MEASURING LINGUISTIC VARIATION COMMENSURABLY

Martijn Wieling and John Nerbonne, University of Groningen

{m.b.wieling,j.nerbonne}@rug.nl

Abstract

The primary data on pronunciation variation – e.g., dialect atlas data – is often recorded incommensurably, i.e. in different ways in different atlases, and even in different ways within the same atlas when teams of fieldworkers and transcribers are involved. In particular these data collections differ in the detail in which pronunciations are recorded, using between 40 and 100 different basic symbols. This study shows that transcription system detail (understood in this sense) increases the linguistic distance measured and therefore must be regarded as a source of bias in assessing pronunciation differences and comparing them across languages. A method is therefore introduced to reduce transcription system complexity, even while retaining faithful assessments of aggregate pronunciation differences. The technique introduced is relevant when comparing within sets that have been transcribed very differently and also when comparing different dialectological datasets, e.g. with respect to the dependence of linguistic difference on geography.

Key words

Pointwise Mutual Information; Linguistic versus geographic distance; Transcription system reduction; Sublinear curve; Normalizing transcriber differences

1. Introduction and Motivation

This paper proposes a technique to remove one source of distortion that may confound the comparison of phonetic transcriptions, namely the use of different numbers of phonetic symbols. We first motivate the work by looking at dialectological theory and by demonstrating that the problem genuinely occurs in examining dialect atlases of different language areas. In the same section we introduce a second potential application for our technique, by noting a single atlas in which different fieldworkers used varying numbers of phonetic symbols. Second, we suggest a means of identifying pairs of symbols that are then merged. By applying the technique iteratively we reduce the size of the phonetic inventory. Third, we examine the result of applying this procedure to several datasets, verifying that the resulting dialectometrical analyses correlate well with the measures using the original phonetic inventory. Fourth, and finally, we examine in the conclusions and discussion section one apparent alternative and also discuss potential further confounds.

This paper is a contribution to a special issue on perception, production and attitudes concerning language variation, and its specific contribution to this topic is a means of comparing perceptions of variation (transcripts) even when they have been compiled on the basis of different segmental inventories (transcription alphabets).

Our primary motivation for attempting to remove a confound due to phonetic inventory size is theoretical, namely the ambition to examine the influence of geography on linguistic variation in different language areas, and in fact to quantify that influence in commensurable fashion. In this point, we should like to go beyond the consensus view *that* geography influences linguistic variation to a measurement of the strength of that relation and to models of the form it takes. Trudgill (1974) suggested one such form, namely a “gravity model” in which the tendency of varieties to share features decreases as an inverse square of their distance to one another and increases as a product of the population size speaking them. Several subsequent qualitative studies provide indications that the population size parameter of the model was sensible, and many others argued that further parameters were needed. Nerbonne and Heeringa (2007) review the literature on the gravity hypothesis in linguistics, and go on to develop a quantitative assessment of the gravity model, showing a sublinear curve mapping geographical distance to aggregate pronunciation distance in the Dutch Low Saxon dialect area, and incidentally finding little effect of population size. They note as well that Séguy (1971), in the first paper using dialectometrical techniques, examined the relation between lexical distance and geography, and likewise observed a sublinear relation. Nerbonne (2010) demonstrates that the sublinear curve of aggregate pronunciation distance is found not only for Dutch, but also for German, Norwegian, (Gabon) Bantu, Bulgarian, and American English (see Figure 1). This line of work suggests that linguistic variation is linked in a law-like fashion to geography, but we will need commensurable measures of linguistic distance if the model is to be formulated and tested more exactly.

