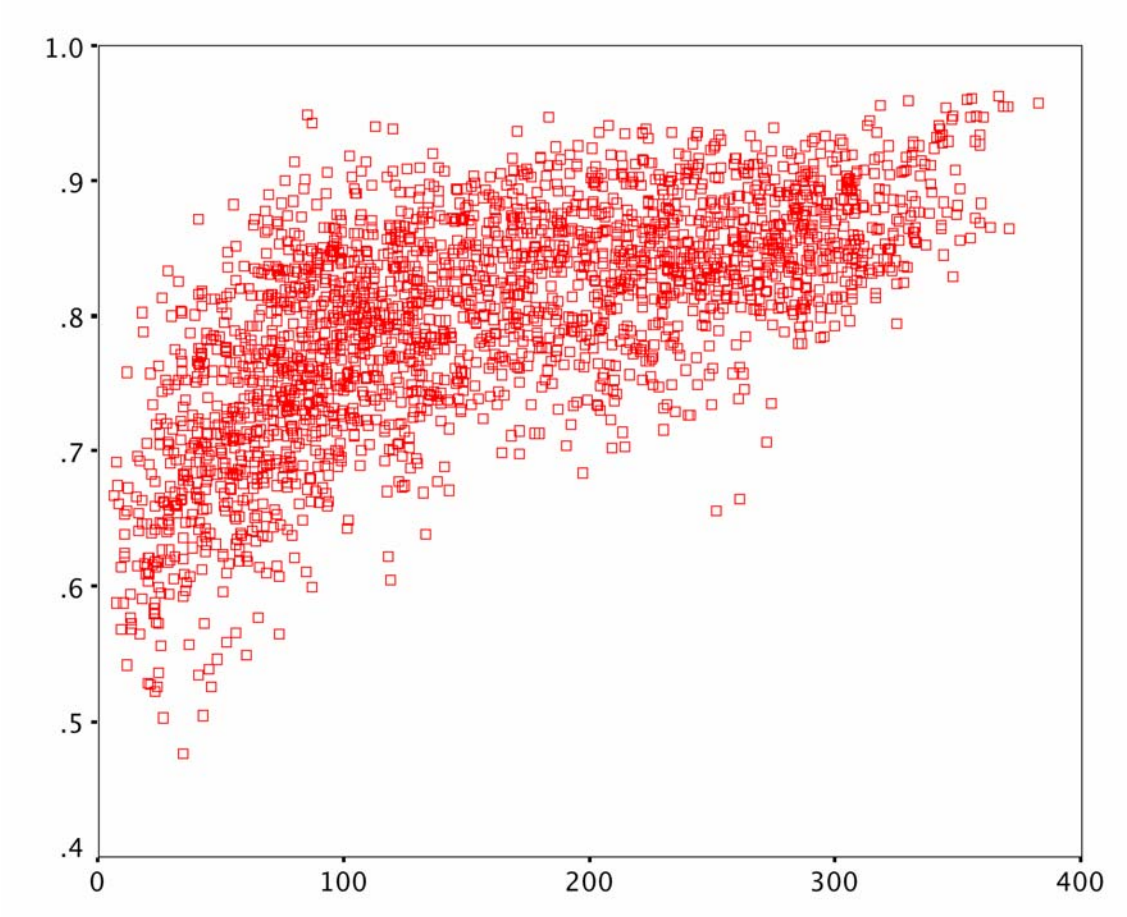

Figure 11. *This scatterplot shows the relation between syntactic GIW distances on the Y-axis and geographical distances on the X-axis.*

Tables

Context:	Complementiser of comparative if-clause										
Variables:	{ of, of dat, dat, as/of + V2, at, as, et }										
Examples:	*'t*	*lijkt*	*wel*	*<u>of</u>*	*<u>dat</u>*	*er*		*iemand*	*in*	*de*	*tuin*	*staat.*
	't	*lijkt*	*wel*	*<u>of</u>*		*er*	*staat*	*iemand*	*in*	*de*	*tuin.*	
Gloss:	'it	looks	[affirm.]	if	that	there	stands	someone	in	the	garden	stands'
Translation:	"It looks as if there is someone in the garden."											

