A Quantitative Analysis of Bulgarian Dialect Pronunciation*

Petya Osenova†, Wilbert Heeringa and John Nerbonne
Humanities Computing, University of Groningen
petya@bultreebank.org, w.j.heeringa@rug.nl, j.nerbonne@rug.nl

20th February 2006

Abstract

We apply a computational measure of pronunciation difference to a database of 36 word pronunciations from 490 sites throughout Stoykov's Bulgarian Dialect Atlases. The result is a comprehensive view of the aggregate pronunciation differences among the 490 sites. This study aims to contribute therefore to Bulgarian dialectology, as well as to the development and testing of the computational technique and its implementation in a software package.

1 Introduction

In recent years computational techniques enable the incorporation of large amounts of dialectal material into studies of language variation. Phonetic measurements of Germanic and Romance dialects (Dutch, Norwegian, Sardinian, German, American) have been successfully carried out, using Levenshtein distance, also known as '(string) edit distance', as a basis. This line of work began in 1995 when Kessler introduced the use of the Levenshtein distance as a tool for measuring linguistic distances among the pronunciations of language varieties. Levenshtein distance is a string edit distance measure, and Kessler applied this algorithm to the comparison of Irish dialects. Later the same technique was successfully applied to Dutch (Nerbonne at. al. 1996; Heeringa 2004, pp. 213–278), Sardinian (Bolognesi and Heeringa, 2002), Norwegian (Gooskens and Heeringa, 2004), German (Nerbonne and Siedle, 2005), and American English (Nerbonne, to appear). Other analyses are underway.

These studies have applied Levenshtein distance to large numbers of words as they are pronounced in many different data collection sites. They have, by and large, vindicated traditional dialectological divisions, even while systematizing the technique of comparison, enormously enlarging the base of data on which

*This work is funded by NWO, Project Number 048.021.2003.009, P. I. J. Nerbonne, Groningen, and also a grant from the Volkswagenstiftung "Measuring Linguistic Unity and Diversity in Europe", P.I. E. Hinrichs, Tübingen.

†Also at the Bulgarian Language Division, Faculty of Slavonic languages, St. Kl. Ohridski University, Sofia, Bulgaria

analyses are systematically based, and providing novel means of understanding the dialect landscapes to which they have been applied.

It is a challenge to see how well the methods developed primarily for other language families perform for Bulgarian, a Slavic language. The task was challenging: first, because there was no digitized data for Bulgarian dialects at our disposal; second, because Bulgarian dialects had not been processed earlier with computational tools, and third, due to the fact that some language specific features had to be taken into account for the first time.

The structure of the paper is as follows: the next section describes the traditional divisions of Bulgarian dialects, providing background. Section 3 focuses on the data source and the preparation of the data. In Section 4 the Levenshtein distance measurement is presented. Section 5 discusses the analytical procedures applied to the distances computed in Section 4, namely clustering and multidimensional scaling. Section 6 focuses on some of the dialect groups in more detail. Section 7 presents the relation of the dialects to standard Bulgarian pronunciation. Section 8 comments on the results obtained on an extended set of data, and Section 9 offers conclusions.

2 Bulgarian Dialect Scholarship

Scholarship has offered several ideas about the geographical distribution of Bulgarian dialects, mostly relying on phonetic criteria. One prominent traditional division follows the pronunciation of the old Bulgarian vowel 'yat'. It divides the Bulgarian dialects into Western, where 'yat' has only the reflection 'e', for example *bel* 'white'- *beli* 'white-pl' and Eastern, where 'yat' has both reflections, 'e' and 'ya'(for example *bʲal* 'white'-*beli* 'white-pl'. This one characteristic is not by itself enough for consistent generalization, but the distinction remains one of the most important features for the comparison of the dialects.

Another phonetics-based classification of Bulgarian dialects reflects the realizations of the old Bulgarian 'big nosovka' (голяма носовка), a nasal vowel. It divides Bulgarian dialects into five groups: ə-dialects (Northeastern and Northwestern Bulgaria and the eastern part of Southeastern Bulgaria); a-dialects (Western Bulgaria and the eastern dialect of Pirdop); ɒ-dialects (the Rodopi mountain); æ-dialects (the Teteven region and two villages in Eastern Bulgaria, Kazichino and Golitsa); and u-dialects (Western Bulgarian areas near the Bulgarian-Serbian border). This classification is admirably simple but encounters exceptions, and which mean that it cannot divide the Bulgarian dialects in a satisfactory way.

There have been other attempts at the phonetic classification of Bulgarian dialects as well, but we are not going to present all of them here. The interested reader is referred to Stoykov (2002, pp. 88–90).

According to most morphological and lexical research Bulgaria is divided into a central part (Northeastern and Central Bulgaria) and a peripheral part (Northwestern, Southwestern and Southeastern Bulgaria).

Because of the instability and conflicting nature of various linguistic criteria Stoykov (2002) suggests a classification of Bulgarian dialects which respects

geographical continuity, as well. In his standard work he distinguishes six, rather than five areas, concluding that:

1. Bulgarian dialects are not separated categorically, but they rather form a continuum.

2. Within Bulgarian dialects there is a central (typical) area and a peripheral (transition) area. Similar situations have been observed for other languages as well, e.g. Dutch (Hoppenbrouwers and Hoppenbrouwers 2001, pp. 66–67).