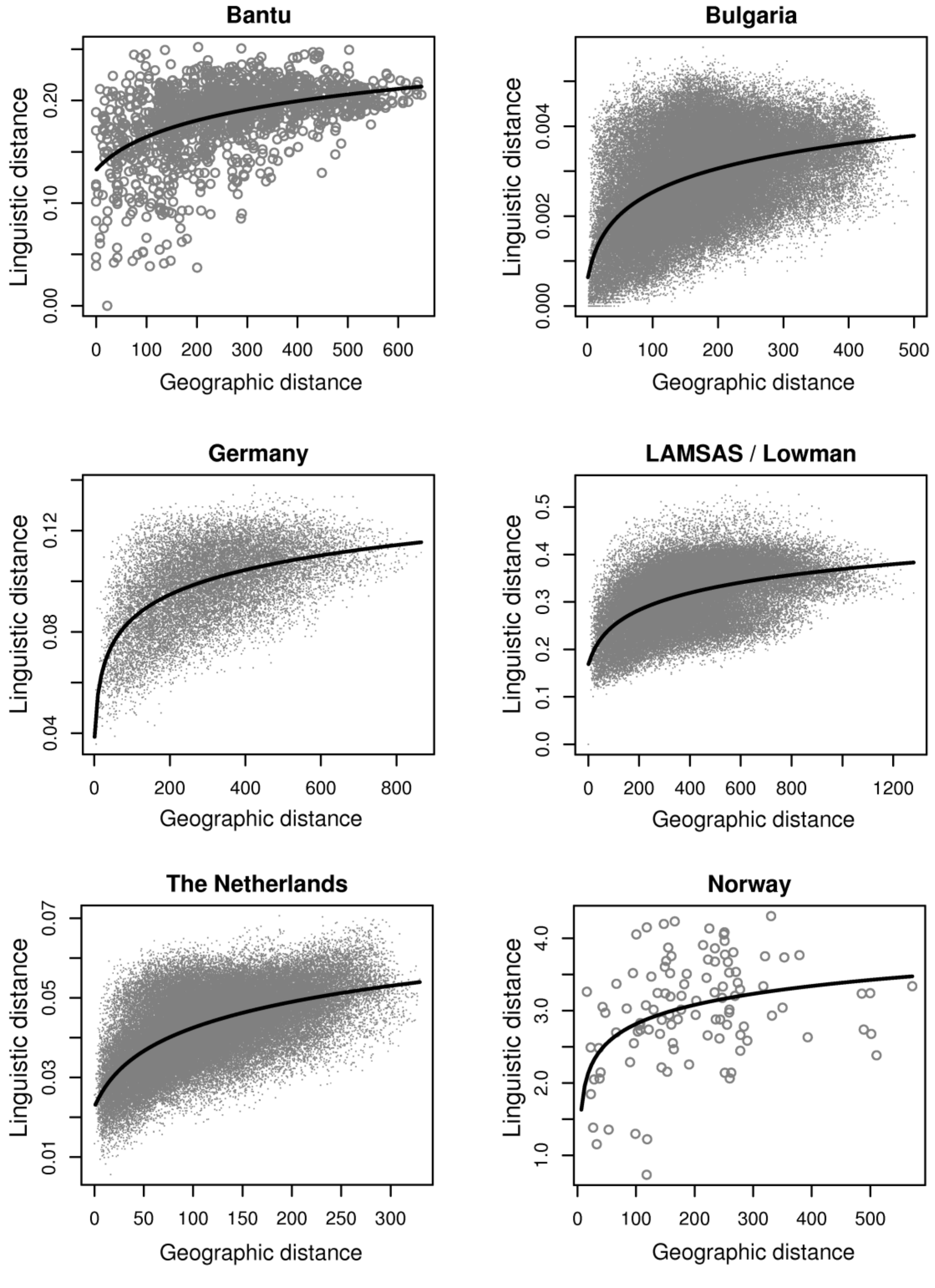

Figure 1. Aggregate pronunciation distance increases as a sublinear function of geography (taken from Nerbonne, 2010). The *x*-axes show the geographic distance in kilometers "as the crow flies" and the *y*-axes show *non-comparable* linguistic distance on the basis of pronunciation data. In each case the logarithmic curve was drawn. The *y*-axes in the different graphs are incommensurable because they come from different procedures, but, as the text argues, also because they are based on differently sized transcription systems. Note that the Norwegian scatter plot is based on 15 varieties and not on the 55 varieties used in this paper.

This paper aims to provide one prerequisite for a more exact formulation of the reliance of linguistic variation on geography. We would like to predict not only the general abstract form of the relation, but also its more specific parameters. Nerbonne (2010) conjectures that the dependence is logarithmic in the aggregate because the dynamic of diffusion is linear in its effect on individual linguistic features. But the logarithmic function that might describe this relation, $ling(x) = m\ log(geo(x)) + b$ has two parameters, *m,* the slope of the logarithmic curve, and *b*, the *y*-intercept. While *b* is presumably the degree of sub-dialectal variation, the slope *m* is a separate population parameter about which we would like to develop more exact hypotheses. Different populations may differ in the degree to which linguistic variation depends on geography, depending on the population density, the mobility of the population, the strictness of its social stratification, the length of time for which language standardizations of some sort exist, the length of time since the population became demographically (or politically) stable, or perhaps on other factors. To move beyond speculation about these factors, we need commensurable measurements across data sets from different languages.

Another motivation to improve commensurability in phonetic transcription is the problem of "field worker isoglosses" (Trudgill, 1982:241ff), the situation in which the analysis of a data collection reveals systematic differences which appear due to the field workers' preferences – or, worse, their errors. While many such cases are irreparable, perhaps even undetectable, one common sort of problem may be addressed, namely one in which phonetic transcription systems differ in the size of the phonetic inventory, or, put differently, the number of distinctions they express. In fact, the large-scale *Goeman-Taeldeman-van-Reenen-Project* (Goeman and Taeldeman, 1996) suffered precisely from this problem, as the dialectal pronunciations in the Netherlands were transcribed in more detail (using 82 sound segments) than the pronunciations in Flanders (using a subset of only 56 segments) and were therefore analyzed separately (Wieling et al., 2007). By applying the reduction method to this dataset, a single analysis of all data is possible.

In the following, we show that the measure of linguistic distance depends on the size of the segment inventory with which the pronunciation data is transcribed. Some data sets distinguish one hundred phonetic segments, others only forty. Given the demonstration that the size of the segment inventory can be influential, the main contribution of the paper is the development of a technique to reduce phonetic inventory size in a way that results in measures that still correlate highly with the original measures. We then report on the success of this technique, which we believe allows us to compare pronunciation distances between different languages validly.

2. Material

To illustrate the effectiveness of our reduction method with respect to investigating the influence of geography on linguistic variation language, we largely use the same linguistic material as used and discussed by Nerbonne (2010). The Bantu data set consists of phonetic transcriptions of 160 words in 53 locations which were collected in Gabon by researchers from the *Dynamique du Langue* lab (http://www.ddl.ish-lyon.cnrs.fr). The data set is described in detail and analyzed by Alewijnse et al. (2007). The Bulgarian data set contains phonetic transcriptions of 156 words in 197 locations and is part of the project *Buldialect* – measuring linguistic unity and diversity in Europe (http://www.sfs.uni-tuebingen.de/dialectometry). Houtzagers et al. (2010) offer a detailed overview and analysis of the data set. The German data set contains phonetic transcriptions of 186 words in 201 locations from the *Kleiner Deutscher Lautatlas – Phonetik* (http://www.uni-marburg.de/fb09/dsa). The data set is analyzed and discussed in detail by Nerbonne and Siedle (2005). The Dutch data set contains phonetic transcriptions of 562 words in 424 locations in the Netherlands. Wieling et al. (2007) selected the words from the *Goeman-Taeldeman-Van Reenen-Project* and also give a detailed overview and analysis of the data set. Our Norwegian data set differs from Nerbonne (2010) since we use all 55 locations as described by Heeringa (2004; Chapter 8), instead of the 15 locations used by Nerbonne (2010) and shown in Figure 1. The Norwegian data set contains phonetic transcriptions of 58 words of the fable “The North Wind and the Sun” (http://www.ling.hf.ntnu.no/nos) and is analyzed and explained in detail by Heeringa (2004; Chapter 8). In all data sets we mostly ignore diacritics and suprasegmentals and focus on the vowels and consonants in the pronunciations. As the transcription system in the American-English LAMSAS data set (included by Nerbonne, 2010) was highly complex and did not allow for straightforward removal of the diacritics, we did not include this data set in the current study. Hence, the results will be based on five dialect data sets: Bantu, Bulgarian, German, Dutch and Norwegian.

