BRILL

Language Dynamics and Change 3 (2013) 147–168

brill.com/ldc

Combining Regular Sound Correspondences and Geographic Spread

Jelena Prokić[a] and Michael Cysouw[b]
a) Forschungszentrum Deutscher Sprachatlas,
Philipps-Universität Marburg, Germany
prokic@uni-marburg.de
b) Forschungszentrum Deutscher Sprachatlas,
Philipps-Universität Marburg, Germany
cysouw@uni-marburg.de

Abstract

In this paper we combine the geographic variation of closely related language variants ('dialects') with the distribution of sound correspondences through the lexicon. One of the central problems with sound correspondences at the dialect level is that they are not very regular, especially when they are investigated in sufficient detail. Sound changes spread both through a language (e.g., from one word to another) and through the population of speakers (in our case through a population of villages with different dialects). Both processes happen at the same time, and the challenge is to reconstruct what has happened from a snapshot of synchronic data. The method described in this paper allows us to track the geographic spread of sound changes and the underlying patterns of linguistic diversity simultaneously. By combining the two, it is possible to detect areas of intensive linguistic contact and gain better insight into the mechanisms of language change.

Keywords

Sound change; diffusion model; geography

1. Introduction

In recent years, the growth of digitally available language data has increased the number of quantitatively oriented studies on language change and evolution. These studies include problems of language evolution both on the macro-level and micro-level. The macro-level studies investigate the relationship between and within major language families. Micro-level studies focus on language changes at the dialect level and are often referred to as *dialectometry* or *quantitative dialectology*. Quantitative methods were first introduced to dialectology in 1971 with the work of French linguist Jean Séguy, who developed the first technique for measuring the distances between dialects by counting the overlapping features between any two sites (Séguy, 1971). Further improvement

DOI: 10.1163/22105832-13030205

in the methodology of dialectometry came with the work of Hans Goebl (Goebl, 1982, 1984), who, among many important innovations, introduced a weighting of linguistic features, and was also the first one to use clustering techniques, Thiessen tiling and chloropleth maps in dialectometry. This line of research is known as the *Salzburg school of dialectometry* (Goebl, 2006). Another important line of research, known as the *Groningen school of dialectometry*, is best known for the application of the Levenshtein distance to automatically measure the distances between dialect varieties (Nerbonne et al., 1996, 1999; Heeringa, 2004). Further developments in the *Groningen school of dialectometry* include, among many, introduction of multidimensional scaling maps (Nerbonne and Heeringa, 1998), noisy or composite clustering (Nerbonne et al., 2008) and several methods used to detect distinctive linguistic features for a given dialect area (Wieling and Nerbonne, 2011; Prokić et al., 2012).

Most of the previous studies in quantitative dialectology have focused on the automatic detection of dialect groups within one language or geographic area, and the relationships among the identified dialect groups. Comparison of the dialects includes research on the phonetic, lexical, morphological and syntactic levels, with the phonetic level being the most frequently investigated. However, the underlying linguistic processes, responsible for the observed dialect distributions, have been addressed to a much lesser extent in dialectometric studies. In this research we adopt an approach that goes beyond the detection of dialect groups ('beyond phylogenies') and instead focuses on quantification of the underlying models of dialect change. We combine the spread of sound change through a population of speakers, i.e., geographic variation of sound change, with the internal variation of sound correspondences in the lexicon. Our starting point is a list of multi-aligned phonetic transcriptions of 152 words collected at 197 villages in Bulgaria. The multi-aligned data allows us to quickly extract all sound correspondences between any two villages in the data set. Although the pronunciation of all words is transcribed in IPA, the phonetic information is ignored for the establishment of similarities between sounds of different dialects. Instead, we calculate the similarity between sounds based solely on how frequently they co-occur in the alignments. The main idea behind this approach is that the regularity of sound change provides a good estimate of the phylogenetic similarity between the sounds: the more often two sounds change together, the stronger the association strength between them. This is a data-driven approach that enables us to automatically identify more or less regular sound changes both in the lexicon and in geographic space.

Regarding the geographic spread of sound change, we investigate the so-called 'diffusion model,' characteristic for regions with a long settlement history, such as the region from which our data originates (Chambers and Trudgill, 1998:

Ch. 11). The diffusion model assumes that linguistic innovations spread via everyday contact in a wave-like manner (Section 4.2). The spread is characterized by the existence of a *focal area* where the sound change originates and is regular. The area in which the change does not happen is the so-called *relic area*. In between is the *transition area* where the change is less regular. For a detailed description on the focal, transition and relic areas see Hock (1991). The patterns of linguistic spread in regions with a short settlement history—one or two centuries long—are known to look quite different (Chambers and Trudgill, 1998), but we do not address that aspect in our model in this research.

2. Similarity Based on Co-occurrence Statistics

In statistics, an association is any relationship between two variables such that some of the variability of one can be accounted for by the other. Association measures can be applied to various types of data to infer the presence or absence of an association. There are dozens of association measures present in the literature, and choosing the best one depends on the data. In quantitative linguistics, different association measures have, for example, been successfully applied to collocation extraction, with the goal of identifying those combinations of words that display idiosyncrasies in their distribution within a corpus (Evert, 2005). The strengths of the association between two words is based on their co-occurrences in a corpus; the more frequently two words co-occur, the stronger the association between them.

In the past decade, several association measures have also been applied to estimate the association strength between sounds in word lists. They are all based on the co-occurrences of two sounds in the aligned words and allow the researcher to infer from the data how similar or dissimilar these two sounds are. A good measure of the similarity or dissimilarity between the sounds leads to a better word alignment and subsequently to a more accurate estimation of the similarities between two words. Below we give a short overview of the association measures used to estimate the association strength between the sounds.

Tiedemann (1999) uses the Dice coefficient, combined with co-occurrence statistics, to automatically construct string similarity matrices. Mackay and Kondrak (2005) use a Pair Hidden Markov Model, a version of the Hidden Markov Model, to compute the similarity between pairs of words. From the word pairs that are known to be similar, the parameters of the model are automatically learned, including the substitution weights between the orthographic symbols, which are calculated using the log-odds ratio. Cysouw and Jung (2007) also use the Dice coefficient. Prokić (2010: Ch. 5, pages 71–87) relies on pointwise mutual information to extract the association between the phones from the phonetically transcribed word lists (note that mathematically the log-odds ratio and the pointwise mutual information are identical). The procedure is iterative: corresponding words are aligned, the association between the phones is calculated and all words are realigned using the learned weights. This process is repeated until there are no more changes in the alignments and in the learned weights for the phones. This procedure results in improved alignments. Wieling and colleagues employ a different version of the pointwise mutual information in order to induce phonetic similarity from pronunciation variation by focusing only on the phones that change (Wieling et al., 2012). An iterative procedure is also used by Steiner et al. (2011) and List (2012) in the task of automatic cognate identification. In these two papers, the association between the phones is calculated using log-likelihood measures.

