Analyzing Dialects Biologically

Jelena Prokić and John Nerbonne (LMU Munich and University of Groningen)

Abstract

This paper sketches some parallels in analyzing linguistic and biological data, emphasizing how methods taken from biology have been applied in dialectometry. We also present a technique developed specifically for detecting distinctive properties in (putative) groups of linguistic varieties (species) as this illustrates how the linguistic enterprise perhaps differs from the biological one.

1 Introduction

Dialectometry is a branch of linguistics whose main goal is the development and application of quantitative methods that enable researchers to explore relationships among dialects in an analytical way while taking into account large amounts of data. Most of the work done so far in dialectometry has focused on the differences between dialect varieties at the lexical and phonetic level, i.e., differences in vocabulary and pronunciation. However, there are projects that were concerned with the differences at the morphological (internal word

structure)and syntactic level (concerned with structure in phrases and sentences). Regardless of the level at which the differences between the dialects are investigated, dialectometry has benefited from related developments in biology, especially those in population genetics and phylogenetics. These include use of sequence alignment techniques, hierarchical clustering, bootstrapping, and dimensionality reduction techniques, just to name a few.

This paper presents a line of research in dialectometry where the distances between the dialects are measured at the phonetic level. Application of quantitative methods to dialect pronunciation data consists of three major steps a) measuring distances, which in this paper will done via string alignment/distance algorithms; b) detection of dialect groups; and c) linguistic interpretation. While any of the three major steps could be a subject of a separate survey paper, we focus here on the last step (c), which is concerned with identifying the linguistic basis for the automatic classification of dialects. We show how recent work that automatically identified characteristic *words* in given regions may easily be extended to allow the automatic identification of *sounds* characteristic of a given region. We also briefly present the first two steps for those less familiar with dialectometry, and we sketch some of the problems present in the quantification of dialect data (language data in general) using some of the

mentioned methods. We first give short description of the data used throughout this paper.

2 Data set

In this paper we use Bulgarian dialect data that comes from the project *Buldialect – Measuring Linguistics Unity and Diversity in Europe*[1] to illustrate various methods used in dialectometry to measure and visualize the data. The Buldialect data set consists of the pronunciation of the 157 words collected at 197 villages distributed all over Bulgaria. Words included are frequent words that were collected from all, or almost all of the 197 sites. Regarding the choice of words, only words which are expected to show some degree of phonetic variation were included. There are in total 39 different dialectal features which have been represented in the chosen 157 words. The full list of 157 words and dialectal features present in these words can be found in Prokić et al. (2009) and Houtzagers, Nerbonne and Prokić (2010). Five words that have lower coverage than the rest of the words were excluded form all experiments presented in this paper, and the total number of words that we work with is 152.

[1] Volkswagen Foundation grant to P.I. Prof.Erhard Hinrichs, Tübingen.

3 String alignment

The fist step in the quantitative analyses of dialect phonetic variation is to measure the distances between various pronunciations in the data set. Hoppenbrouwers & Hoppenbrouwers (2001) address this problem by computing the differences of relative phone frequencies in various dialects (phones are individual sounds such as the ‘p’ sound in ‘pail’, or the ‘l’ sound). They also proposed a similar method based on the differences of the relative frequencies of different articulatory feature values of phones (features are properties, such as the property of being a vowel, or the property of being pronounced using an obstruction of the vocal path at the two lips, as in ‘p’). Both of these frequency-based approaches do not take into account the ordering of the phones in a word. In order to make our measurements sensitive to the ordering of phones in a word, we must first align two pronunciations by means of the sequence (or string) alignment algorithms. We present two approaches to automatic string alignment used in dialectometry in the next two subsections.

3.1 Pairwise alignment

String alignment techniques have been introduced into dialectometry with the work of Brett Kessler who has used Levenshtein algorithm to calculate the pronunciation distance between the Irish Gaelic dia-

lects (Kessler, 1995). Application of the Levenshtein algorithm in dialectometry was later further developed and improved at the University of Groningen and applied to many languages in order to detect main dialect groups: Dutch (Nerbonne at al., 1996; Heeringa, 2004), Sardinian (Bolgnesi & Heeringa, 2002), Norwegian (Gooskens & Heeringa, 2003), German (Nerbonne & Siedle, 2005) and Bulgarian (Osenova et al., 2009).

The Levenshtein, or string edit distance, algorithm (Levenshtein, 1966) is a dynamic programming algorithm used to measure the distance between two strings. The distance between two strings is defined as the smallest number of insertions, deletions and substitutions needed to transform one string to the other. We illustrate how one pronunciation of Bulgarian word *аз* 'I', namely [jɑ] (Aldomirovtsi) can be transformed into another [ɑs] (Asparuhovo):

j	ɑ	-
-	ɑ	s
1		1

The minimal number of required operations is two: [j] has to be inserted/deleted in the word initial position, and [s] has to be inserted/deleted in the word final position. If the cost of each operation is

1, then the Levenshtein distance between these two strings is 2, and 2/3 if the distance is normalized by the length of the alignment. Treating the differences between phones in a binary fashion, i.e. same or not the same, is very simplistic model of sound change and for that reason very often unpopular among linguists. The cost of replacing one segment by another can be made more sensitive by basing it on articulatory features (Heeringa, 2004) or automatically induced from the alignments (Prokić, 2010; Wieling, Margaretha & Nerbonne, 2012). The choice of the operation weights depends on the research goal. Whether or not one uses a segment weighting scheme leads to only minor differences in measurements at the aggregate level (Heeringa, 2004). If one is interested in the more detailed analysis of the alignments, e.g. extraction of regular sound correspondences, then using a differential segment weighting produces more accurate alignments (Wieling, Prokić and Nerbonne, 2009) and is better suited for the dialect analysis at the segment level.