To illustrate our reduction method with respect to the “field worker isoglosses”, we use the same Dutch dataset as explained above, however now also including the 189 locations in Flanders (Wieling et al., 2007).

3. Methods

In the following we will show how we calculate and calibrate the linguistic distances to make them more comparable across different data sets.

3.1. Calculating linguistic distances

The linguistic distance between two locations is based on calculating the Levenshtein distance (Levenshtein, 1965) which measures the minimum number of insertions, deletions and substitutions to transform one string into the other. The following example shows that the Levenshtein distance of [bɪndən] and [bɛində], two Dutch dialectal pronunciations of *binden*, ‘to bind’, is 3.

bɪndən	insert ɛ	1
bɛɪndən	subst. i/ɪ	1
bɛindən	delete n	1
bɛində		
		3

This calculation corresponds with the following alignment:

b		ɪ	n	d	ə	n
b	ɛ	i	n	d	ə	
	1	1				1

The total linguistic distance of two locations is calculated by averaging the Levenshtein distance of all string pairs (i.e. pronunciations) available. Note that in general we enforce a syllabicity constraint to make sure vowels only align with vowels and consonants only with consonants. Nerbonne and Heeringa (2009) review work on measuring language variety differences using *inter alia* the Levenshtein distance.

3.2. Calibrating linguistic distances

Dialect data sets commonly differ in the number of segments (i.e. each is representative of an individual sound) which are used to transcribe the pronunciations. The size of the segment inventory, or segment set, will certainly influence the linguistic distances. To see this, consider the example calculation above. Besides the original segment set which includes /ɪ/ and /i/ assume there is a second segment set which does not distinguish these two sounds. It is obvious that the distance of the alignment above using the second set is reduced with respect to the distance assigned by the original set. In general, using fewer segments reduces the distinctions and will lower the linguistic distance. However, it is unclear how large the effect is, and if the effect is similar for every pair of places, regardless of their distance. To investigate the specific effect of the size of the segment inventory on distance measurements, we merge the segments of the inventory to obtain a smaller number of segments using two different transformation methods, one simple and one more sophisticated.

3.2.1. Simple transformation

The first transformation is extremely simple and consists of reducing each segment set to only two segments, one vowel and one consonant. All vowels reduce to a single vowel and all consonants reduce to a single consonant.

3.2.2. Advanced transformation

The second transformation is more sophisticated and aims to retain as much information as possible from the original dialect distances by reducing the number of segmental distinctions in each data set, but no further than necessary. Intuitively, we reduce the segmental inventory iteratively, mimicking the work a field worker might do if she were told that the segmental inventory she had used was one element too large. Then she would need to consider which distinction is least important among all the distinction made in the data set. As the data set with the minimum number of distinctions is the Bantu data set, we reduce all other segment sets to its number of segments (i.e. 42). We could have attempted this manually, but as defining the most similar sounds is highly subjective, we developed an automatic method based on the Pointwise Mutual Information (PMI) procedure introduced to dialectometry by Wieling et al (2009) to identify the most similar sounds. PMI is a similarity measure for categorical data inspired by information theory (Church and Hanks, 1990). The

Pointwise Mutual Information procedure determines the distance between every segment pair by assessing the relative frequency of every segment pair and comparing this to the relative frequency of the individual segments (i.e. the expected frequency of the segment pair if they are statistically independent). The method consists of the following steps (applied to each dataset having more than the minimum number of segments, in our case 42):

1. The Levenshtein algorithm with syllabicity constraint (see Section 3.1) is used to obtain the initial alignments;
2. For every sound segment pair, we calculate the PMI score:

$$\mathrm{PMI}(x,y) = \log_2(\, p(x,y) \,/\, p(x)p(y)\,)$$

Where:

- $p(x,y)$: the number of times x and y occur at the same position in two aligned strings X and Y, divided by the total number of aligned segments (i.e. the relative occurrence of the aligned segments x and y in the whole dataset). Either x or y can be a gap, representing an insertion or a deletion.
- $p(x)$ and $p(y)$: the number of times x (or y) occurs, divided by the total number of segment occurrences (i.e. the relative occurrence of x or y in the whole dataset). Dividing by $p(x)p(y)$ normalizes the empirical frequency, $p(x,y)$, with respect to the frequency expected if x and y are statistically independent.

The greater the PMI score, the more segments tend to co-occur in correspondences. Negative values indicate that segments do not tend to co-occur in correspondences, while positive PMI values indicate the opposite.

In contrast to Wieling et al. (2009), we ignore identical sound segment pairs in calculating the PMI score, since this improved the quality of the alignments (evaluated against the same gold standard as used by Wieling et al., 2009). Intuitively this also makes sense, since we are only interested in the distances (based on the PMI scores) of non-identical sound segment pairs relative to each other as the distance of identical sound segment pairs is always set to 0.

In order to assign a PMI score to a segment pair which does not occur (i.e. $p(x,y)$ equals 0), we add a very small value to $p(x,y)$, $p(x)$ and $p(y)$. This yields a very low PMI score for these segments, since the denominator is relatively high compared to the numerator. In addition, the effect on the PMI scores of segment pairs which do occur is negligible, since the original denominator and numerator values are relatively high compared to the small increase;

3. We convert the PMI score to a distance by subtracting it from 0 and scaling these values between a value slightly larger than 0 (only identical segments have a distance of 0) and 1. Consequently, a high PMI score yields a low distance and vice versa;
4. The Levenshtein algorithm based on the new segment distances is used to generate the new alignments. Thus instead of using a distance of 1 for an unequal segment pair (as used in the example alignment in Section 3.1), we use the calculated segment distance;
5. Steps 2, 3 and 4 are repeated until the segment distances (and therefore the alignments) remain constant.