In this paper we employ an association measure based on the Poisson distribution, introduced by Quasthoff and Wolff (2002) for collocation extractions. Mayer and Cysouw (2012) use it to simultaneously align a large number of languages based on the co-occurrences of words in a massively parallel text. The main reason to use this association measure is that frequencies in (comparative) linguistics tend to show a strongly skewed, non-normal distribution, mostly Zipfian or negative exponential (Maslova, 2008; Cysouw, 2010; Jäger, 2012). Many widely-used association measures, like the cosine-similarity, the Pearson correlation coefficient, or pointwise mutual information/log-odds ratio, are based on the underlying assumption of a normal distribution of the data. To alleviate this problem, inverse weighting (like inverse document weighting from corpus linguistics) can be used together with those association measures. However, we decided to take a more rigorous probabilistic approach and use an association measure that directly acknowledges the underlying strongly skewed distribution of the frequency of sounds within dialects.

The Poisson distribution is used to define the probability that, for two given sounds, the 'observed co-occurrence' O is different from the 'expected co-occurrence' E. The following formula results in a probability value: a good association will yield values close to zero (e.g., with significance below 0. 01 or so).

$$P(O,E) = \frac{E^O * e^{-E}}{O!} \tag{1}$$

A measure of association is derived from this probability distribution by taking negative logarithm, which results in:

$$-\log\left(\frac{E^O * e^{-E}}{O!}\right) = E + \log(O!) - O * \log(E) \tag{2}$$

However, there are still some unwanted effects of this measure of association. The first problem is that the formula does not reach zero when $O = E$, but the result for $O = E$ turns out to depend on the size of O (which is difficult to interpret).

Fortunately, this can easily be corrected by subtracting the case with $O = E$, leading to the following somewhat surprising result.

$$[E + \log(O!) - O * \log(E)] - [O + \log(O!) - O * \log(O)] = O * \log\left(\frac{O}{E}\right) - (O - E) \tag{3}$$

The second problem is that this formula evaluates the *difference* between O and E, which implies that the result is positive for both $O > E$ and $O < E$. To separate between these crucially different cases, it is necessary to add the correct sign, namely a plus when $O > E$ and a minus when $O < E$. This can be achieved by multiplying the formula with, for example, something like $sign(O - E)$. The resulting POISSON ASSOCIATION measure to be used in this paper then becomes:

$$\text{Poisson Association} = \text{sign}(O - E) * \left(O * \log\left(\frac{O}{E}\right) - (O - E)\right) \tag{4}$$

We calculate both the observed and expected co-occurrences from the multi-aligned word transcriptions, as described in more detail in Section 4.1. The observed co-occurrences are the number of times two phones are found aligned, while the expected frequencies are the number of times two phones are expected to be aligned on the assumption that they are statistically independent of each other. We assume that the co-occurrences of phones follow the Poisson distribution, and therefore apply this association measure.

In Fig. 1 we present the distribution of the association values of all pairs of sound correspondences in our data set for all pairs of neighboring villages. The calculated association values have a strongly skewed distribution with a high number of sound pairs whose association value is < 5 ('uninteresting associations') and a very long tail representing sound correspondences that show a strong association. For comparison, in Fig. 2 we show the distribution of the association values calculated using pointwise mutual information, which show a less differentiated distribution. Although pointwise mutual information has previously successfully been applied in dialectometry to calculate the association strengths between phones, the long-tailed distribution of the Poisson association measure promises to distinguish more strongly between regular and irregular correspondences. The possibility of using yet other association measures and a detailed comparison of different measures remains to be investigated in future work.

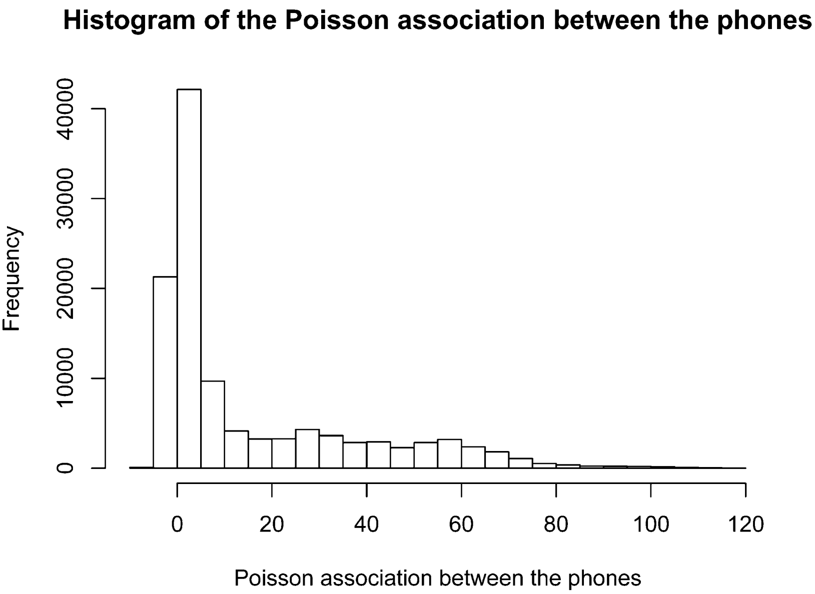

Figure 1. Histogram of the association values, calculated using the Poisson association measure

3. Data Set

3.1. *Bulgarian Dialect Data*

In this paper we use Bulgarian dialect data that consists of the pronunciations of 152 words collected in 197 villages. The data was collected from older, less mobile inhabitants of villages distributed all over Bulgaria. Only the north-eastern part of the country shows a less dense geographical sampling of villages. The distribution of the sampling sites is shown in Fig. 3. The words collected are related to every-day life, belonging to such semantic fields as food and kin terms. All data was transcribed in IPA. Detailed information on the data can be found in Prokić (2010).

3.2. *Automatic Alignment of Phonetic Data*

Phonetic transcriptions of words were aligned automatically using the Alpha-Malig software for multiple string alignment in linguistics (Alonso et al., 2004).