In order to calculate distances between each pair of sites in the data set, each pronunciation of a given word collected at one site is compared to the pronunciation of the same word at the other site by means of the Levenshtein algorithm. The distance between two sites is the mean of all word distances calculated for those two sites. The final result is a *site x site* distance matrix.

3.2 Multi-string alignment

Another approach to string alignment is multiple string alignment where all strings are aligned and compared at the same time. Automatic multiple string comparison is considered *the holy grail* of molecular biology (Gusfield, 1997, 332). This type of string comparison, albeit executed manually, rather than automatically, has played a central role in linguistics ever since the late 19th century and the development of the comparative method of linguistic reconstruction (Campbell, 2004). In the comparative method, identification of regular sound changes has played a major role in the identification of genetically related languages. The correct analysis of sound changes requires the simultaneous examination of corresponding sounds in the multiply aligned strings. Historical linguists align the sequences manually. In recent decade several algorithms for multiple string alignment in linguistics were developed (Bhargava & Kondrak, 2009; Prokić, Wieling & Nerbonne, 2009; Steiner, 2011; List, 2012). In Prokić et al. (2009), the ALPHAMALIG algorithm was applied to dialect pronunciation data for the first time to multi-align word pronunciations. We illustrate the results of automatically aligning six pronunciations of word *аз* 'I':

j ɑ - - - -

- ɑ s - - -

j ɑ z e - -

j ɛ - - - -

j ɑ z e k a

- ɒ s - - -

The advantages of this type of alignment are twofold:

- First, it is easier to detect and process corresponding phones in words and their alternations (like [ɑ] and [ɛ] and [ɒ] in the above example).
- Multi-aligned strings, unlike pairwise aligned strings, contain information on the positions where phones are inserted or deleted in both strings. This leads to different distances between the strings as compared to the pairwise approach. In multi-aligned comparison the number of mismatching phones between [jɑ] and [ɑs] is 2/6 while it is only 2/3 if assayed based on the isolated pair (if distances are normalized).

Evaluation of the alignments automatically produced by ALPHA-MALIG has shown that for the Buldialect data set is above 93 per cent when compared to a manually corrected “gold standard” (Prokić et al., 2009).

The distances between the aligned strings can be calculated by counting the number of mismatching positions in a binary fashion or using some of the weighting schemes mentioned above.

4 Detection of groups

Once the distances between each pair of sites (villages) have been calculated, groups of dialects and their relatedness have to be reconstructed based on the estimated distances. Below we mention some of the methods most frequently used in dialectometry.

4.1 Clustering

A distance matrix that contains information on the distances between each pair of villages in the data set can be analyzed using clustering techniques, and later projected onto a map to check the geographical distribution of the groups obtained. Hierarchical clustering techniques were introduced into dialectometry by Hans Goebel (1982; 1983) who was the first to use clustering in analyses of dialect variation. He performed cluster analysis to detect the most important dialect groups and show their geographical spread by coloring groups detected by clustering differently on the map. Ever since, clustering

has been commonly used to group dialects and analyze their relationship. In Figure 1 we present the dendrogram and the projection of the detected groups on the map of Bulgaria generated using Gabmap dialectometry software (Nerbonne et al., 2011) developed at the University of Groningen.[2]

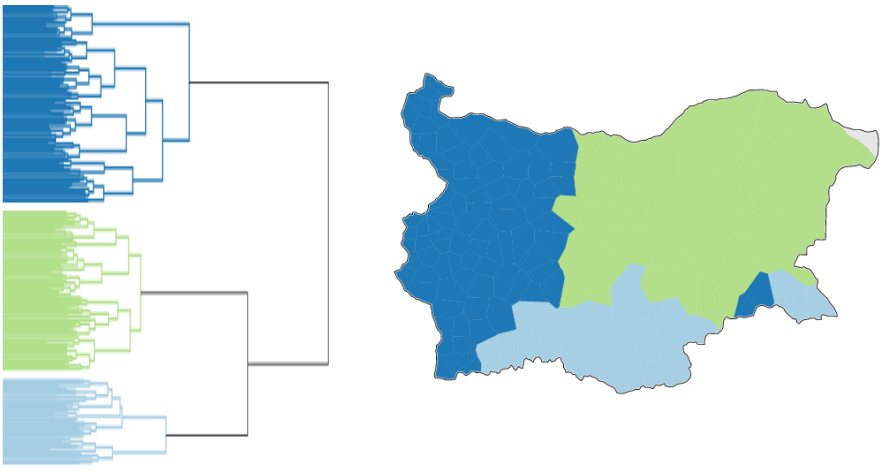

Figure 1: Dialect groups in Bulgaria identified using Ward's clustering method.