After having determined the final segment pair distances, we identify the segment pair having the lowest distance and merge these two sounds (note that a gap is never merged with a sound). As long as the data set contains more sound segments than the desired number of segments (in our case 42), we run the Pointwise Mutual Information procedure anew on the simplified data set, to determine the next segment pair to be merged (i.e. the pair having the lowest distance). Note that since two merged segments are considered as a new individual segment, it is possible that this segment is involved in subsequent mergers, effectively merging more than two segments together.

As an example, consider the following. We have two Dutch dialects, where in each dialect two words are pronounced. The Dutch word *binden* (to bind) is pronounced as [bɪndən] in dialect *A* and [bɛində] in dialect *B*. The Dutch word *heet* (hot) is pronounced as [heɪt] in dialect *A* and [heit] in dialect *B*.

Initially the alignments are based on the Levenshtein algorithm with the syllabicity constraint. Because of this, [bɪndən] and [bɛində] align in two ways (both having a distance of 3):

b		ɪ	n	d	ə	n
b	ɛ	i	n	d	ə	
	1	1				1

b	ɪ		n	d	ə	n
b	ɛ	i	n	d	ə	
	1	1				1

The words [heɪt] and [heit] align in only one way (having a distance of 1):

h	e	ɪ	t
h	e	i	t
		1	

After the initial alignments are generated, the first run of the Pointwise Mutual Information procedure determines the distance between [ɛ] and [ɪ] and the distance between [i] and [ɪ]. It is clear that [ɛ] and [ɪ] align only once and [i] and [ɪ] align twice. Since the frequency of [i] (3) is less than twice the frequency of [ɛ] (2), the increase of the numerator for [i] and [ɪ] is not compensated by the increase of the denominator (relative to [ɛ] and [ɪ]) and hence the PMI score for [i] and [ɪ] will be higher than the PMI score for [ɛ] and [ɪ].[1] Consequently, the distance between [i] and

[1] Note that for the sake of clarity this explanation is somewhat simplified as in the actual algorithm each word (and not each alignment) is assigned the same importance. The general result, however, remains the same.

[ɪ] will be decreased, and in particular made lower than the distance between [ɛ] and [ɪ] so that in the second run of the PMI procedure the second alignment for *binden* will not be generated anymore (since the Levenshtein algorithm only yields the alignment with the minimum distance). After the second run, the PMI scores will not change anymore and the calculated segment distances are used to determine the segment pair which should be merged. In our example segments [i] and [ɪ] will be merged, as there is no other segment pair involving non-identical segments present (the gap is never merged with a sound).

Note that a slight change to this procedure is necessary when it is used to compensate for transcriber differences (e.g., as present in the Dutch dialect dataset). In that case not necessarily the two most similar segments are merged, but a segment used only by one group of transcribers (but not the other) is merged with the most similar sound used by both groups of transcribers. This process is repeated until all segments are used by both groups of transcribers.

3.3. Obtaining the final linguistic distances

After reducing the segment set, we determine the linguistic distances by applying the Levenshtein distance (with syllabicity constraint) to the adapted transcriptions as described earlier. Because different data sets do not necessarily use words with similar lengths, we normalize the Levenshtein distance between two strings by dividing it by the alignment length of the longest transcription (i.e. in the example alignment above, the distance would be 3/7 as there are 7 positions in the alignment).

4. Results

Figure 2 shows the relation between the geographical distance and the logarithm of the linguistic distance (hence the straight lines) (i) based on the original number of segments in every data set, (ii) based on the simple transformation (reduction to two segments) and (iii) based on the advanced transformation (reduction to 42 segments). All slopes were significant ($p < 0.001$) and the association strength ranged between $r = 0.13$ (Norway, 2 segments) to 0.67 (The Netherlands, 42 segments). We can clearly see in Figure 2 that reducing the number of segments decreases both the intercept and the slope.