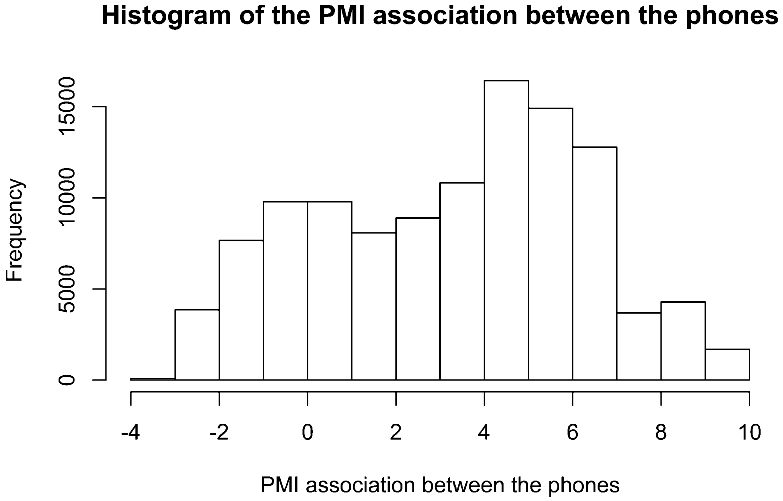

Figure 2. Histogram of the association values, calculated using pointwise mutual information

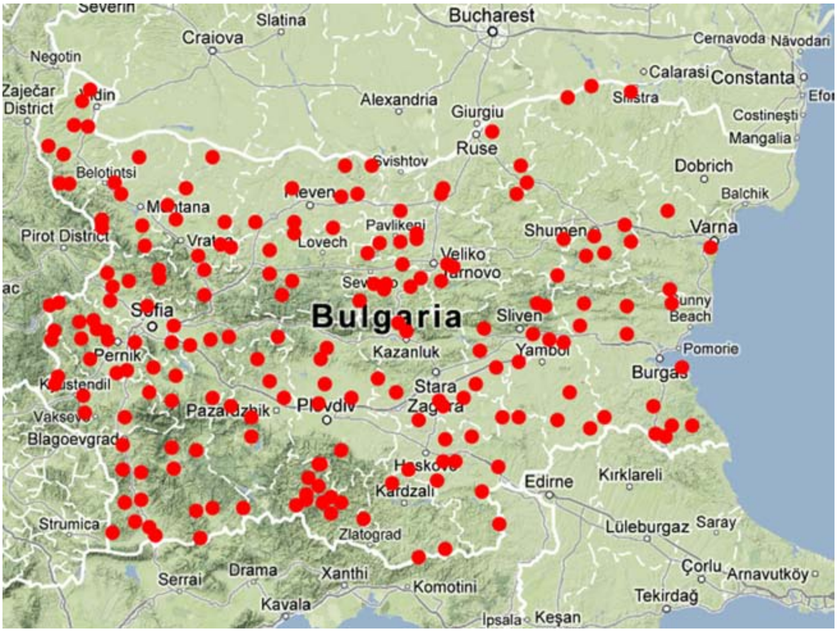

Figure 3. Distribution of the sites in the data set

Multiple string alignment is the standard approach to string comparison in biology and is considered 'the holy grail' of molecular biology (Gusfield, 1997: 332). In multiple string alignment, all strings are aligned and compared at the same time. It is therefore a good technique for discovering patterns, especially those that are weakly preserved and cannot be detected easily from sets of pairwise alignments.

The advantages of multiple over pairwise string alignment and comparison have relatively recently started being recognized in linguistics. Since strings in biology and linguistics are substantially different, with the former being long and comprising a small alphabet and the latter being very short but containing a large number of different phones, algorithms used in biology cannot be directly applied to linguistic data. In the past decade, several approaches to automatic multiple string alignment have been developed for linguistic data (Bhargava and Kondrak, 2009; Prokić, 2010; List, 2012).

In this paper, we use the Bulgarian pronunciation data, which was automatically aligned using the AlphaMalig software and subsequently manually corrected. Earlier experiments with the same data set have shown that automatically produced alignments match the manually corrected data with roughly 98 percent accuracy (Prokić et al., 2009). An example of the multiple aligned strings can be seen in Table 1. In all examples throughout the paper, the sign for stress is moved in front of the stressed vowel. The 152 words resulted in 783 alignments, i.e., 783 aligned phone positions.

Table 1. A scheme of the aligned transcriptions for words *beli* /bˈeli/ 'white' and *vyatar* /vʲˈatɤr/ 'wind' for 3 villages.

Zhelen:	b	e	l	ˈi	v	ˈe	t	ɤ	r
Zheljazkovo:	b	ˈɛ	l	i	vʲ	ˈɛ	t	e	r
Zheravna:	bʲ	ˈe	l	i	vʲ	ˈɑ	t	ɤ	r

For all our calculations we proceed from the multi-aligned phonetic pronunciations, which, in total, comprise 197 rows (one for each village) and 783 columns (for each phone alignment). This format allows us to easily extract information on regular sound correspondences, since the corresponding sounds for each position within words appear in the same column. On the other hand, information on the sounds within each of the sites in the data is organized per row and also easily accessible. We combine the two in order to examine the variation of sounds both within the lexicon and within villages and, ultimately, to detect focal and transitional areas of sound change in Bulgaria.

4. Experimental Setup

We start by comparing all the sounds found in one village with all sounds found in another village, for each pair of villages separately, and calculating the association strength between the sounds in these two villages. In the next step, we examine each of the sound correspondence sets (i.e., each of the word positions), and identify the geographic spread of the sound correspondences and their regularity. Geographic spread is investigated by projecting each village onto a map of Bulgaria and simulating a diffusion model of language change.

4.1. *Co-occurrence Statistics*

While the total number of unique phones in the data set is 98, the number of phones found in each of the 197 villages ranges from 37 for the village Golema Rakovitsa to 58 for the village Svirkovo. In order to estimate how similar the phones found in two villages are, we employ the Poisson association measure described in Section 2. We proceed from multi-aligned word transcriptions by comparing all sounds found in one village to all sounds found in all other villages, for each pair of villages separately, and extracting the co-occurrence frequencies for each pair of sounds. Each time a pair of sounds is found aligned in the data, the co-occurrence frequency for those two sounds increases by one. For each pair of villages, we compute a $n \times m$ matrix, where n is the number of phones for village 1 and m is the number of phones for village 2. The co-occurrence frequencies are used to calculate the probability of each sound recorded in one village being replaced by any other sound recorded in any other village in the data set by means of the Poisson likelihood measure. For each pair of sounds x and y, the observed frequencies O are directly read from the $n \times m$ co-occurrence frequency matrix, and represent the number of times sounds x and y are found aligned in two villages. The expected frequencies E are calculated using the following formula:

$$E_{xy} = \frac{n_x \cdot n_y}{N} \tag{5}$$

where n_x represents the number of times sound x is aligned with any other sound, n_y represents the number of times sound y is aligned with any other sound, and N is the total number of aligned phone pairs in the data. For example, if we look at the aligned transcriptions for the two neighboring villages Sredets and Izvorovo, there are 30 word positions where sound [t] is recorded in Sredets. This sound corresponds 24 times to sound [t] recorded in village Izvorovo, 1 time to palatalized t [tʲ] and 5 times corresponds to an empty position marked with '-' in Table 2, where we give an example of each type of correspondences.