[2] http://www.gabmap.nl/

However, clustering techniques produce unstable results meaning that very small differences in the input matrix can lead to very different groupings of the data (Jain et al., 1988). In biology, in order to obtain stable clustering results a bootstrap procedure is often employed by randomly resampling the observed data (Felsenstein, 2004). In dialectometry, Nerbonne et al. (2008) introduced *noisy* or *composite clustering*, in which small amounts of random noise are added to the matrices during repeated clustering. Tested on dialect data, bootstrapping and noisy clustering produce distance matrices that correlate nearly perfectly ($r = 0.997$). Unlike bootstrapping, noisy clustering can be applied on a single distance matrix which makes it easily applicable in dialectometry if distances between the sites are represented by a *site x site* distance matrix.

4.2 Network representation

There is a problem with using hierarchical clustering to determine the historical relationship among dialects, namely that this approach assumes an underlying tree model of dialect change. The relations between the groups produced by hierarchical clustering are frequently represented by a bifurcating tree diagram called dendrogram (as shown in Figure 1). This representation of language relatedness sug-

gests that the innovations occur exclusively in the process of transmission from a mother language variety to daughter varieties. Just as in biology, bifurcating phylogenetic trees are used to model acquisition by inheritance only. Already in the 19th century, Johannes Schmidt (1872) argued that innovations in languages are spread through borrowing, i.e. argued for non-hierarchical diffusion of linguistic innovations from multiple sources. Borrowings that occur between languages correspond to lateral transfer in biology, and it cannot be modelled using tree representation. In order to visualize evolutionary relationships that include lateral tranfer, biologists use phylogenetic networks. One of the most popular method for reconstructing phylogentic networks is Neighbor-Net, available as part of the Splits Tree software (Huson & Bryant, 2006).[3] In the past ten years there have been an increasing number of studies in linguistics that use this method to infer and visalize the relationships between language varieties. One important property of the Neighbor-Net algorithm is that, if the input distance is circular, it will return the collection of circular splits, i.e. the network. If the input distance is additive, on the other hand, it will return the corresponding tree (Bryant and Moulton, 2004). This property enables researchers to see if the data is tree-like or network-like. In Figure 2 we use Neighbor-Net to analyze the

[3] http://www.splitstree.org/

same distance matrix used to produce dendrogram in Figure 1. By visually inspecting the network, we can identify three groups in the

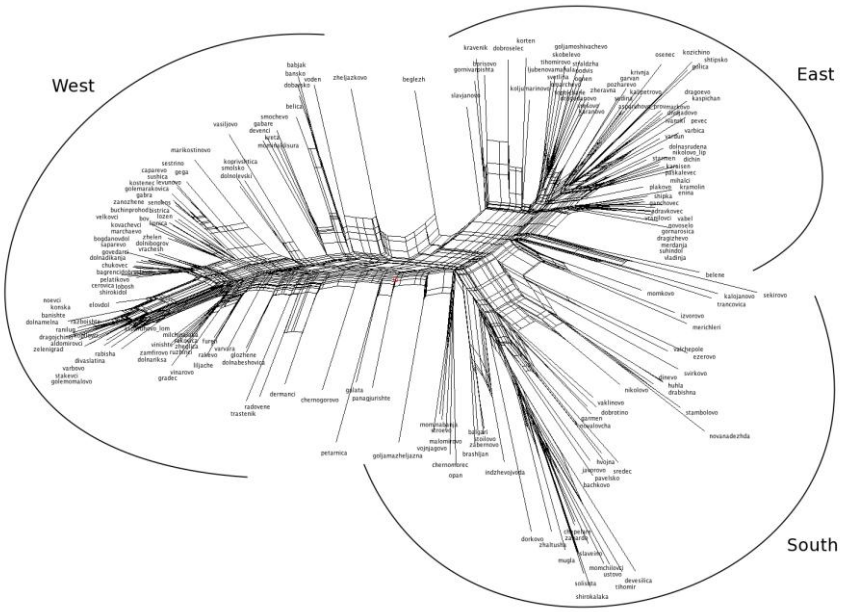

Figure 2: Neighbor-Net detects network-like structure of the data.

data, namely the West, East and South. However, the network representation allows us to see that there are many conflicting signals represented as reticulations (lines connecting radial branches), which makes the data look more network-like than tree-like. This is by no means a surprising result, since dialects often form a continuum ra-

ther than groups of clearly separated varieties (Chambers & Trudgill, 1998). The innovations are spread through borrowing and extensive social contact. For that reason, networks are more realistic representation of the relations between dialect varieties. Unfortunately, there is no direct way to link this kind of representation and geographic data, i.e. project data on the map, which is very important element of the research in traditional dialectology and in dialectometry as well. Another method frequently used in biology, namely multidimensional scaling, allows us to represent dialect variation as a continuum and project the results on the map.