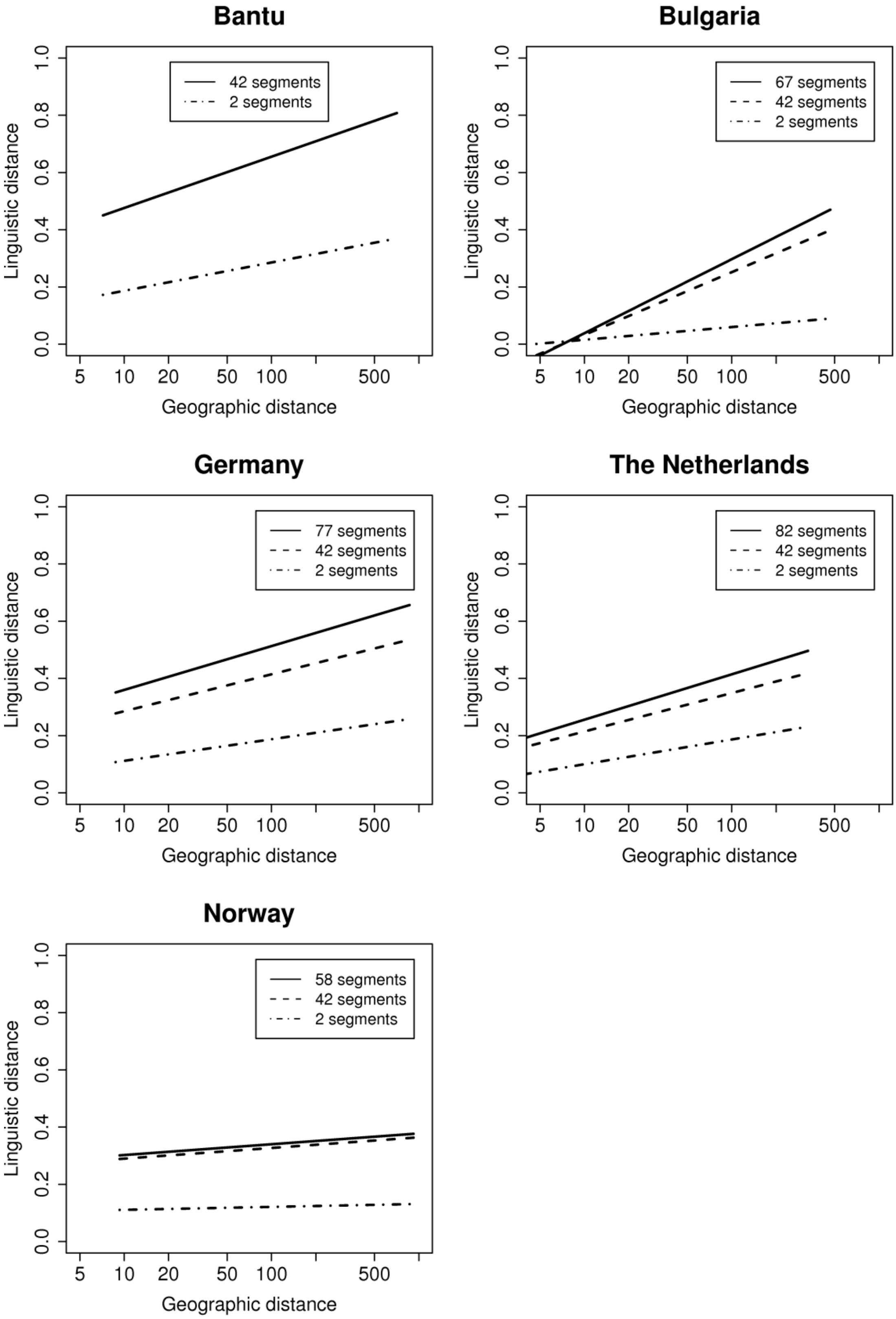

Figure 2. Influence of the segment reduction on the sublinear relationship between geographical distance "as the crow flies" and linguistic distance. Only the curve is shown, the individual data points are omitted for clarity. Note the logarithmic scale of the *y*-axis in all graphs.

To see the effect of the reduction on the original distances, the correlation coefficients between the distances measured using the original segment inventory and the distances measured using the reduced segment inventories are shown below. All correlations are significant ($p < 0.001$).

	Correlation coefficient *r* with reduced data sets	
	42 segments	2 segments
Bantu (42 segments)	1	0.85
Bulgaria (67 segments)	0.98	0.74
Germany (77 segments)	0.96	0.80
Norway (58 segments)	0.995	0.77
The Netherlands (82 segments)	0.97	0.77

It is clear from the high correlations between the original distances and the distances based on 42 segments that most distinctions in the original data set are retained in the transformed data set. It is also clear that the reduction is effectively a linear transformation.

However, when looking at the reduction to two segments, much more variation is lost. To illustrate this effect more precisely, Figure 3 shows a visualization of the similarity between Dutch dialects. Darker lines connect locations which are linguistically more similar. We clearly see the high similarity between the maps based on the original segment set (top-left) and the segment set consisting of 42 segments (middle). However, the map in the bottom-right (based on two segments) is less similar since it shows darker lines (larger distances) as well as increased contrasts (e.g., the diagonal dark line from the northwest to the southeast which divides the map effectively in a light top half and a dark bottom half is not so clear in the original map). We therefore judge the advanced transformation as the better option.