Table 2. A scheme of the aligned transcriptions for 3 words for villages Sredets and Izvorovo. The sound [t] recorded in Sredets and the corresponding sounds, as well as the absence of a corresponding sound, in village Izvorovo are given in bold.

Sredets:	zʲ	ˈe	**t**	dʲ	ˈe	s	e	**t**	s	**t**	r	ˈɑ	h
Izvorovo:	zʲ	ˈe	**tʲ**	dʲ	ˈe	sʲ	ə	**t**	s	-	r	ˈɑ	h

In order to calculate the association strength between sound [t] in Sredets and sound [tʲ] in Izvorovo, we organize the co-occurrence frequencies in a contingency table (Table 3).

Table 3. A contingency table of the sound co-occurrence frequencies. S_S stands for the sounds found in Sredets and S_I stands for the corresponding sounds found in Izvorovo.

	$S_I = [t^j]$	$S_I \neq [t^j]$
$S_S = [t]$	1	29
$S_S \neq [t]$	6	747

This table stores the following information:

- number of times sound [t] from Sredets ($S_S = [t]$) corresponds to the sound [tʲ] from Izvorovo ($S_I = [t^j]$), i.e., observed (O) frequency: 1
- number of times sound [t] from Sredets ($S_S = [t]$) corresponds to any sound other than [tʲ] ($S_I \neq [t^j]$): 29
- number of times sound [tʲ] from Izvorovo ($S_I = [t^j]$) corresponds to any sound other than [t] ($S_S \neq [t]$): 6
- number of correspondences where neither [t] from Sredets ($S_S \neq [t]$) or [tʲ] from Izvorovo ($S_I \neq [t^j]$) are present: 747

From the frequencies list in Table 3, we extract the observed frequency ($O = 1$), and calculate the expected frequency for the correspondence in question:

$$E = \frac{(29 + 1) * (6 + 1)}{(29 + 1 + 6 + 747)} = 0.268199$$

Equation 4 gives us the following value for the Poisson association between these two sounds:

$$Poisson_{[t]:[t^j]} \approx 0.584225$$

This value represents a very low association between two sounds, which is expected given that they are found aligned only once. We apply the described procedure for each pair of sounds, for each pair of villages separately.

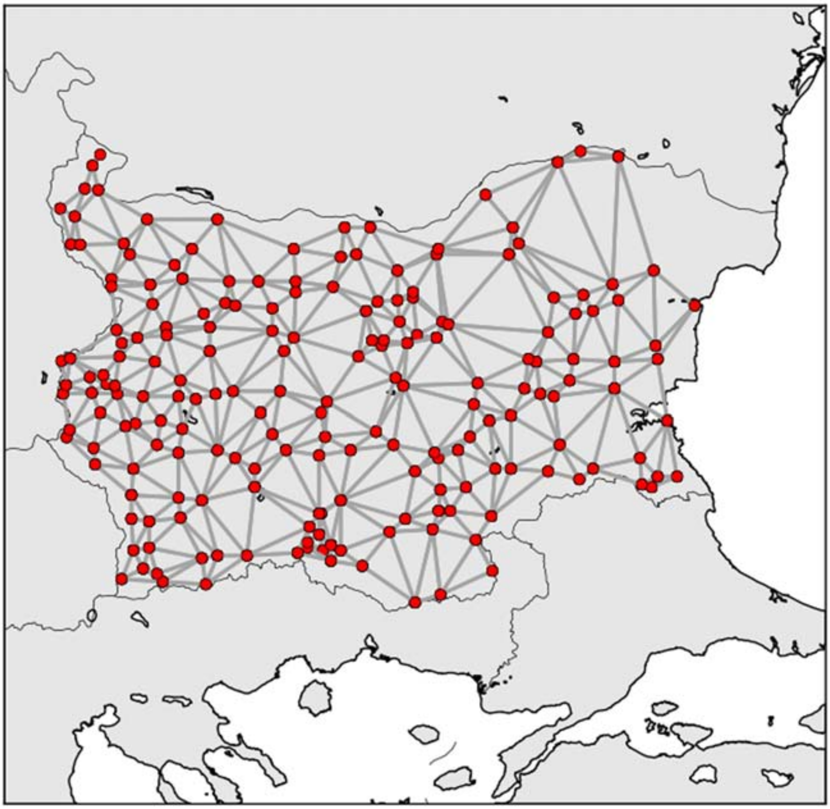

Figure 4. Each site in the data is connected only to neighboring sites

Although all words in our data set are transcribed in IPA, we do not assume any similarity or dissimilarity between the phones based on the IPA notation, i.e., articulatory features of the phones. The association between the phones is based exclusively on how often they co-occur in the multi-aligned strings, i.e., on the regularity of sound change, rather than phonetic similarity. Because we also know the phonetic value of each sound, the results of our method are easily interpretable linguistically. However, in principle we could have used e.g. unique numbers for each phone from each village, and still obtained the same results.

4.2. *Employing Geographic Data in Linguistics*

In order to examine the geographic spread of sound changes and their regularity, we plot all sites in the data set on a map of Bulgaria. We employ the

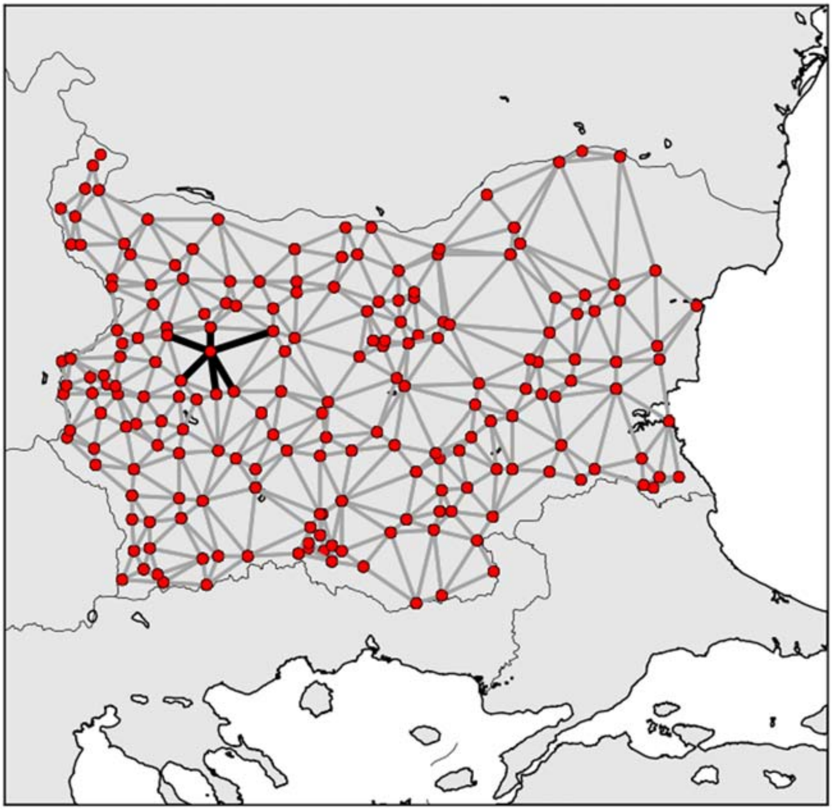

Figure 5. Innovations spread in a wave-like manner to the neighboring villages. Here we illustrate the spread of change from the village of Vrachesh to the neighboring villages

