SOME FURTHER DIALECTOMETRICAL STEPS

John Nerbonne, Jelena Prokić, Martijn Wieling & Charlotte Gooskens

Center for Language and Cognition
University of Groningen

Abstract

This article surveys recent developments furthering dialectometric research which the authors have been involved in, in particular techniques for measuring large numbers of pronunciations (in phonetic transcription) of comparable words at various sites. Edit distance (also known as Levenshtein distance) has been deployed for this purpose, for which refinements and analytic techniques continue to be developed. The focus here is on (i) an empirical approach, using an information-theoretical measure of mutual information, for deriving the appropriate segment distances to serve within measures of sequence distance; (ii) a heuristic technique for simultaneously aligning large sets of comparable pronunciations, a necessary step in applying phylogenetic analysis to sound segment data; (iii) spectral clustering, a technique borrowed from bio-informatics, for identifying the (linguistic) features responsible for (dialect) divisions among sites; (iv) techniques for studying the (mutual) comprehensibility of closely related varieties; and (v) Séguy's law, or the generality of sub-linear diffusion of aggregate linguistic variation.

Keywords: *Phonetic alignment, multi-alignment, spectral clustering, mutual comprehensibility, linguistic diffusion*

1. Introduction

The dialectometric enterprise (Goebl 1982) need not be introduced in a paper in this volume, which is largely dedicated to presenting its contemporary form. We shall only take care to note points at which there may not yet be consensus. The authors of the present contribution view the introduction of exact techniques into dialectology as an extension of the *methods* available within the discipline but do not suggest that dialectology change in its central questions, in particular, questions on the nature of the influence that geography has on language variation. More exact techniques, and especially computational techniques, serve to broaden the empirical base that dialectology can effectively build on, improve the replicability of data analysis techniques, and enable more abstract questions to be addressed with empirical rigor. Nerbonne (2009) elaborates on these opportunities for dialectometry.

We have collaborated especially on developing and applying measures of pronunciation distance derived from edit distance or Levenshtein distance. The distance be-

tween two transcriptions t_1 and t_2 is defined as the sum of costs associated with the least costly set of operations needed to transform t_1 into t_2, and typically one makes use of only three operations, substitution, insertion and deletion. A by-product of the calculation is an alignment of the two transcriptions, where the segments which have been involved in the operations are written one above the other:

[æ ə f t ə n ʉ n] 'afternoon', Georgia (LAMSAS)
[æ f t ə r n u n] 'afternoon', Pennsylvania

In the example here, we see that schwa [ə] corresponds with ø (the null segment), [ø] with [r], and [ʉ] with [u]. The correspondences are extracted automatically from the digitized transcriptions. See Nerbonne & Heeringa (2009) and references there for more extensive explanation and illustration. The technique was developed with a view to analyzing the sort of data typically found in dialect atlases, in particular where the pronunciation of a single word or phrase is elicited, recorded and ultimately transcribed for later analysis. The "Salzburg school" of dialectometry typically devotes a good deal of time to manually extracting correspondences from dialect atlases (a phase Goebl refers to as *Taxierung*, roughly 'appraisal'), a step in methodology which the application of edit-distance largely obviates. So this is a point at which we feel there has been a contribution to dialectometric technique. Nerbonne and Heeringa (2009) reviews work devoted to pronunciation difference measurement.

We see an additional point in favor of the deployment of edit distance, namely that it provides a broader view of the variation in typical atlas data because it incorporates *entire* pronunciations in its measurements instead of relying on the analyst's choice of variables to extract. This means that dialectometrists using edit-distance measures of pronunciation are less likely to fall prey to choosing their variables in a way that biases results. Since we, too, normally deal with dialect atlas data, we are not shielded from the danger of biased data collection completely, but whereas other approaches typically focus on the set of variables the atlas was designed to assay, the edit-distance measurements incorporate all the pronunciations, not merely a one or two segments per word.

In addition we note that pronunciation distance is a true metric, obeying the relevant mathematical axioms. This means that the distances assayed are all non-negative, zero only in case of identity, and that they satisfy the so-called triangle inequality: for all t_1, t_2, there is no t' such that:

$$d(t_1,t') + d(t',t_2) < d(t_1,t_2)$$

The fact that genuine measurements are made instead of predications of identity vs. non-identity has a consequence that less data is required for reliable assessment of the relations among sites in a dialect landscape. Heeringa (2004) shows that about 30 transcriptions per site yields consistent measurements for Dutch (Cronbach's $\alpha \geq 0.8$). One typically needs 100 or more items to attain comparable levels of consistency when comparing at a categorical level. The consistency measure is empirical, and therefore must be recalculated on each new data set, but Heeringa's result has been confirmed in several other data collections. We typically work with sets of 100 words per variety, because we are not merely interested in the distance relations

at an aggregate level, but also in the concrete segmental correspondences which the analysis also yields. These may not turn up in small data sets.

Although we share the general interest in finding groups of most similar sites in language areas, we are also sensitive to the "French" worry that dialect areas may lead ephemeral existences. Following Goebl and others, we have used clustering regularly as a means of seeking groups in our data. We have also been sensitive to the charges of the statistical community that no clustering technique works perfectly (Kleinberg 2003), and have therefore explored versions of clustering that remove the instability inherent in the technique (i.e. the problem that very small differences in input data may lead to major differences in output clusterings), using both bootstrap clustering and "noisy" clustering, a technique we developed independently (Nerbonne et al. 2008, Prokić & Nerbonne 2008). As the last reference demonstrates, there is good reason to be wary of cluster analyses even when they are accompanied by stability-enhancing augmentations.