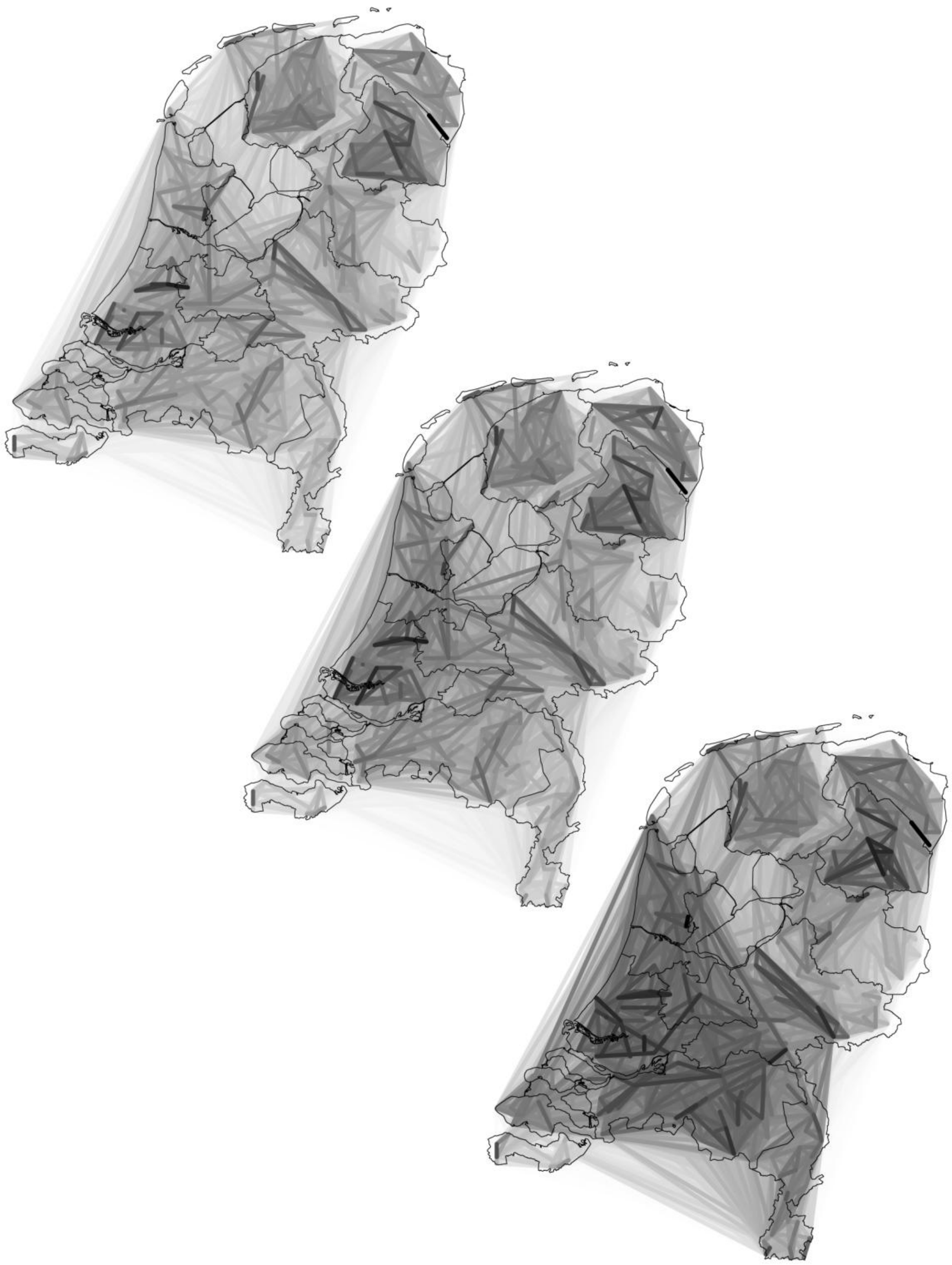

Figure 3. Similarity between Dutch dialects. Darker lines connect dialects which are more similar. The top-left map shows the distances based on the original segment set (82 segments), the middle map shows the distances based on the segment set consisting of 42 segments and the bottom-right map shows the distances based on two segments.

4.1. Segment mergers

We manually verified that the advanced transformation mostly merges sound segments which are linguistically similar. To illustrate that the automatic segment reduction method indeed performs very well, the table below shows the segments which are merged for the German data set (the 32 segments which were not merged are omitted from the table). We can clearly see that the ten segment groups (containing 45 segments) generally consist of similar sounds. For example, the fourth group shows the (sensible) merger of several fricatives in the alveolar and alveo-palatal region.

/ʊ/ /u/ /ɯ/ /ʉ/
/l/ /ʎ/ /ɮ/ /ɬ/
/t/ /d/ /θ/
/s/ /z/ /ʃ/ /ʂ/ /ʒ/ /ç/
/k/ /c/ /ɟ/ /q/ /ɢ/ /g/
/x/ /χ/
/β/ /v/ /ʍ/ /ʋ/ /w/ /f/ /ɸ/ /p/ /b/
/ŋ/ /ɴ/ /ɰ/ /ɲ/
/h/ /ɦ/ /ɥ/
/ɾ/ /r/ /ʁ/ /ɹ/

4.2. Reducing transcriber differences in the Dutch dataset

Applying the reduction method to the *Goeman-Taeldeman-Van Reenen-Project* data gives us the opportunity to analyze the Dutch dialect distances in the complete area, instead of separately for the Netherlands and Flanders (Wieling et al., 2007). Figure 4 visualizes the dialect areas based on the reduced segment set of 56 segments using multidimensional scaling (MDS; See Heeringa, 2004:156). While the new dialect distances[2] correlated highly ($r = .99$, $p < 0.0001$) with the original dialect distances, we are now more confident that the observed differences between the Dutch and

[2] Note that, in contrast to the procedure described in Section 3.3, distances between pronunciations were not normalized and based on the PMI distance between the individual segments as this was found to be one of the best methods to assess dialect distances in a single dataset (Wieling et al., 2009).

Flanders dialects are not caused by transcriber differences due to different-sized segment sets. Apart from the observed similarities and differences already described by Wieling et al. (2007) in the individual countries, we see some differences between neighboring dialects in different countries. In the Limburg area (located in the southeast) dialects seem relatively similar, irrespective of the country in which they are located. More surprisingly, however, the western part of Flanders (with the greenish tint) appears to have some similarities with the northeastern part of the Netherlands (Low Saxon).