Delaunay triangulation method (Delaunay, 1928) to connect a set of points (sites) by a set of triangles. As a result, we obtain a network connecting all localities in a such way that each site is connected only to the neighboring sites (see Fig. 4). The Delaunay triangulation is a method developed in mathematics and geometry, but is frequently used in many domains, including dialectometry where it was first used by Goebl (1984). Network representation of the localities allows us to explore the spread of sound change via diffusion. Each innovation spreads from a village, a potential center of innovation, to the neighboring villages. An example can be seen in Fig. 5. In the next step, the innovation spans further to the neighboring villages of the neighboring villages, in a wave-like manner.

We combine the regularity of sound change and its geographic spread by shading the edges between the sites according to the association strength between the sounds in question. We use light ('transparent') shades if the association between two phones is high and the sound correspondence is regular. The less regular a sound correspondence between neighboring sites, the darker the shades, which allows us to easily detect areas where the sound change is irregular. Examples of maps can be seen in the next section, where we present the results of the analyses. As will be seen, often the same phones in neighboring villages represent a regular correspondence (light shaded) and different phones in neighboring villages represent irregular correspondence (dark shaded in the figures). However, more interestingly, we sometimes observe the same phone in neighboring villages without a regular correspondence between these phones. This could be a sign of a transition area, in which a spread has been only partially established.

5. Results

The multi-aligned word transcriptions of the Bulgarian dialect data analyzed in this paper contain 783 columns, i.e., word positions. Each column represents a set of sound correspondences that has a different pattern of geographic spread and regularity. In our data we distinguish two different types of sound correspondence sets. The first type comprises word positions where there is no or almost no variation. Those are, in the vast majority of cases, consonants in the onset syllable position. Some positions occupied by stressed vowels are also characterized by very small variability. For example, the sound [p] in the word *pat* /pˈɤt/ 'road' remains unchanged in all sites. Plotting this correspondence on a map, we can see that there is a high association between the phones from the neighboring sites, shown by very light lines (Fig. 6). So, in this figure we do not see any irregularity.

The second type of correspondence sets comprises those that show much greater variation, and most of the correspondence sets of this type in our data set are unstressed vowels. Regarding the regularity of this kind of sound correspondences, it can be a) highly regular across the whole examined region, b) irregular only in certain areas, or c) highly irregular in all geographic areas. An example of sound correspondences of type (a) different phones whose correspondences are highly regular in all areas can be seen in Fig. 7, where we show the variation of the standard Bulgarian sound [ˈɤ] in the word *mazh* /mˈɤʒ/ 'man.' Although the correspondence set contains five different phones [ˈu, ˈɑ, ˈa, ˈɔ, ˈɤ], going from one village to any of the neighboring villages, it is highly predictable which phone will be encountered. Only in the northwest we see some irregularity (indicated by the dark lines). Such regular correspondences between different phonetic reflexes indicates a strong vertical phylogenetic signal.

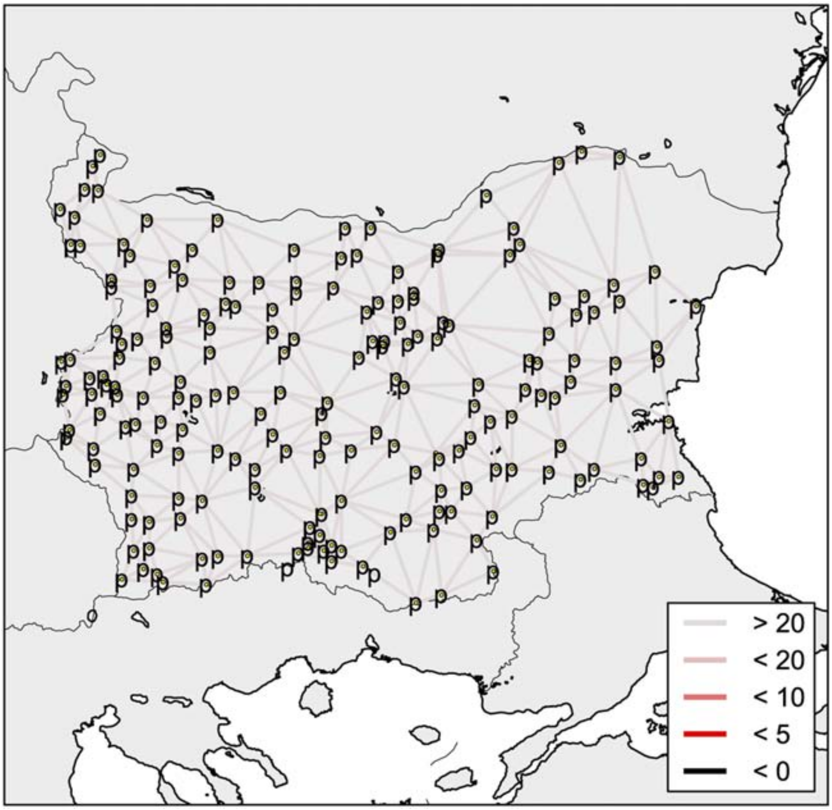

Figure 6. Highly stable sounds, like the sound [p] in word *pat* /p'ɤt/ 'road,' show high association strength. Note that regular association is shown with lighter colours. The depicted values are the negative logarithm of the Poisson association

However, some correspondence sets show much less regularity. This type of correspondence sets is highly interesting, since it allows us to detect transitional areas of sound change and, in some cases, to infer the focal areas from which the change begins to spread. An example of this type of sound change can be seen in Fig. 8, where the variation of the standard Bulgarian sound [e] in the word *kade* /k'ɤde/ 'where' is plotted on the map. The sites with irregular sound correspondences are concentrated in the west of the country, with a few further irregularities in the north-east. Another example of a sound change spread, with a completely distinct geographical distribution of (ir)regular correspondences, is represented in Fig. 9. This map shows the variation of the second ['a] in the word *glava* /glav'a/ 'head.' The irregularity is found in the middle of the country and