3. The most striking distinction of Bulgarian dialects is the distinction between Eastern and Western ones.

In Figure 1 the six most significant geographical groups of Bulgarian dialects are shown as presented in Stoykov (2002, p. 416).

3 The Data

We consider in turn the sources of our data and the selection we made, its preparation, and its conversion to digital form.

3.1 Sources

The data was digitized from the four volumes of Bulgarian dialect atlases which cover the entire country area: Volume I - Southeastern Bulgaria (Stoykov and Bernshteyn 1964), Volume II - Northeastern Bulgaria (Stoykov 1966), Volume III - Southwestern Bulgaria (Stoykov et. al. 1975) and Volume IV - Northwestern Bulgaria (Stoykov et. al. 1981). The atlas materials appear to have been gathered with an eye toward the identification of the historical roots of Bulgarian. This would explain why the data was gathered only from villages with exclusively Bulgarian populations regardless of geography (this is unlike similar atlases for other languages we have worked with). This also means that the sites are not distributed uniformly within Bulgaria. For example, because most of the "purer" Bulgarian sites are in the mountains, there are more mountainous sites than one would expect.

The atlases consist of two parts: maps and commentary on the maps. The maps present general information, while the commentary focuses on deviations and more elaborate characterizations of pronunciation etc. The data was collected in some ways that promise an especially faithful characterization of language variation: the researchers that collected the data did not rely on only one informant, but instead used several, divided into main informants and additional ones. The researchers' approach to informants was not to ask direct questions, but rather to conduct extensive interviews from which material was selected.

The data reflects different linguistic phenomena, and the phenomena are often instantiated by more than one word. Thus material does not coincide across all atlases, and where it does coincide, it may not be presented as instantiating the same linguistic phenomenon. This required us to study the material closely.

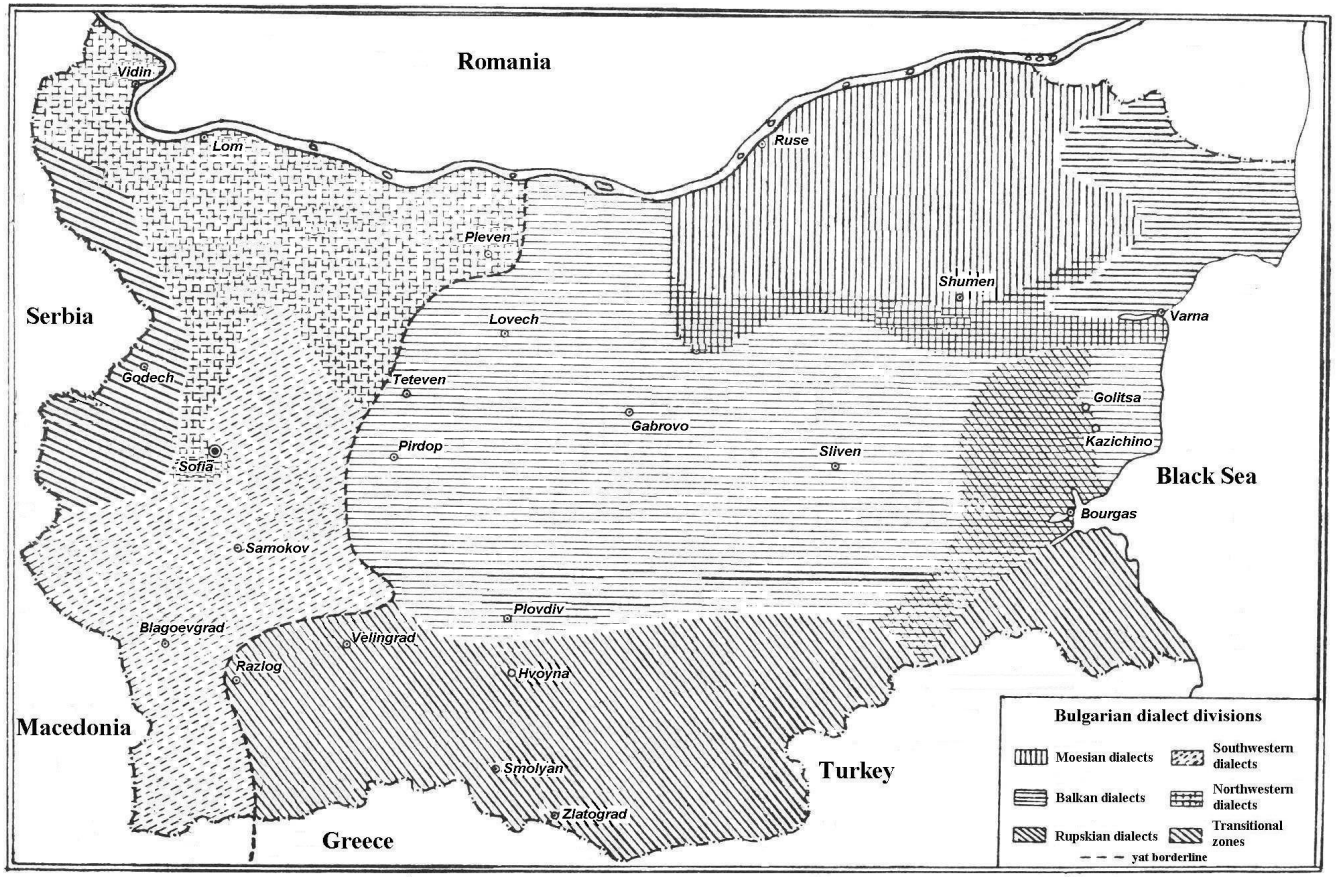

Figure 1: The map of Bulgarian dialect divisions as presented in Stoykov (2002, p. 416). The vertical lines represent Moesian dialects, the horizontal lines represent Balkan dialects, the broken slanting lines represent Southwestern dialects, the crossed lines represent Northwestern dialects. The thick broken line represents the 'yat' borderline that divides the dialects into two major groups: Western and Eastern. The nearly horizontal slanting lines on the left side show the transitional zones, and steeply slanting lines at the bottom of the map represent the Rupskian dialects.