Stimulated by the difficulties of finding groups reliably using clustering, we have emphasized the use of multi-dimensional scaling (MDS) in analyzing the results of dialectometric measurements (Nerbonne, Heeringa and Kleiweg 1999). Black (1973) introduced MDS to linguistics, where it has been applied sporadically since. It is a remarkable fact that the very complicated data of linguist variation, potentially varying in hundreds of dimensions (one for each point of comparison) may normally be faithfully rendered in just two or three dimensions. But this remarkable fact means that it is possible to visualize language variation effectively in scatterplots of two dimensions that make use of a single color dimension, e.g., grey tones to plot the third. Nerbonne (to appear, b) reviews approaches to mapping aggregate linguistic variation, focusing especially on techniques involving MDS.

The remainder of this paper sketches very recent work building on the lines set out above. Section 2 aims to answer a question that we have been wrestling with since the very earliest work applying edit distance to variation data (Nerbonne & Heeringa 1997), namely how to weight operations in such a way that phonetically natural substitutions (and also insertions and deletions) cost less than unnatural ones. Focusing on substitutions, we should prefer to see that the substitution of [i] for [e] should cost less than the substitution of [i] for [a]. Section 3 reports on multi-alignment, the attempt to extend the process of aligning two strings to that of aligning many, potential hundreds. This is potentially very interesting in historical linguistics, where regular sound correspondences play an important role. Section 4 summarizes work on a new effort to identify not only the important groups of varieties, but also —and simultaneously— the linguistic basis of the group. Section 5 reports on efforts to link the work we have done on pronunciation distance to the important issue of the comprehensibility of varieties, effectively asking whether phonetically distant varieties are also less comprehensible. Section 6 attempts to use the dialectometric perspective to view language variation from a more abstract perspective, and asks whether this might allow the formulation of more general laws of linguistic variation. Finally, the concluding section attempts to identify areas which are promising for dialectometric inquiry in the near future.

2. Inducing Segment Distances Empirically

Many studies in dialectometry based on the Levenshtein distance use very simple segment distances, only distinguishing vowels and consonants; then substitution costs for each pair of vowels (or pair of consonants) are the same (e.g. see Wieling et al. 2007). In such studies, the substitution of [e] for [i] has the same weight as a substitution of [e] for any other vowel. While it would clearly be rewarding to use linguistically sensible segment distances, obtaining these is difficult and seems to require some arbitrary decisions. If this seems surprising given the relatively simple charts of phonetic symbols found e.g. in the IPA Handbook, one might consider that dialect atlases employ hundreds of symbols (LAMSAS distinguishes over 1,100 different vowels). A complete segment distance table thus requires tens of thousands of specifications, minimally (and in the case of LAMSAS over 500,000).

In his thesis Heeringa (2004) calculated segment distances using two different approaches. In Chapter 3 he represented every phone as a bundle of features where every feature is a certain phonetic property. In Chapter 4 Heeringa measured the transcription distances using acoustic segment distances calculated from the recordings in Wells and House (1995). This is less arbitrary than the feature representation since it is based on physical measures, but both of these approaches have their disadvantages. The former relies on the selection of a feature system, while the latter requires acoustic data to be available.

In order to avoid these problems, Wieling et al. (2009) proposed using pointwise mutual information (PMI) to automatically induce segment distances from phonetic transcriptions. PMI is a measure of association between two events *x* and *y:*

$$PMI(x,y) = log_2(P(x,y) / P(x)P(y))$$

where the numerator $P(x,y)$ tells us how often we have observed the two events together, while the denominator $P(x)P(y)$ tells us how often we would expect these two events to occur together if we assumed that their occurrence were statistically independent. The ratio between these two values shows us if two events co-occur together more often than just by chance. Wieling et al. (2009) use it to automatically learn the distances between the phones in aligned word transcriptions and also to improve the automatically generated alignments. Applied to aligned transcriptions, $P(x,y)$ represents the relative frequency of two segments being aligned together, while $P(x)$ and $P(y)$ are relative frequencies of segments *x* and *y*.

The procedure of calculating the segment distances and improving the alignments is iterative and consists of the following steps: a) align all word transcription using the Levenshtein algorithm b) from the obtained alignments calculate the PMI distances between the phones c) align all word transcriptions once more using the Levenshtein algorithm, but based on the generated segment distances d) repeat the previous two steps until there are no changes in segment distances and alignments.

Wieling et al. (2009) evaluated the alignments based on the PMI procedure on a manually corrected gold standard set of alignments of Bulgarian dialect data. The results indicated that both at the segment level and at the word level the novel algorithm was a clear improvement over the Levenshtein algorithm with hand-coded segment distances.