4.3 Multidimensional scaling

Multidimensional scaling is a dimensionality-reducing technique used in exploratory data analysis and a data visualization method, often used to look for separations of data groups (Legendre & Legendre, 1998). It analyses the set of distances between elements and attempts to arrange elements in a space within a certain small number of dimensions, which, however, accord with the observed distances. It was used for the first time in linguistics by Black (1973) and in dialectology by Embleton (1993). The plot of the first two extracted MDS dimensions obtained by applying multidimensional scaling to our distance matrix is presented in Figure 3, where

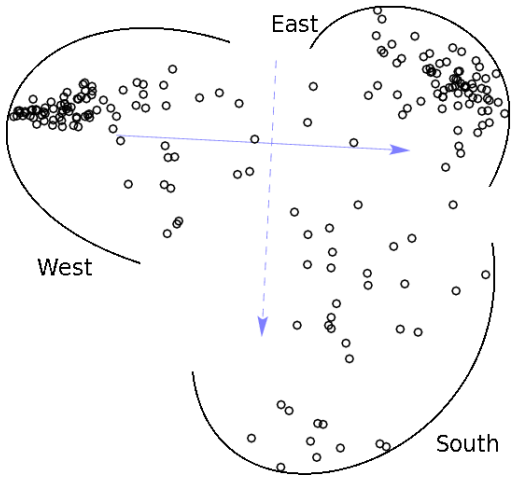

Figure 3: MDS-plot of the first two extracted dimensions

the MDS plot shows two relatively homogeneous groups of varieties, West and East, and a third heterogeneous group that includes varieties from the South of Bulgaria.

Nerbonne and Heeringa (1998) were the first to project the results of MDS on a map by extracting the first 3 MDS dimension and associating each dimension with a color (red, blue and green). Each village in a data set was represented as a mix of these 3 colors depending on its coordinates in the MDS analysis. The space between the sites was colored by interpolation. The results of this technique applied on a Buldialect data set is shown in Figure 4.

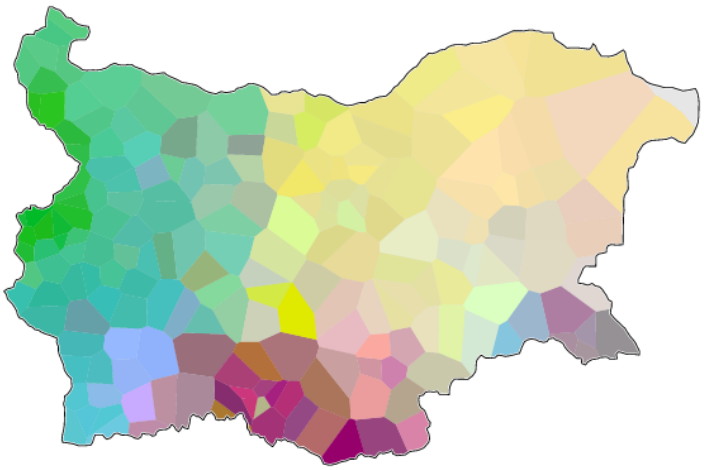

Figure 4: First 3 MDS dimensions projected on a map of Bulgaria.

This visualization technique enables us to detect three main dialect groups and at the same time to portray the degree of their linguistic heterogeneity (spread). It is especially suitable for the data that forms a continuum, like dialect data, rather than clearly separated groups.

5 Cluster determinants

Most of work done in dialectometry so far has been focused on the first two steps: calculation of distances on the aggregate level and detection of dialect groups by means of some of the described methods. Settling on an appropriate linguistic interpretation of an aggre-

gate analysis has always been considered the main drawback of dialectometry and made it less popular among more traditionally oriented dialectologists, who are often not as interested in the aggregate relations among sites, as in the concrete linguistic features that make one dialect distinct from other areas. Previous work in this direction include Nerbonne (2006), Grieve (2009) and Wieling and Nerbonne (2011). In this paper we present a method recently developed by Prokić, Çöltekin and Nerbonne (2012) that proceeds from a group of sites and identifies characteristic features of candidate dialect areas.

5.1 Method

The method proposed by Prokić et al.(2012) is general in that it can be applied to both numerical and to categorical data, requiring only that there be a numerical measure of difference defined for the data. It starts with data where the sites have already been split into groups, and it does not require any information on how the groups were obtained. This makes this method very general and also easily applicable in dialectometry.