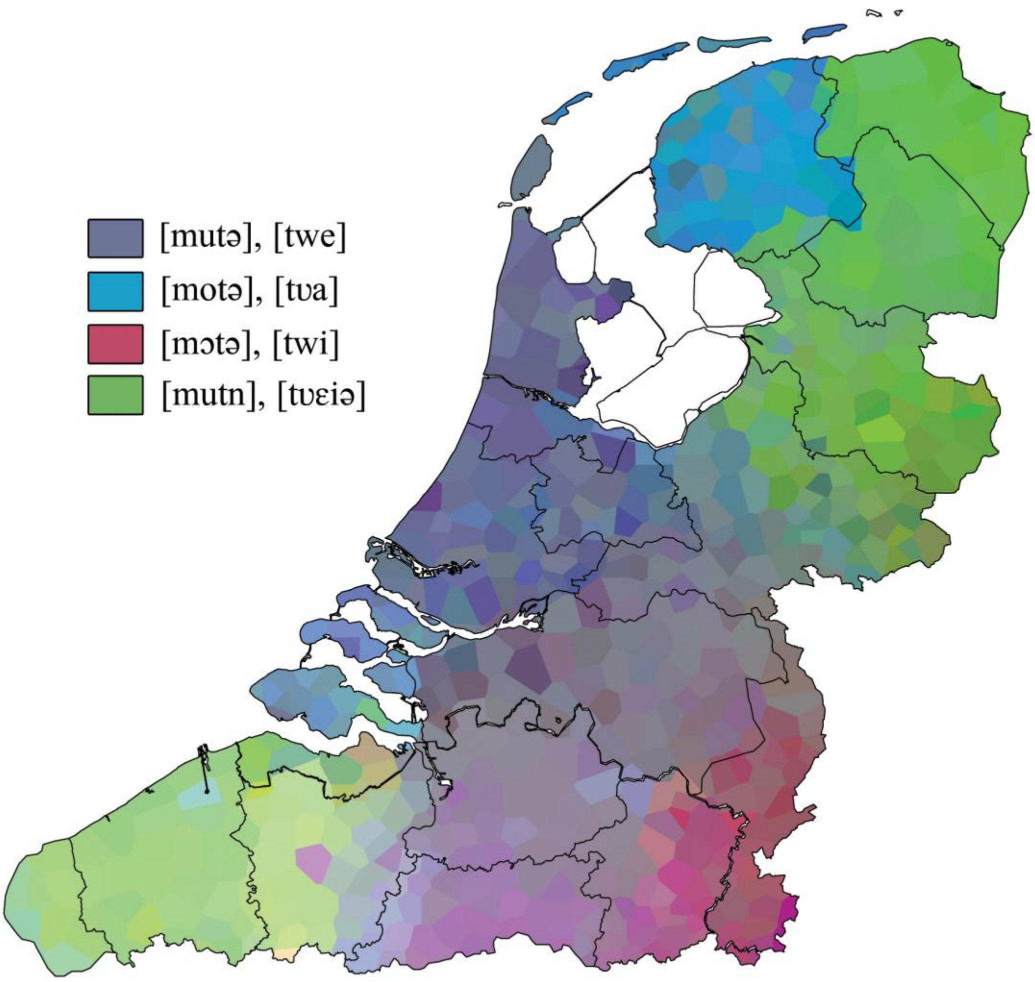

Figure 4. MDS plot of Dutch dialect distances based on 56 segments. The legend shows the approximate pronunciations of the words *moeten* (‘must’) and *twee* (‘two’) in the areas corresponding to the colors. Note the similarity between the Low Saxon area (northeast of the Netherlands) and the area in the western part of Flanders.

5. Conclusions and Discussion

In the previous sections we showed that the number of segments used for recording pronunciation data certainly has an effect on the pronunciation distances measured. With respect to the distribution of linguistic distance as a function of geographical distance, we see effects on both the intercept and the slope of the sublinear curve. It is therefore clear that when attempting to interpret the individual slopes, it is a *sine qua non* that the segment sets are reduced to a comparable size. We examined two segment-set transformations, one simple transformation which reduced all phonetic distinctions to just vowel versus consonant, and one more elaborate transformation which reduced segment inventories iteratively, by repeatedly selecting those two segments with the highest similarity as measured by the information-theoretical pointwise mutual information score.

The simple transformation resulted in segment sets of the same size (i.e. 2 segments) in which obviously a lot of variation got lost. The advanced transformation retained more of the variation of the original data set by removing fewer segmental distinctions. We concluded therefore that the advanced transformation is more suitable for the task of obtaining commensurable measurements of pronunciation difference on data sets which have been transcribed using segment inventories of different sizes.

Two anonymous reviewers suggested a simpler transformation, normalizing so that $x_{norm} = (x\text{-}min) / (max\text{-}min)$, for raw score x. In that transformation the sound segment set is not reduced, but every aggregate dialect distance in a data set is scaled between 0 and 1. While this is a sensible approach in some cases when one wishes to scale the linguistic distances, it is unsuitable for our purposes. As noted in the introduction, one area in which we wished to apply the correction was where we suspected “field-worker isoglosses”, but the application assumes a minimum and a maximum distance. In applying the correction one might use the minima and maxima of comparisons in the entire set of pairwise comparisons, or one might choose to restrict one’s attention to the comparisons between the sites collected by the suspect field worker on the one hand and the sites collected by the non-suspect one the other. In many cases there will not be enough material to be certain that the minimum and maximum are being chosen representatively. For example, in the set we examined, consisting of the Netherlands and Flanders, the set of initially incommensurable pairs of sites are also the sites at the greatest distance from one another – for which one would also expect the greatest linguistic distances. The simple min-max scaling could never cope with this sort of situation. So we are skeptical about applying this technique.

When we turn to scaling for the purpose of investigating a general model, the same problems arise, in addition to some others. One additional problem is that all languages areas would scale from zero to one, while we suspect that the left end of the geography may show interesting differences which we would interpret as different levels of subdialectal variation. A second problem becomes obvious when we consider specific cases, and in particular the slopes of the logarithmic curve which is to be explored further. In Figure 2, the relationship between linguistic and geographical distances for Norway is quite flat. Scaling the linguistic distances between 0 and 1 would increase the slope enormously (since the geographical distances remain the same). In fact we suspect that we would normally obtain a slope 1/max.-geo.-dist., since linguistic distance normally rises monotonically with respect to geographic distance. This would not be a rewarding *explicandum*!

It is clear that there may be further sources of bias in comparing pronunciation transcriptions of different languages. We have introduced a means of controlling for different sizes of segment inventories in this paper, but the segment inventories may also have radically different constitutions, as well. For example, while Germanic languages have complex vowel phoneme systems with twenty or so vowels and diphthongs (Roach, 2000), Slavic languages may have as few as five vowels (i.e. /i/, /e/, /a/, /o/ and /u/). Slavic languages, however, have more complicated sets of consonants, distinguishing two variants of most consonants, one with and one without palatalization (Hamilton, 1980:18). We conjecture that using segment inventories of the same size but of different composition (in the sense just illustrated) need not skew measurements to the same extent as using segment inventories of different sizes does, but we concede that this point deserves attention as we proceed to compare dialect distances from different language areas. It may be worth noting that, if the composition of the segment inventories does indeed bias measurements, then the program which seeks to compare measurements across different languages will be faced with a very difficult problem, as most data collections represent phonemic distinction reliably, and phonemic inventory size is known to vary a good deal (Hay and Bauer, 2007).