Qualitative error analysis has shown that both versions of the Levenshtein algorithm make most errors due to the restriction that vowels and consonants cannot be aligned. Apart from this error, the simple Levenshtein algorithm is not able to distinguish between aligning a vowel with one of the two neighboring vowels, since in the simple version of the algorithm all vowels are equally distant from each other. This also holds for the consonants. Using PMI-induced segment distances solves this problem since the algorithm learns that the distance between [n] and [ŋ] is smaller than the distance between [n] and [k] (see Figure 1). Correction of these types of errors is where the PMI procedure improves the performance of the simple Levenshtein algorithm and generates more correct alignments.

v	ˈɤ	-	n	-		v	ˈɤ	n	-	-
v	ˈɤ	ŋ	k	ə		v	ˈɤ	ŋ	k	ə

Figure 1

Erroneous alignment produced by the simple Levenshtein algorithm (left) and the correct alignment produced by Levenshtein PMI algorithm (right)

The alignments produced using the segment distances arrived at via the pointwise mutual information procedure improves the alignment accuracy of the Levenshtein algorithm, and consequently enables us to obtain better distances between each pair of sites calculated from the transcriptions. The next step would be to improve this procedure and enable the algorithm to calculate the distances between vowels and consonants. In that way, the quality of the alignments could be further improved by minimizing the number of errors caused by the restriction on the alignments between vowels and consonants.

3. Multi-Alignment

While the technique described in the previous section aims at improving pairwise alignments, Prokić et al. (2009) introduced an algorithm that is used to produce multiple sequence alignments. It is an adapted version of the ALPHAMALIG algorithm (Alonso et al. 2004) modified to work with phonetic transcriptions.

Pairwise string alignment methods compare two strings at a time, while in multiple string alignment (MSA) all strings are aligned and compared at the same time. MSA is an especially effective technique for discovering patterns that can be hard to detect when comparing only two strings at once. Both techniques are widely used in bioinformatics for aligning DNA, RNA or protein sequences in order to determine similarities between the sequences. However, as noted in Gusfield (1997), the multiple string alignment method is more powerful. Gusfield calls it "the holy grail" of sequence algorithms in molecular biology.

In recent years there has been increasing interest in using phylogenetic methods to analyze linguistic data, especially language change and variation (Gray and Atkinson 2003; Atkinson et al. 2005; Warnow et al. 2006). This is possible because of the similarities between the evolution of languages and the evolution of species —they

are both passed on from generation to generation accompanied by changes during the process. As they change, both languages and biological populations can split into new subgroups, becoming more and more distant from each other and from common ancestor(s). In order to apply methods from biology directly to pronunciation data, which we are particularly interested in, it is essential to preprocess the data in order to identify all the correspondences. This means that we need to be able to derive multiply aligned strings of phonetic transcriptions. An example of multiply aligned transcriptions can be seen in Figure 2:

village 1:	j	'ɑ	-	-	-	-
village 2:	j	'ɑ	z	e	-	-
village 3:	-	'ɑ	s	-	-	-
village 4:	j	'ɑ	s	-	-	-
village 5:	j	'ɑ	z	e	k	a
village 6:	j	'ɛ	-	-	-	-
village 7:	-	'ɒ	s	-	-	-

Figure 2

Multiply aligned phonetic transcriptions
of the Bulgarian word *az* 'I' collected at 7 villages

The advantage of multiply aligned strings over pairwise alignments is two-fold: a) it is easier to detect and process sound correspondences (e.g. [ɑ], [ɛ] and [ɒ] are very easy to detect and extract from the second column in Figure 2); b) the distances between strings are more precise if calculated from multiple aligned strings since they preserve information on the sounds that were lost (the last two columns in all transcriptions —except for village 5's transcription— preserve the information that the villages commonly lack the last two sounds, which would be lost in the pairwise alignments).

The automatic alignment of strings was carried out using the ALPHAMALIG algorithm, originally designed for bilingual text alignment. It is an iterative pairwise alignment program that merges multiple alignments of subsets of strings. Although originally developed to align words in texts, it can work with any data that can be represented as a sequence of symbols of a finite alphabet. In Prokić et al. (2009) the algorithm was adapted in order to work with phonetic transcriptions. The distances between the phones were set in such a way that vowels can be aligned only with vowels and consonants only with consonants.

Since there is no widely accepted way to evaluate the quality of multiple alignments Prokić et al. (2009) suggested two new methods for comparing automatically produced alignments against manually aligned strings, the so called *gold standard*. Both methods compare the content of corresponding columns in two alignments. One method, called the *column dependent method,* takes into account the order of columns in two alignments and the content of the columns as well. In other words, it looks for a perfect match. The other method is not sensitive to the order of col-

umns and takes into account only the content of two corresponding columns. It is based on Modified Rand Index (Hubert and Arabie 1985), one of the most popular methods for comparing two different partitions.

The results for the Bulgarian data set show that automatically generated alignments are of a very high quality, scoring between 92% (first method) and 98% (latter method). Error analysis has revealed that most of the alignment errors are due to the restriction that vowels and consonants cannot be aligned. In order to avoid this problem, the algorithm would need information on the distances between the phones, which is not straightforward to obtain (see Section 2). Although both evaluation methods could be improved further, they both estimate alignment quality well.

Studies in historical linguistics and dialectometry where string comparison is used could benefit from tools for multiple sequence alignment by speeding up the process of string aligning and making it suitable to work with large amounts of data.

4. Spectral Graph Clustering

Until recently, almost all aggregate dialectometric analyses have focused on identifying the most important geographical groups in the data. While it is important to identify the areas which are linguistically similar and those which differ, the aggregate approach does not expose the linguistic basis of the groups.

The aggregate approach averages over the distances between pairs of large numbers of aligned words to obtain pairwise dialect distances. After obtaining an MDS map of Dutch dialects, Wieling et al. (2007) correlated the distances based on each individual word in the dataset with all MDS dimensions to find the most characteristic words for each of the three MDS dimensions. While finding the most characteristic word is certainly informative, linguists are also interested in finding the most important sound correspondences. Nerbonne (to appear, a) used factor analysis to identify the linguistic structure underlying the aggregate analysis of southern American dialects, focusing his analysis on vowels and showing that aggregate distances based only on vowels correlated highly with distances based on all sound segments.