The following information is presented in three different sorts of maps. The first sort of information maps a single phonetic or grammatical phenomenon which is not connected to specific lexical material. The second sort characterizes a certain phonetic or grammatical phenomenon within a restricted set of words, and the third and final sort presents a certain phonetic or grammatical phenomenon exemplified by only one word. The basic principle in the presentation of the information is the default to 'accept more general information unless otherwise stated'. Consequently, we sometimes interpreted inexplicitness as confirmation that nothing special needed to be said.

In the present paper we extract words from these atlases which we then compare in pronunciation. Our method (described below) relies on transcriptions of entire words, which we took from the atlas as best we could (see discussion above). Where we needed to extrapolate, we always did this conservatively, e.g. using little phonetic detail.

3.2 Preparation of the data

3.2.1 The maps

For the present experiment we sought the transcribed pronunciations of a common set of words for all Bulgarian areas. In view of this, we proceeded in the following manner: First, we chose 490 dialect sites within Bulgaria with as much geographic distribution as possible. In addition we included the pronunciation Standard Bulgarian.[1] The sites were selected with respect to two main criteria: maximally complete coverage of the area covered by the atlas, and a representative number of varieties and sub-varieties. There are altogether 1682 sites in the atlas, so that our selection of 490 constitutes roughly one third of all the sites. See the map in Figure 2.[2]

Second, we digitized a set of 54 words, which turned out not to be instantiated in every sites, but which includes a subset of 36 words that were instantiated in all the atlases. This differentiation of two sets arose because, as noted above, the lexical material differs a great deal across the four atlases. But we decide also to view the 36-word and 54-word sets as a test of the consistency of the measurements. So while we concentrate here on reporting the results of the experiment with a common set of 36 words, we recognize that some might consider this to be too small a sample (we would disagree). We shall therefore likewise report on the experiments with 54 words, and in particular on the degree to which the two sets of measurements correlate. Based on our experience using pronunciation distance measures in other areas, we expect a high correlation between the two sets of measurements.

The words selected represent various parts of speech (nouns, adverbs, adjectives, verbs, numerals, participles) and different word forms (singular and plural, 3rd person verb forms, past tense forms, etc.). The subset of 36 words which are common to all sites is as follows:

[1] The standard pronunciations are in accordance with Popov et.al. (1998).

[2] Note that in all maps presented here, the boundary lines indicate administrative divisions, not dialect areas.

Figure 2: The distribution of the Bulgarian sites selected.

бъчва ˈbətʃva, ‘barrel’	зълва ˈzəlva, ‘sister-in-law’
дошъл doˈʃəl, ‘has come-he’	жълт ʒəlt, ‘yellow’
зъб zəb, ‘tooth’	събота ˈsəbota, ‘Saturday’
къща ˈkəʃta, ‘house’	бяла ˈbʲala, ‘white’- fem
бели ˈbɛli, ‘white’- pl	язди ˈjazdi, ‘ride’-3per
неделя nɛˈdɛlʲa, ‘Sunday’	млекар mlɛˈkar, ‘milkman’
грешка ˈgrɛʃka, ‘mistake’	венчило vɛnˈtʃilo, ‘married life’
ключ klʲutʃ, ‘key’	чаша ˈtʃaʃa, ‘glass; cup’
път pət, ‘road’	жаби ˈʒabi, ‘frogs’
нощви ˈnoʃtvi, ‘hutch’	поляна poˈlʲana, ‘glade’
овче ˈovtʃe, ‘sheep’s’	тънко ˈtənko, ‘narrow-neut’
гуляй guˈlʲaj, ‘feast’	овчар ovˈtʃar, ‘shepherd’
кон kon, ‘horse’	сън sən, ‘dream’
отишъл otiˈʃəl, ‘has gone-he’	вътре ˈvətrɛ, ‘inside’
тенджера ˈtɛndʒɛra, ‘pot’	джоб dʒob, ‘pocket’
няма ˈnʲama, ‘there is no’	череша tʃɛˈrɛʃa, ‘cherry’
гръб grəb, ‘back’	живя ʒiˈvʲa, ‘lived’
сол sol, ‘salt’	ден dɛn, ‘day’

These words represent many of the most important phonetic features of Bulgarian. They reflect the following phenomena:

1. the reflections of ‘yat’ in different phonetic contexts (stressed and un-

stressed, word-finally, after fricatives, etc.): ˈbʲala, ˈbɛli, ˈgrɛʃka, mlɛˈkar, ˈvətrɛ, vɛnˈtʃilo

2. the reflections of the etymological 'ja': ˈjazdi or ˈjezdi , poˈlʲana or poˈlena, guˈlʲaj or guˈlej

3. palatal-nonpalatal-semipalatal distinction word-finally: sol or solʲ, pət or pətʲ, kon or konʲ, dɛn or denʲ

4. the realizations of 'schwa' under stress: ˈbətʃva or ˈbotʃva or ˈbatʃva etc. The same for the other words: ˈzəlva, sən, ˈtənko, otiˈʃəl, doˈʃəl

5. the realizations of the nasal vowel: zəb or zob or zab etc. Similarly for other words: ˈkəʃta, ˈsəbota

6. the metatheses 'əl-lə' and 'ər-rə': grəb or gərb, ʒəlt or ʒlət.