The method proposed seeks the features which differ very little within the group in question and a great deal outside that group. It examines one candidate group g at a time that consist of $|g|$ sites among a larger area of interest G consisting of $|G|$ sites. This larger

area *G* includes the sites *s* within the cluster of interest and also those outside the cluster of interest *g*. The method assumes a measure of difference *d* between sites, always with respect to a given feature *f*. A mean difference with respect to *f* is calculated within the group in question

$$\bar{d}_f^g = \frac{2}{|g|^2 - |g|} \sum_{s,s' \in g} d_f(s,s')$$

and also involving elements outside the group in question

$$\bar{d}_f^{\not g} = \frac{1}{|g|(|G| - |g|)} \sum_{s \in g, s' \notin g} d_f(s,s')$$

Characteristic features are those with relatively large differences between $\bar{d}_f^g$ and $\bar{d}_f^{\not g}$. The values obtained are sensitive to the size of the group under examination and the number of elements compared, which can be affected by missing data. Most importantly, the feature differences may systematically be influenced by differences in the natural variability of the data. For example, it appears that vowels are naturally more variable than consonants. To abstract away from this last influence, both $\bar{d}_f^g$ and $\bar{d}_f^{\not g}$ are standardized by calculating the

difference between *z-scores*. The mean and standard deviation of the difference values are estimated from all distance values calculated with respect to feature *f*. As a result the following measure is used:

$$\frac{\bar{d}_f^{\,\$} - \bar{d}_f}{sd(d_f)} - \frac{\bar{d}_f^{\,g} - \bar{d}_f}{sd(d_f)}$$

where d_f represents all distance values with respect to *f*. The scores are normalized for each feature separately.

The Buldialect set is blessedly complete, with data missing for very few pronunciations at very few sites. This means that we need not ask ourselves how often a given feature must be instantiated in a given region before we are willing to ask whether it might be characteristic. Prokić et al.(2012) discuss this problem.

5.2 Experimental setup

The method described is tested on the Buldialect pronunciation data (Section 2). Pronunciations of the 152 words from this data set were multi-aligned using ALPHAMALIG algorithm. The automatically obtained alignments were very accurate, with scores ranging between 93 and 97 per cent depending on the evaluation method. We manually post-processed the alignments since we are primarily interested in

the performance of the 'cluster determinants' method. However, because of the good quality of the automatically generated alignments the post-processing step could be avoided in future research. By proceeding from the multi-aligned data we assure that every position within a word is treated as a separate feature *f*. This is, incidentally, the point at which the present paper extends the work in Prokić et al. (2012).

The distances between each two sites were calculated by comparing the phones in each position in the multi-aligned pronunciations and taking the average of all obtained distances. The phones were compared based on the following weighting scheme: same phones have distance 0, same phones with different diacritics have distance 0.5 and different phones have distance 1. We use Gabmap software to do all calculations and obtain a *site* × *site* distance matrix.

In order to determine the optimal number of dialect groups in the data, we analyzed the distance matrix by means of MDS and Neighbor-Net (shown in Figures 3 and 4) which both revealed 3 relatively distinct groups. We tested several hierarchical clustering algorithms on our data by coloring the points in the MDS plot according to the 3-way divisions suggested by each of the clustering algorithms. In Figure 5 we present the 3-way division detected by Ward's algorithm, which we have found to be optimal for this data. The dialect

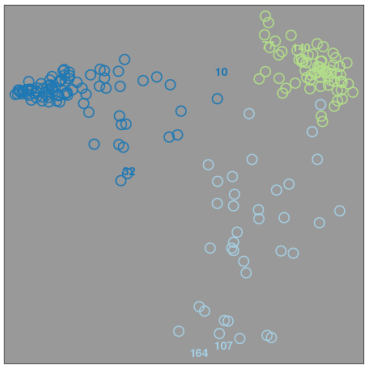

Figure 5: 3-way division derived by Ward's clustering method projected on MDS plot. Note that the colors correspond with those in the map in Fig.1.

regions detected can be seen on the map in Figure 1. These results conform with the traditional Bulgarian dialectology (Stoykov, 2004) and quantitative analyses of the Buldialect data set (Houtzagers et al., 2010), which both distinguish western, eastern and southern dialects as the most important dialect groups.

In the final step, we apply the cluster determinants method described in section 5.1 in order to determine the linguistic bases of this 3-way division. Since our input data is multi-aligned, the features *f* that we are trying to recover using the described method are single positions in words. This is a step beyond Prokić, Çöltekin and Nerbonne (2012), which sought *words* characteristic of a given region. We now apply the same technique to seek characteristic *sounds*. In the next section we present the results.

5.3 Results

For each of the three dialect groups we calculated the most important linguistic feature, i.e. cluster determinants. In Table 1 we present the top five determinants for the western dialects. Each feature, i.e. word position, is presented within the word it occurs and marked in bold. We also present the standard pronunciation of the word in question.

Table 1: The five most important determinants for western dialects. We also give the word in which the feature appears and mark the feature itself with bold font.