Prokić (2007) went a step further and extracted the ten most frequent non-identical sound correspondences from the aligned transcriptions of pairs of Bulgarian dialects and used the relative frequency of each of these sound correspondences to assign a score to each site (each site had multiple scores; one for each sound correspondence). When the pairwise distances were calculated on the basis of these scores and correlated with the aggregate distances, she was able to identify how characteristic each sound correspondence was for the aggregate view.

All the aforementioned methods have in common that the process of determining the linguistic basis is *post-hoc*; the aggregate distances are calculated first and the linguistic basis is determined later. This is less than optimal as we wish to know which features or sound correspondences really serve as the linguistic basis for the site grouping. Consequently, linguists have also not been convinced by these approaches and have been slow to embrace the aggregate approach.

Another approach was taken by Shackleton (2007), who used principal component analysis to group features (instead of sound correspondences) in the pronunciation of English dialects. For each variety the component scores were calculated and

groups of varieties were distinguished based on the presence of the grouped features. Hence, in this case, first the groups of features are determined, after which the geographical groups are identified. Shackleton did not use sound correspondences, but he used self-selected features of both consonants and vowels and also (in a separate experiment) variants determined by English linguists. While this is certainly insightful, there remains a great deal of subjectivity in categorizing and selecting the features.

To counter these drawbacks, Wieling and Nerbonne (2009, to appear) introduced a new method to simultaneously cluster geographic varieties as well as the concomitant sound correspondences (compared to a reference variety). The BIPARTITE SPECTRAL GRAPH PARTITIONING method they applied was first introduced by Dhillon (2001) to co-cluster words and documents and is based on calculating the singular value decomposition (SVD) of a word-by-document matrix. The left and right singular vectors obtained from this procedure are merged and clustered into the desired number of groups using the *k*-means algorithm. A detailed explanation as well as an example is shown in Wieling and Nerbonne (to appear).

The variety-by-sound correspondence matrix of Wieling and Nerbonne (2009, to appear) was based on alignments for 423 Dutch varieties with respect to a reference pronunciation close to standard Dutch using the PMI algorithm discussed in Section 2. All sound correspondences present in the alignments for a variety were counted and in the matrix the presence (frequency of at least 1) or absence of a sound correspondence in a variety was stored. We did not use the frequencies in the matrix as this seemed to have a negative impact on performance, possibly because of the presence of some very high frequencies. In their first study, Wieling and Nerbonne (2009) reported a fair geographical clustering in addition to sensible sound correspondences, based on "eyeballing" the data. In a subsequent study, Wieling and Nerbonne (to appear) developed a method to rank the sound correspondences to identify the most important ones for each cluster based on representativeness (i.e. the proportion of varieties in a cluster containing the sound correspondences) and distinctiveness in a cluster (i.e. the number of varieties within as opposed to outside the cluster containing the sound correspondence). They concluded that their method to rank the sound correspondences conformed to a great extent with the subjectively selected sound correspondences in the previous study (Wieling and Nerbonne 2009). While this method still has some drawbacks (e.g., incorporating frequency information has a negative effect on the results), it is certainly a step forward in identifying the linguistic basis of aggregate dialectometric analyses.

5. Intelligibility of contrasting varieties

Dialectometrical techniques are useful tools for investigating the role that linguistic distances play in the mutual intelligibility among speakers of closely related language varieties. The techniques allow the researcher to test the relationship between intelligibility on the one hand and objective linguistic similarity on the other. It seems likely that the greater the linguistic resemblance is between two languages or dialects, the greater the degree of mutual intelligibility will be. However, only a moderately strong correlation ($r = -0.65$, $p < 0.01$) was found between intelligibil-

ity scores of 17 Scandinavian dialects by Danish listeners and the perceived distances to these dialects from the listeners' own varieties (Beijering, Gooskens and Heeringa 2008). This suggests that perceived distance and intelligibility scores are two different measurements that cannot be equated with each other. In other words, the (dis)similarity of another language variety to one's own, is only a moderately successful predictor of the how intelligible this variety is.

Methods for testing and measuring the effect of linguistic distance are becoming increasingly sophisticated. Methods include web-based experiments and computational techniques. Intelligibility can be measured by asking listeners to answer open or closed questions about the content of a spoken text or by having subjects translate a spoken text or word lists. By means of open questions about a text, the mutual intelligibility of three Scandinavian standard languages (Danish, Norwegian and Swedish) and three West-Germanic languages (Afrikaans, Dutch and Frisian) were tested in Gooskens (2007). The percentage of correctly answered questions per listener-language combination (e.g. Danes listening to Swedish) was correlated with the corresponding phonetic distances measured with the Levenshtein algorithm. There was a negative correlation of –0.64 ($p < 0.01$) between linguistic distance and intelligibility when all data were included but a stronger correlation ($r = –0.80$, $p < 0.01$) when only the Scandinavian data were included.