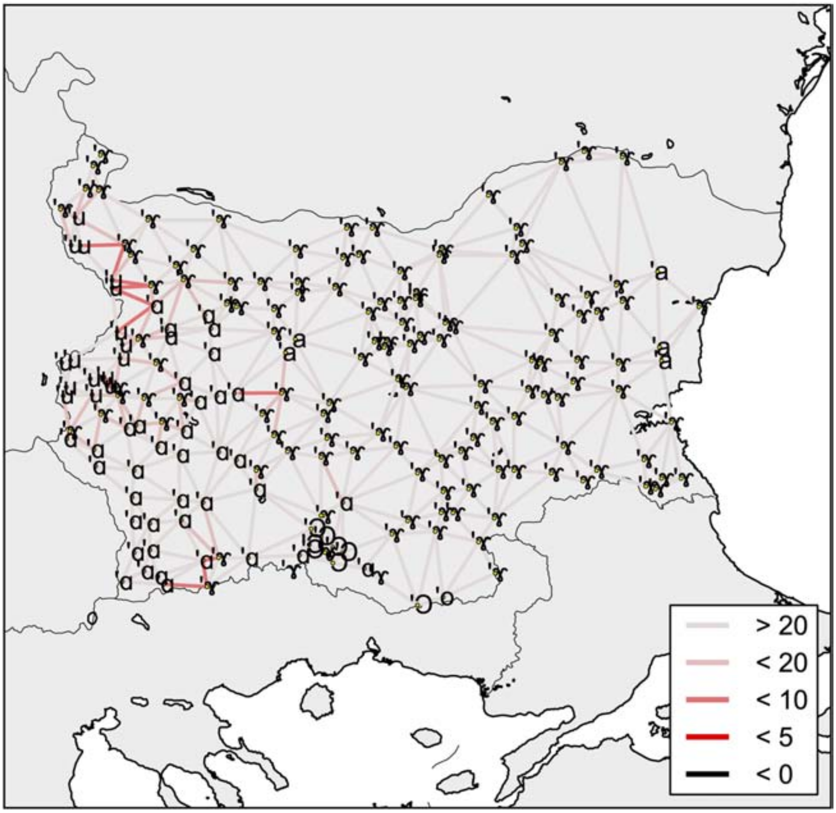

Figure 7. Some highly variable sounds, like the sound [ˈɤ] in the word *mazh* /mˈɤʒ/ 'man,' still show high association strength, which indicates that the change has already spread

follows the so-called 'yat' line, marked with a black thick line on our map, which goes from the north to the south, and is the most important dialect border in Bulgaria according to traditional dialectologists (Stoykov, 2002). This irregularity on the boundary indicates that there has been local diffusion of individual words changing their pronunciation. There is some additional irregularity in the southeast of the country. Of course, there are also sound correspondence sets that show high irregularity in all geographical areas, but this type is not helpful in the detection of geographical distributions of sound change.

Examining individual word positions allows us to gain a detailed and precise insight into the geographic spread of sound correspondences and their regularity. Different sound changes have different focal areas and a different strength of

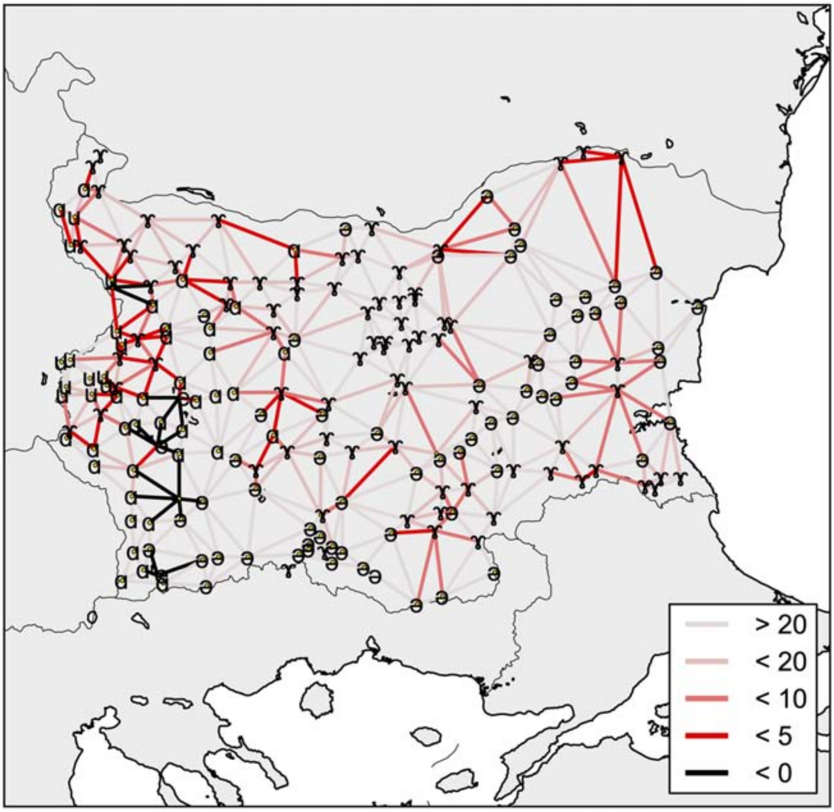

Figure 8. Here we illustrate the variation of sound [e] in word *kade* /kˈɤde/ 'where.' Sites with irregular correspondences are connected with dark lines, which indicate transition areas. The depicted values are the negative logarithm of the Poisson association

spread through a population of villages. They can arise at any time in any location and spread in very different, even opposite, directions (Hock, 1991). However, for a variety of cultural, political and historical reasons, some places are sources of numerous innovations that can spread to a geographically very distant area. In order to abstract away from the individual cases and look for the geographical areas that are more likely to be centers of observed innovations, in the next step, we aggregate over all word positions in our data. We seek areas that are characterized by highly irregular sound correspondences in the majority of word positions examined. For each site we calculate the number of times we encounter irregular correspondences between a site and any of the neighboring sites, which gives us an *irregularity index* for each edge between the neighboring sites in the