7. the realizations of various vowels in different contexts: ˈovtʃe or ˈovtʃo, klʲutʃ or klitʃ

8. the reduction of the open vowels in unstressed position: mlɛˈkar or mliˈkar, vɛnˈtʃilo or vinˈtʃilo etc.

The full list of 54 words includes the 36 words above plus the following 18 words, which were not common to all sites. More precisely, they were available only for the sites in the first and the second atlases. The phonetic information that this set adds is the behavior of the fricative 'x' in some of the words. For example, the presence or absence of 'x' is indicated in ˈxladno and its alternative ˈladno:

шепа ˈʃɛpa, 'handful'	две dve, 'two'
ясла ˈjasla, 'manger'	шапка ˈʃapka, 'hat'
жив ʒiv, 'alive'	почивам poˈtʃivam, 'rest-I'
широк ʃiˈrok, 'wide'	зет zɛt, 'son-in-law'
език ɛˈzik, 'tongue'	добър doˈbər, 'good'
отивам oˈtivam, 'go-I'	мързелив mərzeˈliv, 'lazy'
ябълка ˈjabəlka, 'apple'	хайде ˈxajdɛ, 'let's'
хладно ˈxladno, 'cold'	снаха snaˈha, 'daughter-in-law'
беряха bɛˈrʲaxa, 'were picking up-they'	дадох ˈdadox, 'gave-I'

3.2.2 Digitization

When we started the task, the data was available only in printed form. Since we wished to analyze the data computationally, digitization became a very important subtask. The data was converted to the X-SAMPA (Wells) representation of IPA (the alphabet of the International Phonetic Association, see Handbook IPA 2003), because X-SAMPA is used within the Levenshtein-based toolkit we used (`www.let.rug.nl/kleiweg/L04/`). X-SAMPA encodes IPA, but is ASCII-based, and therefore easily processed by a virtually all software on all platforms.

We also provided a more permanent form in an XML format using the facilities of the CLaRK System (Simov et. al. 2004), and using Unicode and attempting to standardize geographic references. We hope to report on this separately.

4 Measuring Pronunciation Distance

4.1 Method

Levenshtein distance is a technique to compare a pair of strings (words) and to assay their distance from each other. Two dialects are compared by comparing the pronunciation of the same words in the two dialects and then averaging the distances of the pairs of words.

An effective way to understand Levenshtein distance is to consider how one pronunciation may be transformed into the other by means of inserting, deleting or substituting individual sounds (symbols). Weights are assigned to these three operations. In the simplest form of the algorithm, all operations have the same cost. We illustrate this with an example of two varieties of a word pronunciation in Northwestern dialects. *Черешa* ('cherry') is pronounced[3] as [ˈtsrɛʃnʲa] in some dialects in the most Northwestern part (near Kula, Belogradchik) and the transitional area (around Tryn), and as [t͡ʃɛˈrɛʃa] in other dialects such as those near Sofia. Changing one pronunciation into the other may proceed as follows (ignoring suprasegmentals and diacritics for this moment):

t͡srɛʃna	subst. t͡s by t͡ʃ	1
t͡ʃrɛʃna	insert ɛ	1
t͡ʃɛrɛʃna	delete n	1
t͡ʃɛrɛʃa		
		3

In fact many sequence operations map [t͡srɛʃna] to [t͡ʃɛrɛʃa]. The power of the Levenshtein algorithm is that it always finds the cost of the cheapest mapping. Levenshtein distance is then the distance assigned by the Levenshtein algorithm, the cost of the least expensive means of mapping one string to another.

Comparing pronunciations in this way, the distance between longer pronunciations will generally be greater than the distance between shorter pronunciations. The longer the word, the greater the chance for differences with respect to the corresponding word in another variety. In order not to overemphasize the importance of longer words, we normalize the raw Levenshtein distance using word length. Thus, the sum of the operations is divided by the length of the longest alignment which gives the minimum cost. The longest alignment has the greatest number of matches. In our example we have the following alignment:

t͡s	∅	r	ɛ	ʃ	n	a
t͡ʃ	ɛ	r	ɛ	ʃ	∅	a
1	1				1	

[3]The encodings in this paper are in IPA (Handbook of the International Phonetic Association 2003)

The total cost of 3 (1+1+1) is now divided by the length of 7. This gives a word distance of 0.43 or 43%.

The simplest versions of this method are based on a notion of phonetic distance in which phonetic overlap is binary: non-identical phones contribute to phonetic distance, identical ones do not. In this simplest version the pair [ɪ,ɒ] counts as different to the same degree as [ɪ,e]. In more sensitive versions gradual segment distances are used as weights. The segment distances are based on comparison of feature values or acoustic measurements. In fact, Heeringa (2004) shows that the phone-based methods outperform most of the methods which are using gradual segment distances as operation weights (see p. 186 and p. 194). We use this simple version of the Levenshtein algorithm in this paper.