In another study (Beijering, Gooskens and Heeringa 2008), the intelligibility of 17 Scandinavian dialects by speakers of Standard Danish was tested using a translation task. The percentage of correctly translated words in a short story (per language variety) was correlated with phonetic distances from Standard Danish to each of the 17 dialects. Previous applications of the Levenshtein algorithm typically employed a word length normalization, which means that the total number of operations (insertions, deletions and substitutions) is divided by the number of alignment slots for a word pair. The effect of normalization is that a pronunciation difference counts more in a short word than in a long word. In our investigation, we correlated the intelligibility scores with normalized as well as non-normalized Levenshtein distances. The results showed higher correlations than in the previous study ($r = –0.86$, $p < 0.01$ for the normalized and $r = –0.79$, $p < 0.01$ for the non-normalized distances). The phonetic distances between each of the 17 dialects and Standard Danish were correlated with perceptual distances as judged by the listeners on a 10-point scale. The normalized Levenshtein distances correlated more strongly with the intelligibility scores than with perceived distances, and the difference was significant ($r = –0.86$ versus $r = 0.52$, $p < 0.05$). The non-normalized Levenshtein distances showed the same tendency, but the difference between normalized and non-normalized distances was not significant ($r = –0.79$, $p < 0.01$ versus $r = –0.62$, $p < 0.01$ respectively). These results suggest that Levenshtein distance is a good predictor of both intelligibility and perceived linguistic distances. However, the algorithm seems to be a better predictor of intelligibility than of perceived linguistic distances. Word length normalization is important when modeling intelligibility since segmental differences in short words presumably have a larger impact on intelligibility than segmental differences in long words. On the other hand, perceived distance is likely to be dependent on the total number of deviant sounds regardless of word length and therefore correlations are higher with the non-normalized distances.

Mutual intelligibility among the Scandinavian languages is fairly high, comparable to the mutual intelligibility in many dialect situations. Several investigations have been carried out in order test how well Scandinavians understand each other (e.g. Maurud 1976; Börestam 1987; Delsing & Lundin Åkesson 2005; Gooskens, Van Heuven, Van Bezooijen and Pacilly accepted). Results repeatedly show asymmetric intelligibility scores between Scandinavian language pairs. Especially Swedes have more difficulties understanding Danes than vice versa. The techniques for distance measurements used for the investigations discussed above cannot capture this asymmetry. A conditional entropy measure has therefore been developed as a measure of remoteness to model asymmetric intelligibility (Moberg, Gooskens, Nerbonne and Vaillette 2007). In the conditional entropy measure semantically corresponding cognate words are taken from frequency lists and aligned. The conditional entropy of the phoneme mapping in aligned word pairs is calculated. This approach aims to measure the difficulty of predicting a phoneme in a native language given a corresponding phoneme in the foreign language. The results show that a difference in entropy can be found between language pairs in the direction that previous intelligibility tests predict. Conditional entropy as a measure of remoteness thus seems a promising candidate for modeling asymmetric intelligibility.

In the investigations mentioned above, intelligibility was measured on the basis of whole texts and aggregate phonetic distance measures were applied. As a consequence, no conclusions could be drawn about the nature of the phonetic differences that contribute most to intelligibility. However, it is desirable to gain more detailed knowledge about which role various linguistic factors play in the intelligibility of closely related languages. Gooskens, Beijering and Heeringa (2008) reanalyzed the data from the intelligibility tests with 17 Scandinavian dialects (see above), now measuring the consonant and the vowel distances separately. A higher correlation was found between intelligibility and consonant distances ($r = -0.74$, $p < 0.01$) than between intelligibility and vowel distances ($r = -0.29$, $p < 0.05$), which confirms the claim that consonants play a relatively important role for intelligibility of a closely related language.

Kürschner, Gooskens and Van Bezooijen (2008) focused on the comprehension of 384 isolated spoken Swedish words among Danes, and examined a wide variety of potential linguistic predictors of intelligibility, including the similarity of the foreign word's pronunciation to one's own variety's pronunciation. The strongest correlation was found between word intelligibility and phonetic distances ($r = -0.27$, $p < 0.01$). This is rather low in comparison with the correlations found for the intelligibility of whole texts with aggregate phonetic distances. Including more linguistic factors like word length, foreign sounds, neighborhood density and word frequency in a logistic regression analysis improved the predictive power of the model. However, a large amount of variation still remains to be explained. We attribute this to the high number of idiosyncrasies of single words compared with the aggregate intelligibility and linguistic distance used in earlier studies. While at the aggregate level we are well able to predict mutual intelligibility between closely related language varieties, it is a challenge for future work to develop phonetic distances that are better able to express the communicative distances between closely related languages at the word level.

6. Séguy's law

In the paper which launched the dialectometric enterprise, Jean Séguy (1971) observed that the measure of aggregate lexical distance which he defined in the same paper increased directly with geographic distance, but in a sub-linear fashion. Not long after Séguy's paper, Trudgill (1974) advocated that "gravity" laws of the sort then popular in the social sciences might underlie the dynamics of linguistic diffusion. He explicitly advocated that dialectology seek more general accounts of linguistic variation. But where Séguy had measured a sublinear effect of geography on linguistic diversity, Trudgill postulated an attractive force which weakened with the square of linguistic distance. Trudgill clearly investigated geography as a means of studying the effect of social contact, a point at which we, and we suspect nearly all dialectologists, heartily agree. This can be seen in the fact that Trudgill also predicted that population size would play a role in strengthening the tendency of two sites to adopt each other's speech habits. Nerbonne & Heeringa (2007) explicitly contrasted the two views of geography, arguing their incompatibility based on a study of 52 locations in the northeast Netherlands, and showing that aggregate Dutch dialect distance followed the a logarithmic curve not unlike the sublinear curve that Séguy had used to model his Gascony data.