Once we are confident that we are in possession of a measure of pronunciation distance that yields commensurable scores across different languages, we are in a position to remove some “field-worker isoglosses”, as we demonstrate above. We are also in a position to address questions about the different factors influencing the distribution of varietal differences with respect to geography in different languages. These may concern population density, population mobility, the degree of social stratification, the length of time since language standardization, the length of time since the population has become demographically (or politically) stable, or perhaps other factors. In this paper we have attempted to lay the groundwork for investigation of such factors that can move beyond speculation, by providing a technique for obtaining commensurable measurements across data sets from different languages.

We have not examined the measurement of geographical distances in this paper, but it is also clear that this topic deserves careful consideration as we do not imagine that distance directly influences the tendency of language varieties to differ, but rather, indirectly, in reducing the likelihood of social contact. We have measured geographical distance very simply to-date, using the “as-the crow-flies” great circle distance on the earth’s surface, with no attention to specific aspects of geography. It is clear that the great circle distance is not optimal as it disregards natural barriers limiting individuals in their mobility (Handley et al., 2007). A possible improvement in this area would be to use travel distances or travel duration (automatically) calculated using a travel planner (but this may have limited effectiveness; see Gooskens, 2005), or alternatively use friction matrices indicating how difficult it is to travel from one location to the next (Handley et al., 2007).

Acknowledgements

We thank the anonymous reviewers for their helpful and generous comments and suggestions on the previous draft of this paper. Peter Kleiweg implemented a second version of the PMI-based segment weighting which raised valuable questions.

References

ALEWIJNSE, Bart, John NERBONNE & Lolke VAN DER VEEN (2007) “A Computational Analysis of Gabon Varieties”, in: Petya Osenova et al. (eds.) *Proceedings of the RANLP Workshop on Computational Phonology Workshop at Recent Advances in Natural Language Processing*, Borovetz, 3-12.

CHURCH, Kenneth & Patrick HANKS (1990) “Word association norms, mutual information, and lexicography”, *Computational Linguistics*, 16(1), 22–29.

GOOSKENS, Charlotte (2005) “Traveling time as a predictor of linguistic distance”, *Dialectologia et Geolinguistica*, 13, 38-62.

GOEMAN, Anton & Johan TAELDEMAN (1996) "Fonologie en morfologie van de Nederlandse dialecten. Een nieuwe materiaalverzameling en twee nieuwe atlasprojecten." *Taal en Tongval* 48, 38-59.

HAMILTON, William (1980) *Introduction to Russian Phonology and Word Structure*, Columbus: Slavica.

HANDLEY, Lori, Andrea MANICA, Jérôme GOUDET & François BALLOUX (2007) “Going the distance: human population genetics in a clinal world”, *TRENDS in Genetics*, 23(9), 432-439.

HAY, Jennifer & Laurie BAUER (2007) “Phoneme inventory size and population size”, *Language*, 83(2), 388-400.

HEERINGA, Wilbert (2004) *Measuring Dialect Pronunciation Distances using Levenshtein Distance*, PhD thesis, Rijksuniversiteit Groningen.

HOUTZAGERS, Peter, John NERBONNE and Jelena PROKIĆ (2010) “Quantitative and Traditional Classifications of Bulgarian Dialects Compared”, *Scando-Slavica*, 56(2), 29-54.

LEVENSHTEIN, Vladimir (1965) “Binary codes capable of correcting deletions, insertions and reversals”, *Doklady Akademii Nauk SSSR*, 163, 845-848.

MANNI, Franz, Wilbert HEERINGA, Bruno TOUPANCE & John NERBONNE (2008) “Do Surname Differences Mirror Dialect Variation?”, *Human Biology*, 80(1), 41-64.

NERBONNE, John (2010) “Measuring the Diffusion of Linguistic Change”, *Philosophical Transactions of the Royal Society B: Biological Sciences*, special issue from the "Cultural and Linguistic Diversity" conference, AHRC Centre for Evolution of Cultural Diversity, London, Dec. 9-13, 2008.

NERBONNE, John & Wilbert HEERINGA (2007) "Geographic Distributions of Linguistic Variation Reflect Dynamics of Differentiation", in S. Featherston and W. Sternefeld (eds.), *Roots: Linguistics in Search of its Evidential Base*, Berlin: Mouton De Gruyter, 267-297.

NERBONNE, John & Wilbert HEERINGA (2009) "Measuring Dialect Differences", in J. E. Schmidt and P. Auer (eds.), *Language and Space: Theories and Methods* in series *Handbooks of Linguistics and Communication Science*. Berlin: Mouton De Gruyter, Chapter 31, 550-567.

NERBONNE, John & Christine SIEDLE (2005) "Dialektklassifikation auf der Grundlage aggregierter Ausspracheunterschiede", *Zeitschrift für Dialektologie und Linguistik*, 72(2), 129-147.

ROACH, Peter (2000) *English Phonetics and Phonology: a Practical Course*, Cambridge: Cambridge University Press.

SÉGUY, Jean (1971) "La relation entre la distance spatiale et la distance lexical", *Revue de Linguistique Romane*, 35(138), 335-357.

TRUDGILL, Peter (1974) "Linguistic Change and Diffusion. Description and Explanation in Sociolinguistic Dialect Geography", *Language in Society*, 2, 215-246.

TRUDGILL, Peter (1982) "The contribution of Sociolinguistics to Dialectology", *Language Sciences* 4(2), 237-250.

WIELING, Martijn, Wilbert HEERINGA & John NERBONNE (2007) "An Aggregate Analysis of Pronunciation in the Goeman-Taeldeman-van Reenen-Project Data", *Taal en Tongval*, 59(1), 84-116.

WIELING, Martijn, Jelena PROKIĆ & John NERBONNE (2009) "Evaluating the pairwise string alignment of pronunciations", in Lars Borin and Piroska Lendvai (eds.), *Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education*, Workshop at the 12th Meeting of the European Chapter of the Association for Computational Linguistics. Athens, 30 March 2009, 26-34.