4.2 Results

Given a sample of word pronunciations from two sites, we average the distances of all pairs of corresponding words to obtain an estimate of the aggregate pronunciation difference between any two sites. We repeat this for all $(490 * (490 - 1))/2 (= 119,805)$ pairs of sites. An important issue is whether the data sample is large enough for us to extract a reliable signal. As a measure of reliability we used Cronbach's α method (for details see Heeringa (2004, pp. 170–173)), for which a widely accepted threshold is 0.70. Our results show a value of 0.84 for the set of 36 words. We therefore view the data sample large enough to provide a reliable view of pronunciation differences. This is incidentally the justifcation for the remark above (§ 3.2.1) that we regard the 36-word sample as large enough.

Conceptually, our application of the pronunciation difference measurements results in a 490×490 table of pronunciation differences we have measured using the Levenshtein algorithm. Naturally, we may restrict our attention to half of the table, since the distances in it are symmetrical. Once we know the (pronunciation) distance from Plovdiv to Sofia, we know the distance in return from Sofia to Plovdiv.

In Figure 3 the connections between the dialects are shown based on the Levenshtein distances for 36 words. Darker lines mark pairs of sites which are more similar to each other. Even at this level, without any clustering or further analysis, the important division between Western and Eastern dialects emerges clearly in the map. It suggests that the Western dialect groups are closer and more coherent than the Eastern ones. It also shows some close varieties in the East: the southern Rodopian group, the Central Balkan group and the central part of the Moesian dialects. In addition, this map shows moderately strong connections between the Balkan and Moesian dialects, which appear here as merged.

There is little discussion in the literature as to whether the Western dialects are phonetically more cohesive than the Eastern ones as a whole. On the one hand, in West there have not been as many large migrations as in East, which would promote coherence. On the other hand, this area is supposed to be more coherent with respect to the 'yat' realizations (but less coherent with respect to developments of the Old Bulgarian nasal). The aggregate pronunciation

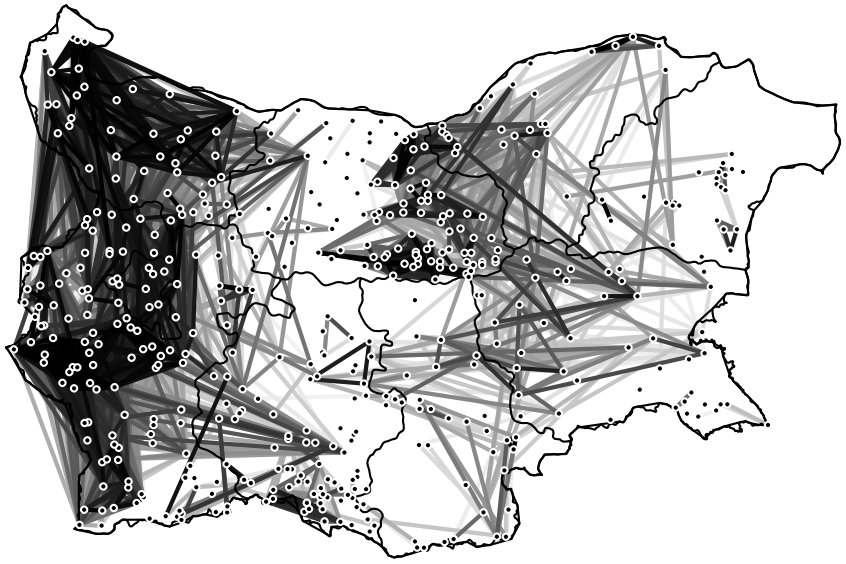

Figure 3: The average Levenshtein distances between 490 Bulgarian dialects are shown for 36 words. Darker lines indicate close varieties, while lighter lines indicate more remote ones.

differences highlighted by the average Levenshtein distance demonstrate this to be a dialectally cohesive area.

5 Analyzing Dialect Distances

From the line map, discussed above, we have already obtained some useful information about the dialects, but it is not sufficient for comparison to the results of traditional maps. For that reason, we continue our analysis by exploring the results further. On the one hand we will use CLUSTER ANALYSIS in order to divide the dialects into similarity classes. The resulting classification allows us to compare our findings with the results of dialectological scholarship, which has focused on the identification of dialect areas. We additionally examine the result of the pronunciation differences using a COMPOSITE CLUSTERING techniques, intended to mitigate chance effects in clustering.

We also analyze the pronunciation distance table using MULTIDIMENSIONAL SCALING, which complements clustering nicely and provides a view of the dialect landscape as a continuum. These techniques are presented in more detail below.

5.1 Cluster analysis

5.1.1 Method

Clustering is commonly applied in many disciplines as an exploratory technique in data analysis, one intended to expose natural classes of similar instances. The result of clustering is a DENDROGRAM, a tree incorporating all the input

elements, and in which more similar elements are grouped lower in the tree, i.e., closer to the input items, the leaves (the finest divisions). See Fig. 4.