Nerbonne (to appear, c) examines five more dialect areas, namely Forest Bantu (Gabon), Bulgaria, Germany, Norway, and the U.S. Eastern seaboard, and shows that the distribution of aggregate linguistic distance in each of these is a logarithmic function of geography, just as the Netherlands and Gascony. Figure 3 shows the six curves.

Six areas of linguistic variation that display the same sub-linear relation between geographic distance and aggregate linguistic distance, first noted by Séguy (1971). Logarithmic curves are drawn. Because different measuring schemes were applied, the *y*-axes are not comparable.

The paper goes on to examine the relation between the dynamic influencing individual linguistic variables and the curve representing aggregate variation, showing that the sub-linear aggregate curve is indeed incompatible with a dynamic obeying an inverse square law, but that it is also compatible with a dynamic in which the force to differentiate decreases linearly with distance.

The general, cross-linguistic relation between geography and linguistic variation is clearly of great potential interest to dialectology and the study of linguistic diffusion and deserves further attention. Progress can be made by obtaining a broader selection of linguistic case studies on which to base the general views, by the development of a metric that is applicable cross-linguistically without confounding effects, and by the development of hypotheses concerning the more exact form of the curves.

7. The proximate horizon

Dialectology has always based its appeal on its attention to the incredible detail and multi-faceted nature of linguistic variation. Dialectometry serves dialectology by improving the efficiency of its data archiving techniques, and by developing efficient and replicable data analysis techniques, which in turn broaden the empirical base on which theoretical dialectology can build. Finally, the opportunity to measure linguis-

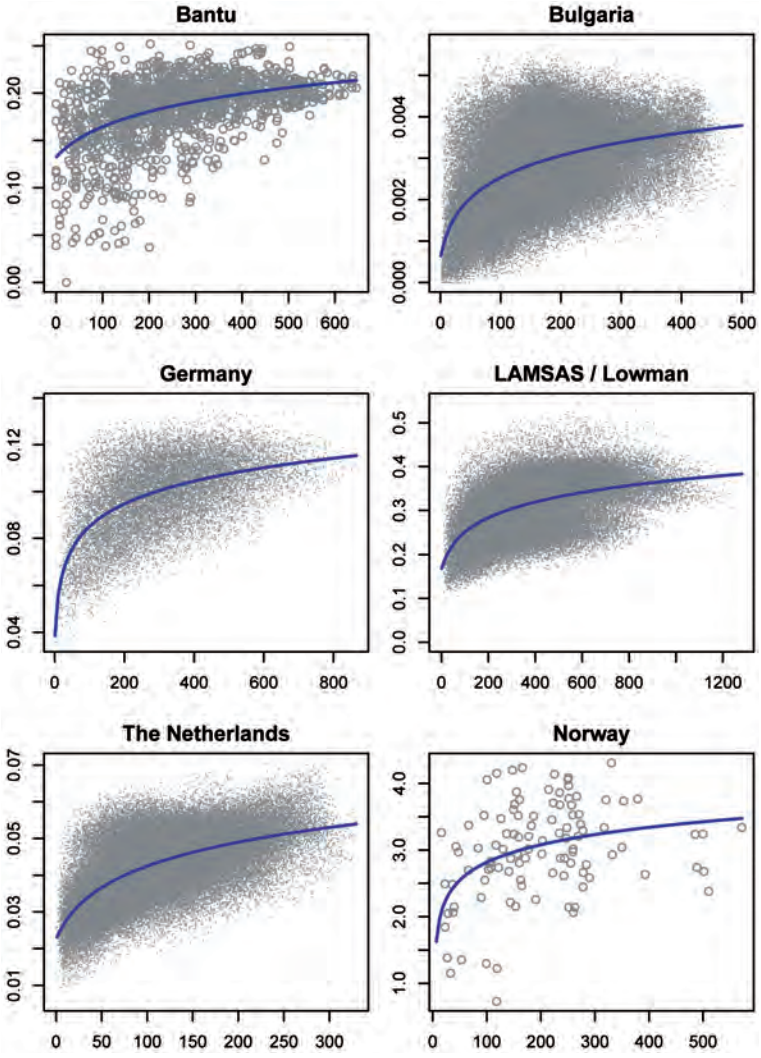

Figure 3

Séguy's law

tic variation validly and consistently may open the door to more abstract and general characterizations of dialectology's central ideas.

Several topics deserve spots high on the dialectometrical agenda. We mention first two practical matters. First, although the Salzburg VDM package and the Groningen L04 package are being used profitably at other centers, it would be worthwhile to develop new packages or new versions of these packages which are easier for linguists with standard dialectological training to use. Second, it would be worthwhile devel-

oping techniques to extract variation from the increasingly available corpora which are collected for the purposed of pure and applied linguistic research (Szmrecsanyi 2008).

Turning to the more abstract scholarly topics, we should mention third, the relation between synchronic variation and historical linguistics. There is widespread consensus that language variation and language change belong together as topics, and there is increasing interest in the use of exact techniques in historical linguistics (Nakleh, Ringe & Warnow 2005). It would be only natural to see some cross fertilization in these two areas, for example in the deployment of the alignment techniques discussed here and the application of phylogenetic inference, which to-date has mostly used lexical data and/or manually prepared phonetic or phonological data.