Determinants	In cluster	Outside cluster
m lʲ a k **o** t o	o	u ʊ
bʲ a x m e	b	bʲ
dʊ **o** b	e	o i u a ʊ ɤ ɪ ʌ
nʲ a m a	n	nʲ
d ɤ n **o**	o	u ʊ

In Figure 6 we present the distribution of the first phone [o] from word *млякото* /mlʲakoto/ ‘milk’ (colored blue on map) that was the highest scoring feature for the western dialects:

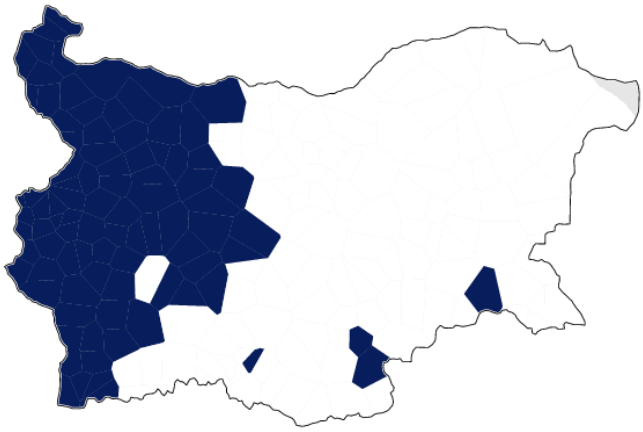

Figure 6: Distribution of the phone [o] in the second syllable of /mlʲakoto/, which clearly separates the western areas from the rest of the country.

The map in Figure 6 clearly shows that in the word *млякото* /mlʲakoto/ ‘milk’, the first /o/ (the vowel in the second syllable) is always realized as [o] in the west and as [u] or [ʊ] in the rest of the country. For that reason the distances among the varieties in the west with respect to this feature are very small when compared to the distances between those same varieties and varieties elsewhere.

In Table 2 we present five most important determinants for the eastern dialects:

Table 2: The five most important determinants for eastern dialects. The second one represents elision of [j].

Determinants	In cluster	Outside cluster
ts^{j} a l	ts^{j}	ts
- a z	-	j
g r o z d **e**	i	e ə ɤ ɪ
e d n o	i	e ə ɤ ɪ ɑ
d ɤn **o**	o	u ʊ

In Figure 7 we show the distributional map for the most important cluster determinant for eastern dialects, realization of /ts^{j}/ in word *цял* /ts^{j}al/ ‘whole’. In the east, it is always realized as [ts^{j}] and in the west and south as [ts].

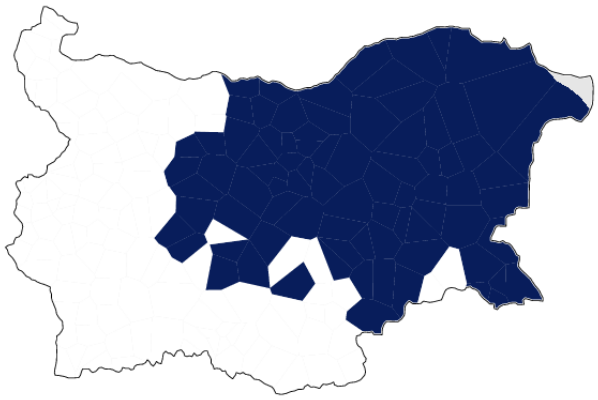

Figure 7: The segment /ts^{j}/ is realized as [ts^{j}] in the east and as [ts] in the west and south.

The list of five most important determinants for the southern dialects is given in Table 3:

Table 3: The five most important determinants for southern dialects.

Determinants	In cluster	Outside cluster
tʃ e t **ɤ**	ə	a e i ɑ ɤ
r ʊ ts **e**	i ɨ	e ɑ ə ɛ
t **o** v a	-	i o u ɑ ə ɤ ʊ
d e r **a**	ə	a e ɑ ɤ ʊ
r ɤ **ts** e	c c^{j} k	ts ts^{j}

In Figure 8 we present the distribution of the most important determinant for the southern dialects. In word *чема* /tʃetɤ/ 'read – 1st sg', the segment /ɤ/ has realization [ə] in the south and [a], [e], [i], [ɑ], and [ɤ] in the rest of the country.

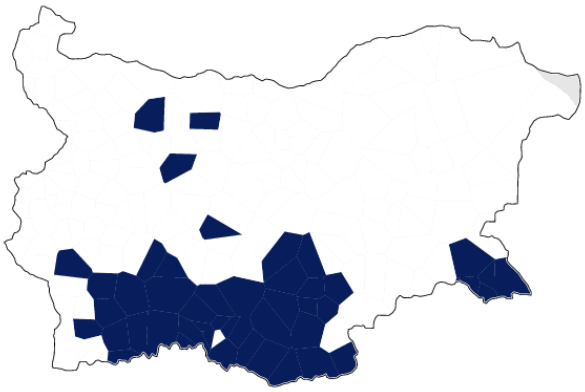

Figure 8: Segment /ɤ/ in word *чема* /tʃetɤ/ 'read – 1st sg' is realized as [ə] in the south.

The results presented suggest that this method is successful in recovering the most distinctive features for the area in question. In this paper, we have used multi-aligned data as input and treated each position in a word as a feature. However, this method can have any type of data as input, as long as the distances can be quantified. For example, Prokić et al. (2012) use whole words as features and quantify the distances between them using Levenshtein algorithm.