Table 1. *Map 14b in SAND1 shows seven syntactic variables in the context of complementiser of comparative if-clause.*

Context:	Subject doubling 2 singular											
Variables:	{ V_{FINITE} __, __ V_{FINITE} __, C __, $C_{COMPARATIVE}$ __ }											
Examples:	*<u>Ge</u>*	*gelooft*	*<u>gij</u>*	*zeker*	*niet*	*dat*	*hij*	*sterker*	*is*	*as*		*<u>gij</u>.*
	<u>Ge</u>	*gelooft*	*<u>gij</u>*	*zeker*	*niet*	*dat*	*hij*	*sterker*	*is*	*as*	*-<u>ge</u>*	*<u>gij</u>.*
Gloss:	‘you_{weak}	believe	you_{strong}	certainly	not	that	he	stronger	is	than	you_{weak}	you_{strong}’
Translation:	“You do not seem to believe that he is stronger than you.”											

Table 2. *Map 54a in SAND1 shows four syntactic variables in the context of subject doubling 2 singular.*

Alignment	*[hart]*	*[ærtə]*	*Edit operation*	*Cost*
1	h		delete h	1
2	a	æ	substitute æ for a	1
3	r	r		0
4	t	t		0
5		ə	insert ə	1
				—
			Levenshtein distance between [hart] and [ærtə] =	3 / 5 = 0.6

Table 3. *String alignment and Levenshtein distance calculation between two pronunciations of the Dutch word hart 'heart'.*

concept	*Middelstum*	*Ommen*		*matches*		*comparisons*		*GIW*
vriend	kaməruˑt̬	kaməɾɔːt	1 -	(140	/	354)	=	0.60
schip	sxɪp	sxɪp	1 -	(353	/	360)	=	0.02
duwen	støˑt̬n̩	drʏk̬ŋ̍		-		-	=	0
								———
								0.62 / 3 = 0.21

Weigthed similarity between Middelstum and Ommen:

Table 4. *Weighted similarity calculation between two dialects based on word choices for the three concepts of vriend 'friend', schip 'ship' and duwen 'to push' using the gewichteter Identitätswert (GIW) measure.*

Linguistic level	*Number of variables (n_{var})*	*Cronbach's alpha (α)*
Pronunciation	125	0.97
Lexis	107	0.75
Syntax	106	0.94

Table 5. *Reliability coefficients (α) of our measurement results at the pronunciational, lexical and syntactic levels.*

Linguistic level 1	*Linguistic level 2*	*Correlation (r)*	*Explained variance (r^2 * 100)*
Pronunciation	Lexis	0.617	38%
Lexis	Syntax	0.496	25%
Syntax	Pronunciation	0.648	42%

Table 6. *Associations between aggregate pronunciational, lexical and syntactic distances.*

Linguistic level	*Correlation (r)*	*Explained variance (r^2 * 100)*
Pronunciation	0.685	47%
Lexis	0.575	33%
Syntax	0.669	45%

Table 7. *Correlations between geographical distances and pronunciational, lexical and syntactic distances.*

Linguistic level 1	*Linguistic level 2*	*Correlation (r)*	*Explained variance (r^2 * 100)*
Pronunciation	Lexis	0.374	14%
Lexis	Syntax	0.183	3%
Syntax	Pronunciation	0.350	12%

Table 8. *Associations between aggregate pronunciational, lexical and syntactic distances controlling for the influence of geography as an underlying factor*.

Linguistic level 1	*Linguistic level 2*	*Geographical influence*
Pronunciation	Lexis	39 %
Lexis	Syntax	63 %
Syntax	Pronunciation	46 %

Table 9. *The percentage of the correlation attributable to geography.*

$$\alpha = \frac{n_{\text{var}} \times r}{1 + (n_{\text{var}} - 1) \times r}$$

Equation 1. *Cronbach's alpha (α) is a function of the number of linguistic variables (n_{var}) and the average inter-correlation value among the variables (r).*

$$r = \frac{\sum_{i=2}^{n_{\text{var}}} \sum_{j=1}^{i-1} r\left(\text{var}_i, \text{var}_j\right)}{\frac{n_{\text{var}} \times \left(n_{\text{var}} - 1\right)}{2}}$$

Equation 2. *The average inter-correlation value (r) is based on all Pearson's correlation coefficients between each pair of variables r(var$_i$, var$_j$).*

$$\textit{Geographical influence} = \left(1 - \frac{\text{correlation controlling for influence of geography}}{\text{correlation not controlling for geography}}\right) * 100$$

Equation 3. *Influence of geography underlying the associations between the linguistic levels as a percentage.*