A fourth opportunity for progress in dialectometry lies in the further validation of techniques. By validation, we mean the proof that the techniques are indeed measuring what they purport to measure, an undertaking which presupposes that one has been explicit about what this is. A good number of studies are undertaken without being explicit on this point: the researchers seem content to show that 81% of the vocabulary is shared between sites s_1 and s_2, etc. But the ease with which alternatives are developed and implemented makes further reflection on this point absolutely imperative. Even in the case of share vocabulary, we ask whether it is enough that some sort of cognate is found, whether the order of preferences for certain lexicalizations above others has been taken into account, whether the significance of shared vocabulary might not need to be corrected for frequency effects, etc. These considerations led Gooskens and Heeringa (2004) to suggest that dialectometric measurements be understood as measuring the signals of provenance which dialect speakers provide, which in turn led them to view to validate the measurements on the basis of dialect speakers' abilities to identify a dialect as like or unlike their own. Further studies along these lines would be most welcome, if for no other reason, than to guard against depending too much on a single behavioral study done on speakers of a single, Scandinavian language.

A fifth, but related opportunity certainly lies in the further investigation of the relation between comprehensibility (as discussed above in Section 5) and the signal of provenance that was central in Gooskens and Heeringa (2004). It is clear that comprehensibility and the signal of "likeness" correlate to a high degree, giving rise to the question of whether the two are ultimately the same, or, alternatively where they part their ways and with respect to which (linguistic) phenomena. More empirical studies investigating the overlap would certainly advance the field.

Sixth, we do not mean to suggest that the spectral clustering techniques presented in Section 4 above should be regarded as closing the book on the issue of how to identify the linguistic basis of dialect distributions. We are optimistic that these techniques are promising, but they should be compared *inter alia* to techniques such as Shackleton's (2007) use of principal component analysis. The quality of the groups detected needs further investigation, and the impact of factors such as frequency would be worthwhile examining closely.

Seventh, and finally, we hope that further research into the general relation between geographical distance and dialectal difference is a promising avenue for further work, as we hope to have suggested in Section 6 above.

References

Alonso L., Castellon, I., Escribano, J., Messegeur, X. & L. Padro, 2004, «Multiple sequence alignement for characterizing the linear structure of revision», in *Proceedings of the 4th International Conference on Language Resources and Evaluation*.

Atkinson, Q., Nicholls, G., Welch, D. & R. Gray, 2005, «From words to dates: water into wine, mathemagic or phylogenetic», *TPS* 103, 193-219.

Beijering, K., Gooskens, C. & W. Heeringa, 2008, «Predicting intelligibility and perceived linguistic distances by means of the Levenshtein algorithm», in M. van Koppen & B. Botma (eds.), *Linguistics in the Netherlands*, John Benjamins, Amsterdam, 13-24.

Black, P., 1976, «Multidimensional Scaling applied to Linguistic Relationships», *Cahiers de l'Institut de linguistique de Louvain* 3, 43-92.

Börestam Uhlmann, U., 1991, *Språkmöten och mötesspråk i Norden* [Language meetings and the language of meetings], Nordisk språksekretariat, Oslo.

Delsing, L-O. & K. Lundin Åkesson, 2005, *Håller språket ihop Norden? En forskningsrapport om ungdomars förståelse av danska, svenska och norska* [Does the Language Keep the Nordic Countries Together? A Research Report on How Well Young People Understand Danish, Swedish and Norwegian], Nordiska ministerrådet, Copenhagen.

Dhillon, I., 2001, «Co-clustering documents and words using bipartite spectral graph portioning», in *Proceedings of the seventh ACM SIGKDD international conference on knowledge discovery and data mining*, ACM, New York, 269-274.

Goebl, H., 1982, *Dialektometrie: Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie im Bereich der Dialektgeographie*. Osterreichische Akademie der Wissenschaften, Wien.

Gooskens, C., 2007, «The contribution of linguistic factors to the intelligibility of closely related languages», *Journal of Multilingual and Multicultural Development* 28 (6), 445-467.

—, Beijering, K. & W. Heeringa, 2008, «Phonetic and lexical predictors of intelligibility», *International Journal of Humanities and Arts Computing* 2 (1-2), 63-81.

— & W. Heeringa, 2004, «Perceptive evaluation of Levenshtein dialect distance measurements using Norwegian dialect data», *Language Variation and Change* 16(3), 189-207.

—, Heuven, V. Van, Bezooijen, R. van & J, Pacilly, accepted, «Is spoken Danish less intelligible than Swedish?», *Speech Communication.*

Gray, R. & Q. Atkinson, 2003, «Language-tree divergence times support Anatolian theory of Indo-European origin», *Nature* 405, 1052-1055.

Gusfield, D., 1997, *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. CUP.

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

Hubert, A. & P. Arabie, 1985, «Comparing partitions», *Journal of Classification* 2, 193-218.

Kleinberg, J., 2003, «An impossibility theorem for clustering», in S. Becker, S. Thrun & K. Obermaier (eds.), *Advances in Neural Information Processing Systems* 15. Avail. at http://books.nips.cc/papers/files/nips15/LT17.pdf.

Kürschner, S., Gooskens, C. & R. van Bezooijen, 2008, «Linguistic determinants of the intelligibility of Swedish words among Danes», *International Journal of Humanities and Arts Computing* 2 (1-2), 83-100.

Maurud, Ø., 1976, *Nabospråksforståelse i Skandinavia. En undersøkelse om gjensidig forståelse av tale- og skriftspråk i Danmark, Norge og Sverige* [Mutual intelligibility of languages in Scandinavia. A study of the mutual understanding of written and spoken language in Denmark, Norway and Sweden], Nordiska Rådet, Stockholm.