5.4 Conclusions

In this paper we have presented number of techniques taken from biology that are now standard tools in dialectometry, which is primarily concerned with measuring the distances between dialect varieties and their classification. Although biological and linguistic data differ to a great extent, techniques taken from biology have proven valuable in language data analyses. They have enabled us to analyze large amounts of data and overcome some of the methodological problems in earlier dialectology, which focused on identifying distinguishing individual features. Advances in the field of dialectometry have not been generally accepted by traditional dialectologists, perhaps because aggregate dialectometric analyses offer too little insight into the details they have focused on. In this work we have tested a new method that can overcome this problem, analyzing large amounts of data while at the same time preserving and sharpening a view on the linguistic details. This method can also be applied in other branches of linguistics that deal with quantitative language comparison. Clearly a great deal remains for future work. The technique should be applied to more data sets to gather more insight into its strengths and weaknesses, exposing further how it works and how it might be improved. One example of a point where a wider range of data must be examined is the parameter specifying how often a feature must be

instantiated in a given region if it is to qualify at all as being "characteristic".

The present paper has examined specific positions (sounds) in specific words in an effort to find characteristic elements (the vowel in the second syllable of *млякото* /mlʲakoto/ 'milk') for a given cluster, while at the same keeps track of the context in which the element occurs. A great deal of linguistic interest is attached to the question of regular segment correspondences with respect to generally characterized contexts (the /o/:/u/ correspondence in unstressed syllables) and we hope that the present paper has taken a step in that direction.

Finally, we should prefer to evaluate the work with respect to some independent criterion, perhaps the reactions of dialect speakers (positive or negative) to given correspondences, or perhaps to their characterizations of the one or the other variant as like their own variety, or as rather different.

References

Bhargava, Aditya and Grzegorz Kondrak (2009) Multiple Word Alignment with Profile Hidden Markov Models. *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics, Companion Volume: Student Research Workshop and Doctoral Con-*

sortium (NAACL-HLT 2009), Boulder. 43-48.

Black, Paul (1973) 'Multidimensional scaling applied to linguistic relationships', in *Cahiers de l'Institut de Linguistique Louvain*, Volume 3 (Montreal). 13-92. Expanded version of a paper presented at the Conference on Lexicostatistics. University of Montreal.

Bolognesi, Roberto and Wilbert Heeringa (2002) De invloed van dominante talen op het lexicon en de fonologie van Sardische dialecten. In: *Gramma/TTT: tijdschrift voor taalwetenschap*, 9(1): 45—84.

Bryant, David and Vincent Moulton (2004) Neighbornet: an agglomerative algorithm for the construction of plenar phylogenetic networks. *Molecular Biology and Evolution*, 21:255-265.

Campbell, Lyle (2004) *Historical Linguistics: An Introduction*. Edinburgh University Press, second edition.

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

Embleton, Sheila (1993) Multidimensional Scaling as a Dialectometrical Technique: Outline of a Research Project. In Reinhard Köhler &

Burghard Rieger (eds.) *Contributions to Quantitative Linguistics,* Dordrecht: Kluwer. 267_276.

Felsenstein, Joseph (2004) *Inferring Phylogenies*, Sinauer Associates, Inc.

Goebl, Hans (1982) Ansätze zu einer computativen Dialektometrie. In: Werner Besch, Ulrich Knoop, Wolfgang Putschke en Herbert E. Wiegand (eds.) *Ein Handbuch zur deutschen und allgemeinen Dialektforschung*. Handbücher zur Sprach- und Kommunikationswissenschaft, Volume I, 778-792. Berlin and New York: de Gruyter Mouton.

Goebl, Hans (1983) "Stammbaum" und "Welle". Vergleichende Betrachtungen aus numerisch-taxonomischer Sicht. In: *Zeitschrift für Sprachwissenschaft* 2, 3-44.

Grieve Jack (2009) *A Corpus-Based Regional Dialect Survey of Grammatical Variation in Written Standard American English*. Ph.D. Dissertation. Northern Arizona University.

Gusfield, Dan (1997) *Algorithms on Strings, Trees and Sequences: Computer Science and Computational Biology*. Cambridge Universi-

ty Press.

Heeringa, Wilbert and Charlotte Gooskens (2003) Norwegian dialects examined perceptually and acoustically. In: John Nerbonne and William Kretzschmar (eds.), *Computers and the Humanities*, Kluwer Academic Publishers, Dordrecht, 37 (3), 293-315.

Heeringa, Wilbert (2004) Measuring dialect pronunciation differences using Levenshtein distance. Ph.D. dissertation University of Groningen.

Hoppenbrouwers, Cor and Geer Hoppenbrouwers (2001) *De indeling van de Nederlands streektalen: dialecten van 156 steden en dorpen geklasseerd volgens de FFM*. Assen: Koninklijke Van Gorcum.

Houtzagers, Peter, John Nerbonne and Jelena Prokić (2010) Quantitative and Traditional Classifications of Bulgarian Dialects Compared. *Scando-Slavica* 56 (2), pages 29-54.

Huson, Daniel H. and David Bryant (2006) Application of phylogenetic networks in evolutionary studies, *Molecular Biology and Evolution*, 23(2):254-267.