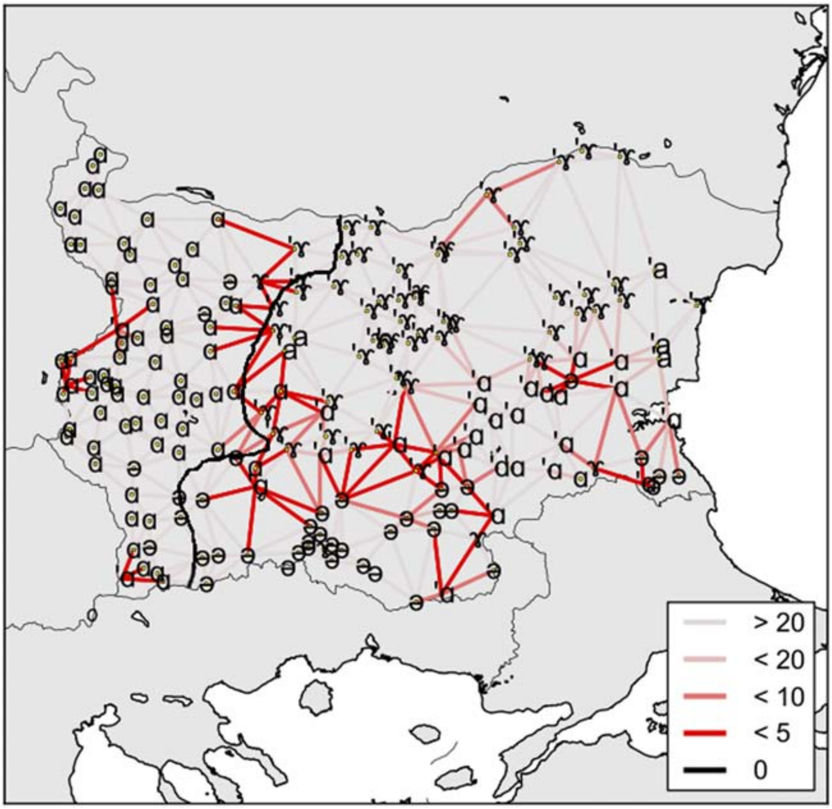

Figure 9. This map shows the variation of the second [ˈa] in word *glava* /glavˈa/ 'head.' Irregular correspondences marked with dark lines, indicating transition areas, are found along the north-to-south 'yat' line, marked with the black tick line, and in the south

data. To establish the irregularity index, we use an extremely conservative cut-off point of the association values (lower than 5), i.e., we are only counting the number of neighboring correspondences that are strongly irregular. Each time a sound correspondence below a certain threshold is encountered, the irregularity index increases by one. In the last step, the index is normalized so that all values lie between 0 and 1:

$$Irr = \frac{N_{irr} - max}{max - min} \quad (6)$$

where N_{irr} represents the number of irregular correspondences between two sites, *max* represents the maximal number of irregular correspondences in the

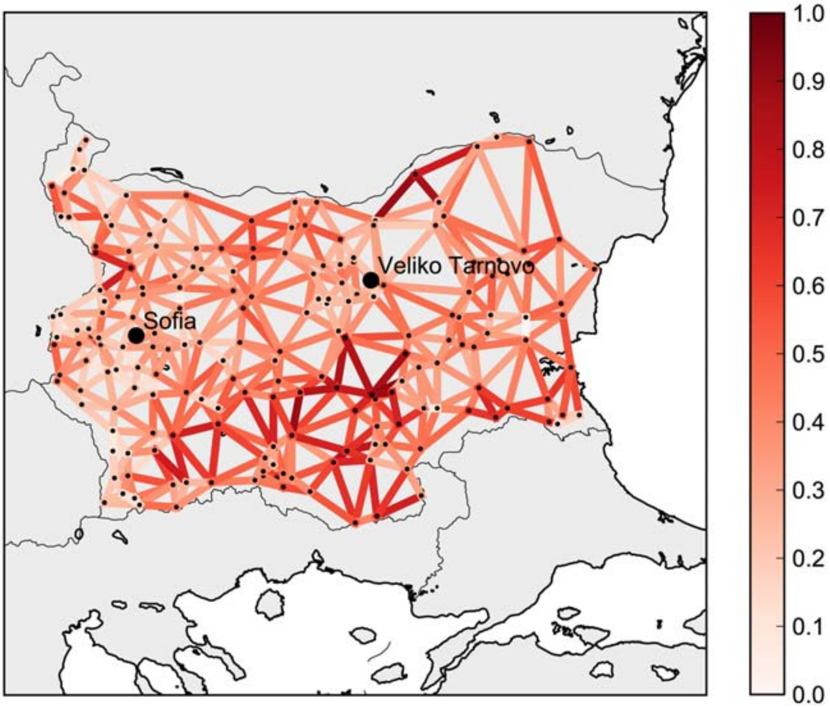

Figure 10. Edges with a higher number of correspondences whose association strength is below 5 are colored darker and are found in the south of the country. Areas in the north-west, south-west and central-east are characterized by highly regular sound correspondences and we represent them with lighter shades

data, and *min* represents the minimal number of irregular correspondences in the data. In order to visualize the *irregularity index (Irr)*, we connect all sites in the data to all the neighboring sites, and color the edges between them according to the value of the index. We present the results in Fig. 10.

The map reveals no sharp borders. It confirms that the spread of linguistic change goes in different directions and that every area can be *focal* in some cases and *transitional* or *relic* in others. However, there are three areas on the map in Fig. 10 that show higher regularity than the rest of the country (indicated by lighter lines), namely the western part of the country, especially the south-western and north-western parts, and the central area that spreads to the north-east. The high regularity of sound correspondences suggests that these three areas are centers of many innovations that spread in a wave-like manner to the rest of the country. In two of these three areas we actually find the two biggest towns in Bulgaria, Sofia and Veliko Tarnovo—the former being the biggest city and the capital since 1879, and the latter having been a very important educational and cultural center of Bulgaria for centuries. The dialect of Veliko Tarnovo had prestigious status for centuries and is nowadays the closest to Standard Bulgarian. The spread of innovations from these two centers can be seen very clearly in our analyses. At the same time, an area in the south of the country shows higher irregularity of sound changes and has much darker shades. This area is the least likely to be an area of sound innovations, since it is characterized by high irregularities of sound correspondences, which is typical of transitional areas. It is an area of high mountains; in traditional literature on Bulgarian dialects, we find that this area exibits large dialectal diversity due to, among other factors, a geographical landscape that makes everyday communication harder, compared to more flat areas (Stoykov, 2002). Our findings are in line with this characterization of the southern dialect area in Bulgaria.

6. Conclusions

In this paper we present an approach to dialectometry that uses co-occurrence statistics combined with methods taken from geographical sciences in order to answer the questions of the geographic origins of sound change and its propagation through lexicon and space. Unlike many previous studies in this field, the main interest of this research is not on the identification of the main dialect groups, but on the underlying models of linguistic change. By combining statistical methods and visualization, we are able to track the spread of individual sound changes, but also to aggregate over many of them and detect *focal*, *transitional* and *relic* areas of the spread of change.

The central innovation of our study is to investigate regular co-occurrences between sounds without relying on the phonetic similarity between them. This approach approximates the notion of regular sound correspondence, which is central to historical linguistics and dialectology. So, instead of looking for superficial sound similarity, we are focussing on regularity of the correspondences (possibly between different sounds). We employ the diffusion model of language change in our quantitative analysis by comparing each site only to the neighboring sites, which gives us good insight into the underlying linguistic processes. However, we are able to abstract away from the individual features, both individual sound correspondences and one-to-one village relations, and search for more general patterns in the data while basing our analyses on transparent, linguistically-motivated models.