Clustering is most easily understood procedurally. At each step of the procedure we select the shortest distance in the matrix of Levenshtein distances obtained above, and we fuse the two data points which gave rise to it. Since we wish to iterate the procedure, we have to assign a distance from the newly formed cluster to all remaining points. Although there are many so-called "matrix-updating algorithms," (Jain and Dubes, 1988) we may be content with simply averaging the lengths from the two points being fused to other elements being clustered[4]

We can measure the quality of a clustering result by comparing the distances in the dendrogram (e.g. the number of nodes which have to be traversed to move from one site to another in the dendrogram) to the distances in the pronunciation distance table via a COPHENETIC CORRELATION COEFFICIENT. This is simply the Pearson correlation coefficient of the two distances. The clustering technique we used obtained a cophenetic correlation coefficient (0.71). The dendogram obtained with this method explains $(0.71)^2 \times 100 = 50.4\%$ of the variance of the original Levenshtein distances.

A disadvantage of clustering is that it is statistically not stable: small differences in input values may lead to substantial differences in dendrograms. We use it nevertheless for its value in allowing us to compare our results to those of traditional scholarship.

5.1.2 Results

Here we present the results from the clustering in two views: in a dendrogram (Fig. 4) and in a classification map (Fig. 5). The numbers which indicate clusters in the dendrogram correspond to the numbered areas in the area map to facilitate the comparison.

As the distances among a large number of varieties (490) were calculated, only one general dendrogram is presented here. The dendrogram in Fig. 4 shows five clusters, which represent the same divisions as in the dialect area map (see Fig. 5 for comparison).

Classification Map

We likewise present a classification map, showing the varieties assigned to dialect groups. It represents the projection of the dendrogram in Fig. 4 onto Bulgarian geography. In particular we project the five-way division from the dendrogram in Fig. 4 in the map in Figure 5.

The map in Fig. 5 shows that the Rodopian group is not dialectally uniform. The 'yat' border to the north of Teteven is also not represented, and as a consequence, the Northwestern dialects appear to be rather indistinct from the Eastern ones. Although we need to exercise caution in interpreting the results of clustering, since real distinctions may not be reflected well, still we note one

[4]For those wanting technical detail, we note that we used the 'weighted pair group method using arithmetic averages' (UPGMA). See Jain and Dubes (1988).Nerbonne and Siedle (2005, p. 9) argue that this technique is less sensitive to irregularities in geographic sampling.

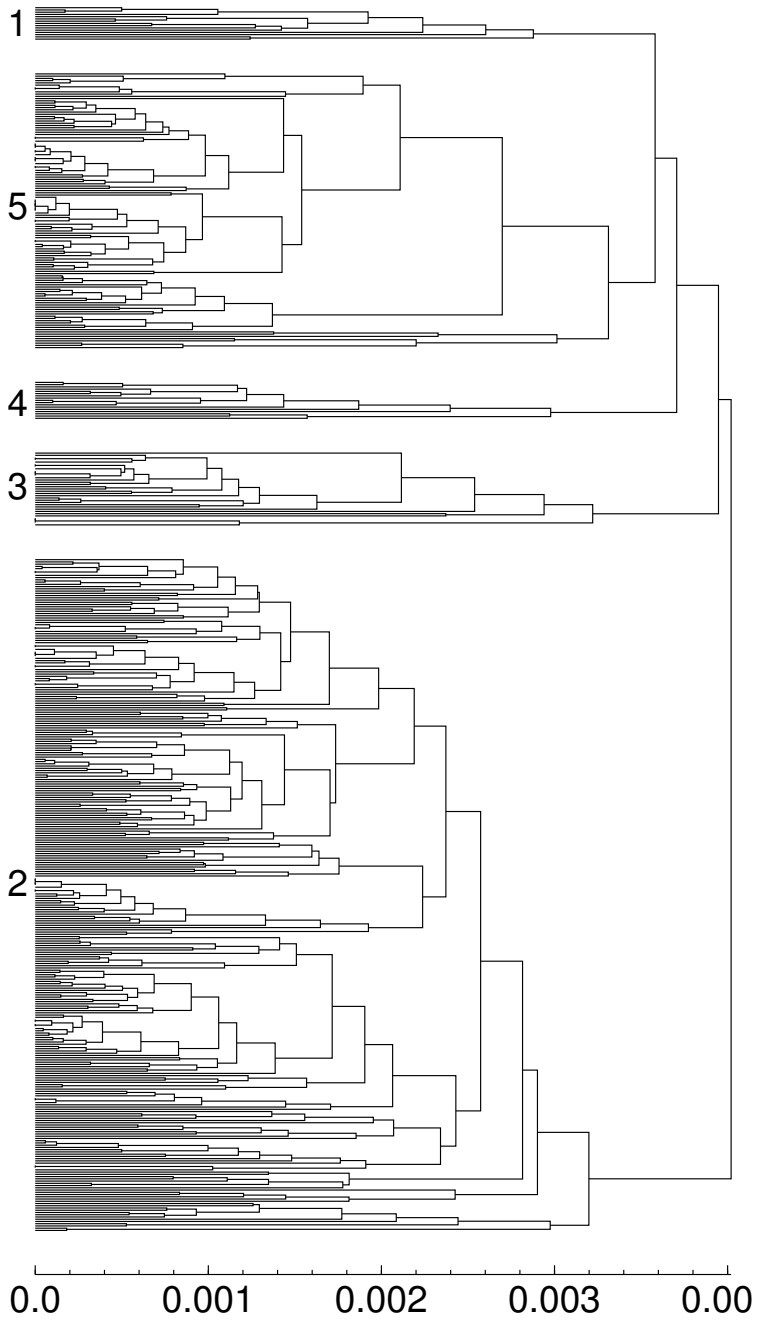

Figure 4: This dendrogram shows the five most significant clusters of Bulgarian dialects. The "leaves" on the left correspond to the individual sites in the sample, which are gradually fused into subclusters as one follows the diagram to the "root" on the right (WPGMA clustering was used, see text). The scale distance shows average Levenshtein distances as a fraction. The tree structure explains 50.4 % of the average Levenshtein distances.