Moberg, J., Gooskens, C., Nerbonne, J. & N. Vaillette, 2007, «Conditional Entropy Measures Intelligibility among Related Languages» in P. Dirix, I. Schuurman, V. Vandeghin-

ste & F. van Eynde (eds.), *Computational Linguistics in the Netherlands 2006: Selected papers from the 17th CLIN Meeting*, LOT, Utrecht, 51-66.

Nakleh, L., Ringe, D. & T. Warnow, 2005, «Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages», *Language* 81(2), 382-420.

Nerbonne, J., 2009, «Data-Driven Dialectology», *Language and Linguistic Compass*. 3(1), 2009, 175-198. DOI: 10.1111/j.1749-818x2008.00114.x

— (to appear, a), «Various Variation Aggregates in the LAMSAS South», in C. Davis & M. Picone (eds.) *Language Variety in the South III,* University of Alabama Press, Tuscaloosa.

— (to appear, b), «Mapping Aggregate Variation», in S. Rabanus, R. Kehrein & A. Lameli (eds.), *Mapping Language,* De Gruyter, Berlin.

— (to appear, c), «Measuring the Diffusion of Linguistic Change», *Philosophical Transactions of the Royal Society B: Biological Sciences*.

— & W. Heeringa, 1997, «Measuring Dialect Distance Phonetically», in J. Coleman (ed.), *Workshop on Computational Phonology,* ACL, Madrid, 11-18.

— & —, 2007, «Geographic Distributions of Linguistic Variation Reflect Dynamics of Differentiation», in S. Featherston & W. Sternefeld (eds.), *Roots: Linguistics in Search of its Evidential Base*, De Gruyter, Berlin, 267-297

— & —, 2009, «Measuring Dialect Differences», in J.-E. Schmidt & P. Auer (eds.), *Language and Space: Theories and Methods* in series *Handbooks of Linguistics and Communication Science*, Chap. 31, De Gruyter, Berlin, 550-567.

—, — & P. Kleiweg, 1999, «Edit Distance and Dialect Proximity», in D. Sankoff & J. Kruskal (eds.), *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison,* 2nd ed., CSLI Press, Stanford, v-xv.

—, Kleiweg, P., Heeringa, W. & F. Manni, 2008, «Projecting Dialect Differences to Geography: Bootstrap Clustering vs. Noisy Clustering», in Ch. Preisach, L. Schmidt-Thieme, H. Burkhardt & R. Decker (eds.), *Data Analysis, Machine Learning, and Applications. Proc. of the 31st Annual Meeting of the German Classification Society,* Springer, Berlin, 647-654.

Prokić, J., 2007, «Identifying linguistic structure in a quantitative analysis of dialect pronunciation», *Proceedings of the ACL 2007 Student Research Workshop*, ACL, Prague, 61-66.

— & J. Nerbonne, 2008, «Recognizing Groups among Dialects», *International Journal of Humanities and Arts Computing*, 153-172. DOI: 10.13366/E1753854809000366.

—, Wieling, M. & J. Nerbonne, 2009, «Multiple string alignments in linguistics», in L. Borin & P. Landvai (chairs), *Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH-SHELT&R 2009),* EACL Workshop.

Séguy, J., 1971, «La relation entre la distance spatiale et la distance lexicale», *Revue de Linguistique Romane* 35, 335-357.

Shackleton, R. G., 2007, «Phonetic Variation in the Traditional English Dialects», *Journal of English Linguistics* 35(1), 30-102. DOI: 10.1177/0075424205297857.

Szmrecsanyi, B., 2008, «Corpus-Based Dialectometry: Aggregate Morphosyntactic Variability in British English Dialects», *International Journal of Humanities and Arts Computing* 2, special issue on *Computing and Language Variation*, ed. by J. Nerbonne, C. Gooskens, S. Kürschner & R. van Bezooijen, 261-278.

Trudgill, P., 1974, «Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography», *Language in Society* 2, 215-246.

Warnow, T., Evans, S., Ringe, D. & L. Nakhleh, 2006, «A stochastic model of language evolution that incorporates homoplasy and borrowing», in P. Foster and C. Renfrew (eds.), *Phylogenetic Methods and the Prehistory of Languages*, MacDonald Institute for Archeological Research, Cambridge.

Wells, J. & J. House, 1995, *The Sounds of the International Phonetic Alphabet*. Dept. Phonetics & Linguistics, University College London. Booklet with tape.

Wieling, M., Heeringa, W. J. and J. Nerbonne, 2007, «An aggregate analysis of pronunciation in the Goeman-Taeldeman-van Reenen-Project data», *Taal en Tongval* 59, 84-116.

— & J. Nerbonne, 2009, «Bipartite spectral graph partitioning to co-cluster varieties and sound correspondences in dialectology», in M. Choudhury et al. (eds.) *Proceedings of the 2009 Workshop on Graph-based Methods for Natural Language Processing (TextGraphs-4)*, ACL-IJCNLP, Singapore, 7 August 2009, 14-22.

— & — (to appear), «Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features», *Computer Speech and Language*.

—, Prokić, J. and J. Nerbonne, 2009, «Evaluating the pairwise string alignment of pronunciations», in L. Borin & P. Landvai (chairs), *Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH-SHELT&R 2009),* EACL Workshop.