We test our method on the dialect data, but the suggested procedure is not restricted to dialect data. It can easily be applied to any language data, provided that at least some more or less regular sound correspondences can be identified and that geographic information about the locations is available. In this work we rely on the Poisson association measure to infer the similarity between the sounds and on the diffusion model to map the spread of change. The usage of alternative association measures for extracting phone similarity and models other than the diffusion model can be easily operationalized, but remains to be investigated in more detail in future work.

Acknowledgments

This work was supported by the ERC starting grant 240816: 'Quantitative modelling of historical-comparative linguistics.'

References

Alonso, Laura, Irene Castellón, Jordi Escribano, Xavier Messeguer, and Lluís Padró. 2004. Multiple sequence alignment for characterizing the linear structure of revision. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)*, 403–406.

Bhargava, Aditya and Grzegorz Kondrak. 2009. Multiple word alignment with Profile Hidden Markov Models. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Student Research Workshop and Doctoral Consortium*, Boulder, Colorado: Association for Computational Linguistics, 43–48, URL http://www.aclweb.org/anthology/N/N09/N09--3008.

Chambers, J.K. and Peter Trudgill. 1998. *Dialectology*. Cambridge: Cambridge University Press.

Cysouw, Michael. 2010. On the probability distribution of typological frequencies. In Ebert, Christian, Gerhard Jäger, and Jens Michaelis (eds.) *The Mathematics of Language*, Berlin: Springer, 29–35.

Cysouw, Michael and Hagen Jung. 2007. Cognate identification and alignment using practical orthographies. In Nerbonne, John, T. Mark Ellison, and Grzegorz Kondrak (eds.) *Proceedings of Ninth Meeting of the ACL Special Interest Group in Computational Morphology and Phonology*, Prague: Association for Computational Linguistics, 109–116.

Delaunay, Boris. 1928. Sur la sphère vide. In *Proceedings of the International Mathematical Congress Held in Toronto, August 11–16*, 695–700.

Evert, Stefan. 2005. *The Statistics of Word Cooccurrences: Word Pairs and Collocations*. PhD dissertation, Institut für maschinelle Sprachverarbeitung, University of Stuttgart.

Goebl, Hans. 1982. *Dialektometrie: Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie in Bereich der Dialektgeographie*. Wien: Österreichische Akademie der Wissenschaften.

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

——. 2006. Recent advances in Salzburg dialectometry. *Literary and Linguistic Computing* 21: 411–435.

Gusfield, Dan. 1997. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge: Cambridge University Press.

Heeringa, Wilbert. 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. PhD dissertation, University of Groningen.

Hock, Hans Henrich. 1991. *Principles of Historical Linguistics*. Berlin: Walter de Gruyter.

Jäger, Gerhard. 2012. Power laws and other heavy-tailed distributions in linguistic typology. *Advances in Complex Systems* 15.

List, Johann-Mattis. 2012. LexStat: Automatic detection of cognates in multilingual wordlists. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, Avignon: Association for Computational Linguistics, 117–125.

Mackay, Wesley and Grzegorz Kondrak. 2005. Comparing word similarity and identifying cognates with Pair Hidden Markov Models. In Dagan, Ido and Daniel Gildea (eds.) *Proceedings of the 9th Conference on Natural Language Learning (CoNLL)*, Ann Arbor, MI: Association for Computational Linguistics, 40–47.

Maslova, Elena. 2008. Meta-typological distributions. *Sprachtypologie und Universalienforschung* 61: 199–207.

Mayer, Thomas and Michael Cysouw. 2012. Language comparison through sparse multilingual word alignment. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, Avignon: Association for Computational Linguistics, 54–62.

Nerbonne, John and Wilbert Heeringa. 1998. Computationele vergelijking en classificatie van dialecten. *Taal en Tongval; Tijdschrift voor Dialectologie* 50: 164–193.

Nerbonne, John, Wilbert Heeringa, Eric van den Hout, Peter van de Kooi, Simone Otten, and Willem van de Vis. 1996. Phonetic distance between Dutch dialects. In Durieux, Gert, Walter Daelemans, and Steven Gillis (eds.) *CLIN VI, Papers from the Sixth CLIN Meeting*, Antwerpen: University of Antwerpen, 185–202.

Nerbonne, John, Wilbert Heeringa, and Peter Kleiweg. 1999. Edit distance and dialect proximity. In Sankoff, David and Joseph Kruskal (eds.) *Time Warps, String Edits and Macromolecules: The Theory and Practice of Sequence Comparison*, Stanford: CSLI Press, v–xv.

Nerbonne, John, Peter Kleiweg, Wilbert Heeringa, and Franz Manni. 2008. Projecting dialect differences to geography: Bootstrap clustering vs. noisy clustering. In Preisach, Christine, Lars Schmidt-Thieme, Hans Burkhardt, and Reinhold Decker (eds.) *Data Analysis, Machine Learning, and Applications. Proceedings of the 31st Annual Meeting of the German Classification Society*, Berlin: Springer, 647–654.

Prokić, Jelena. 2010. *Families and Resemblances*. PhD dissertation, University of Groningen.

Prokić, Jelena, Çağrı Çöltekin, and John Nerbonne. 2012. Detecting shibboleths. In *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*, Avignon: Association for Computational Linguistics, 72–80, URL http://www.aclweb.org/anthology/W12--0211.

Prokić, Jelena, Martijn Wieling, and John Nerbonne. 2009. Multiple sequence alignments in linguistics. In *Proceedings of the EACL 2009 Workshop on Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH—SHELT&R 2009)*, Stroudsburg, PA: Association for Computational Linguistics, 18–25, URL http://www.aclweb.org/anthology/W09--0303.

Quasthoff, Uwe and Christian Wolff. 2002. The Poisson collocation measure and its applications. In *Proceedings of the 2nd International Workshop on Computational Approaches to Collocations*, Vienna.

Séguy, Jean. 1971. La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane* 35: 335–357.

Steiner, Lydia, Peter F. Stadler, and Michael Cysouw. 2011. A pipeline for computational historical linguistics. *Language Dynamics and Change* 1: 89–127, URL http://www.ingentaconnect.com/content/brill/ldc/2011/00000001/00000001/art00004.

Stoykov, Stoyko. 2002. *Balgarska dialektologiya* [Bulgarian Dialectology]. Sofia: Marin Drinov Academic Publishing.

Tiedemann, Jörg. 1999. Automatic construction of weighted string similarity measures. In *Proceedings of the Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora (EMNLP/VLC)*, College Park, MD: University of Maryland, 213–219.

Wieling, Martijn, Eliza Margaretha, and John Nerbonne. 2012. Inducing a measure of phonetic similarity from pronunciation variation. *Journal of Phonetics* 40: 307–314.

Wieling, Martijn and John Nerbonne. 2011. Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguistic features. *Computer Speech and Language* 25: 700–715.