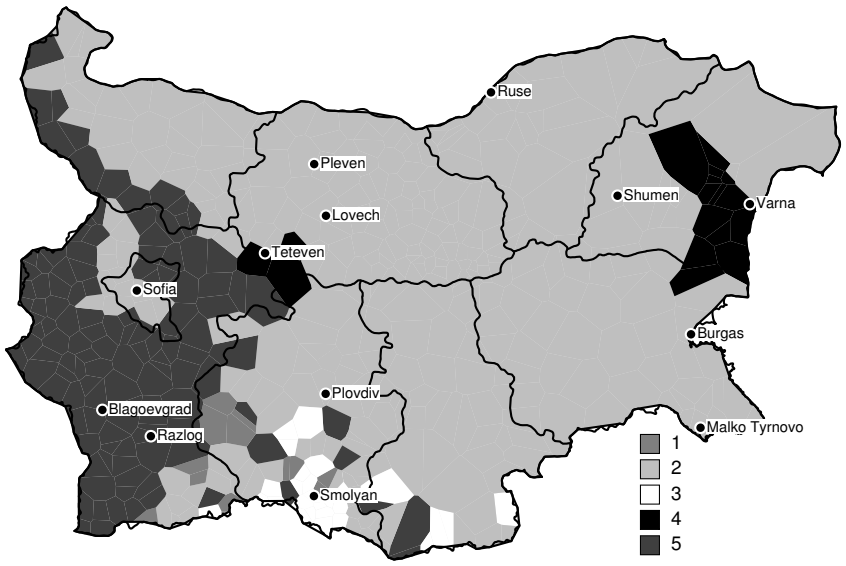

Figure 5: Five areas are distinguished, roughly the following: 1. some Rodopi dialects, 2. Eastern dialects, 3. other Rodopi dialects, 4. the Teteven region and the Northeastern Balkan region and 5. Western and South Central dialects.

possible explanation for this: in the sample the similarity of the Northwestern with the Eastern dialects is reflected extensively in the development of the stressed schwa. Northwestern and Eastern dialects have a schwa in contrast to Southwestern dialects, where we find an [a]. The map emphasizes a Northeastern Balkan dialect area near Varna and Dobrich. This detail deserves attention, because this group behaves distinctly in the continuum map, too (see below). It also distinguishes one of the Central Balkan sub-dialects of the Eastern part, namely the Teteven dialect. Recall that this dialect was noted as one of the divisions with respect to the 'big nosovka' development – the open /e/ vs. /æ/. Finally, the West region around Sofia is distinguished, which surprisingly shows similarities with Eastern areas.

Composite Cluster Map

The composite cluster map (Fig. 6) is obtained by repeatedly clustering while adding random small amounts of noise to the input distance table. By repeatedly clustering using noise, we overcome the instability inherent in clustering (see above). See Kleiweg, Nerbonne and Bosveld (2004) for details. The repetitions of the clustering procedure result in maps which are superimposed on one another in Fig. 6. The darker the line, the more frequently the boundary appeared in one of the repetitions of clustering.

A composite cluster map for Bulgarian was created with the same clustering method used to obtain the dendogram in Fig. 4. This map shows the most significant divisions of the groups, where the most significant border is again the one that divides the dialects into Western and Eastern. Its shape is certainly similar to the one on the traditional map in Fig. 1. However, the contrast of

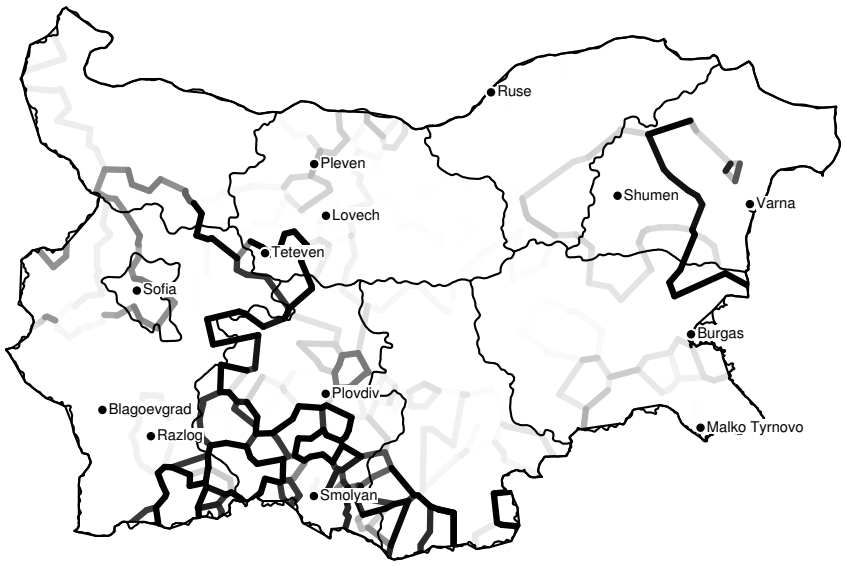

Figure 6: Composite cluster map for 36 words. The darker lines indicate more likely linguistic borders.

the northern part of the 'yat' border near to Pleven is interestingly suppressed. At the same time, the map shows that the most heterogenous group is Rodopi. Additionally, this map distinguishes the Northeastern Balkan dialect area near Varna and Dobrich, the Teteven dialect and the Plovdiv dialect.