Jain, Anil K. and Richard C. Dubes (1988) *Algorithms for Clustering Data* (New Jersey).

Kessler, Brett (1995) Computational dialectology in Irish Gaelic. In: *Proceedings of the European ACL*, 60-67. Dublin: Association for Computational Linguistics.

Legendre, Pierre and Louis Legendre (1998) *Numerical Ecology*, second ed. (Amsterdam).

Levenshtein, Vladimir I. (1966) Binary codes capable of correcting insertions, deletions and reversals. *Cybernetics and Control Theory*, 10(8):707– 710. Russian orig. in *Doklady Akademii Nauk SSR* 163(4), 845–848, 1965.

List, Johann-Mattis (2012) Multiple sequence alignment in historical linguistics. A sound class based approach. *Proceedings of ConSOLE XIX*.

Nerbonne, John (2006) Identifying Linguistic Structure in Aggregate Comparison. *Literary and Linguistic Computing* 21(4). 463-476. (J.Nerbonne & W.Kretzschmar, Jr. (eds.) *Progress in Dialectometry: Toward Explanation*)

Nerbonne, John, Wilbert Heeringa, Eric van den Hout, Peter van de Kooi, Simone Otten and Willem van de Vis (1996) Phonetic Distances between Dutch Dialects. In: G.Durieux, W.Daelemans, & S.Gillis (eds.) *CLIN VI: Proc. of the Sixth CLIN Meeting.* Antwerp, Centre for Dutch Language and Speech (UIA), 185-202.

Nerbonne, John and Wilbert Heeringa (1998) Computationele vergelijking and classificatie van dialecten. In: *Taal en Tongval; Tijdschrift voor Dialectologie* 50(2): 164-193.

Nerbonne, John und Christine Siedle (2005) Dialektklassifikation auf der Grundlage Aggregiert Ausspracheunterschiede. *Zeitschrift für Dialektologie und Linguistik* 72(2), 129-147.

Nerbonne, John, Peter Kleiweg, Wilbert Heeringa and Franz Manni (2008) Projecting Dialect Differences to Geography: Bootstrap Clustering vs. Noisy Clustering. In: Christine Preisach, Lars Schmidt-Thieme, Hans Burkhardt & Reinhold Decker (eds.) *Data Analysis, Machine Learning, and Applications. Proc. of the 31st Annual Meeting of the German Classification Society* Berlin: Springer. 647-654. (*Studies in Classification, Data Analysis, and Knowledge Organization*).

Nerbonne, John, Rinke Colen, Charlotte Gooskens, Peter Kleiweg and Therese Leinonen (2011) Gabmap — A Web Application for Dialectology. *Dialectologia*. Special Issue II, 2011: 65-89.

Osenova, Petya, Wilbert Heeringa and John Nerbonne (2009) A Quantitative Analysis of Bulgarian Dialect Pronunciation. *Zeitschrift für slavische Philologie* 66(2), 425-458.

Prokić, Jelena, Martijn Wieling and John Nerbonne (2009) Multiple string alignments in linguistics. In: Lars Borin & Piroska Landvai (chairs) Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education (LaTeCH - SHELT&R 2009) EACL Workshop.

Prokić, Jelena John Nerbonne, Vladimir Zhobov, Petya Osenova, Kiril Simov, Thomas Zastrow and Erhard Hinrichs. The computational analysis of Bulgarian dialect pronunciation. *Serdica Journal of Computing*, Sofia, 2009.

Prokić, Jelena (2010) *Families and Resemblances*. PhD thesis. University of Groningen.

Jelena Prokić, Cagri Coltekin and John Nerbonne. Detecting Shibboleths. In: *Proceedings of the EACL 2012 Joint Workshop of LINGVIS & UNCLH*. Avignon. 2012.

Schmidt Johannes (1872) *Die Verwandtschaftsverhältnisse der Indogermanischen Sprachen*, Weimar.

Steiner, Lydia, Peter Stadler & Michael Cysouw (2011) A Pipeline for Computational Historical Linguistics. *Language Dynamics and Change* 1(1).

Stoykov, Stoyko (2004) *Bulgarska dialektologiya*. Sofia, 4th ed.

Wieling, Martijn, Eliza Margarethe and John Nerbonne (2012) Inducing a Measure of Phonetic Similarity from Pronunciation Variation. *Journal of Phonetics* 40(2).307-314.

DOI: http://dx.doi.org/10.1016/j.wocn.2011.12.004

Wieling, Martijn and John Nerbonne (2011) Bipartite spectral graph partitioning for clustering dialect varieties and detecting their linguis-

tic features. *Computer Speech and Language* 25, 700-715. DOI:10.1016/j.csl.2010.05.004.

Wieling, Martijn, Jelena Prokić and John Nerbonne (2009) Evaluating the pairwise string alignment of pronunciations. In Lars Borin & Piroska Landvai (chairs) *Language Technology and Resources for Cultural Heritage, Social Sciences, Humanities, and Education* (LaTeCH - SHELT&R 2009) EACL Workshop. 26-34.