5.2 Multidimensional scaling

5.2.1 Method

On the basis of geographic coordinates one can derive distances between any two locations. The reverse is also possible: on the basis of distances, an coordinate system can be proposed within which the data points may be located. The last is realized by a technique known as 'multidimensional scaling' (MDS), which, unlike clustering *is* statistically stable. In an MDS plot, strongly related dialects are close to each other, while very different dialects are located far away from each other (Kruskal and Wish, 1984).

As input each dialect is viewed as located by a set of distances, namely the distance to itself and the distances to other dialects. The other sites correspond thus to dimensions, so that if we have 490 dialects, we get 490 dimensions. Using MDS these 490 dimensions can be reduced to two, three or four. Each site is then identified via coordinates in 2-, 3-, or 4-dimensional space.

MDS also allows the definition of a color map where the resulting MDS dimensions are assigned colors, and the intensity of a color at a given point reflects the MDS coordinates assigned to that site. If our 490 dimensions are scaled to 3 dimensions, and we let the 3 dimensions be the intensities of respectively red, green and blue, then each dialect site is assigned a mixture of these three colors.

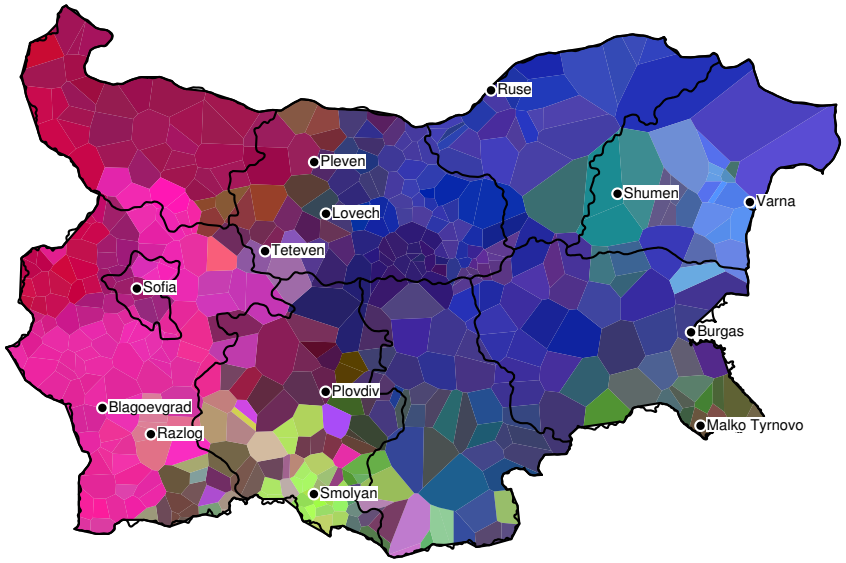

Figure 7: This continuum map highlights two subgroups in Western Bulgaria (Northeastern and South Western) as well as the most Northeastern subgroup.

The result is a map which reflects the gradual changes in dialect characteristics (see Figure 7). The map is not constrained to reflect a dialect continuum, which is why the polygons surrounding a given site may change color abruptly. To the degree that the color shifts are not abrupt, we are dealing with a continuum.

5.2.2 Results

The map in Fig. 7, which does not depend at all on clustering, represents the dialects as forming a continuum, a view which most dialectologists find congenial. The map suggests that Eastern Bulgaria is a rather uniform area, where only the far Eastern part around Varna and Shumen is distinguished. Also, the Western part is clearly divided into two parts: Northwestern and Southwestern. The dialectally least uniform area remains Rodopi.

5.3 Comparison

In general, the important border between Western and Eastern dialects is present in all the maps. Within the Western group two varieties are regularly detected: Northwest and Southwest groups, in conformity with the traditional division. In Eastern Bulgaria the Moesian and Balkan dialects are not distinguished sharply, which may reflect the fact that a number of migrations have taken place in this area. Of course, we should bear in mind the fact that our procedure is based exclusively on pronunciation, and not all on morphology or the lexicon. The Balkan group around Varna and Shumen is partly distinguished, which may due to the irregularity in the grid of data collection sites. We regard our results as "partial" when compared to the view embodied in the map on

Fig. 1. The most heterogenous groups seem to be the Rodopian group, and the Eastern Balkan group. We examine these separately below.

If we compare the maps we have derived to the maps in the generalized volume of Bulgarian dialects (Kochev 1988), then we can see the following: our maps reflect e.g. the 'yat' border between Western and Eastern dialects (see map 2/1[5]) and also the borders in the Atlas map 2/2, which distinguishes the Balkan area around Varna and Shumen and the varieties within the Rodopian group and the neighboring Southwestern groups. The features which were taken into account in these two traditional dialect maps were the oppositions: palatal vs. non-palatal consonants and closed vs. open front vowels. These were reflected in our data as well.

6 The Rodopian and Eastern Balkan groups

Before seeking groups each of 490 sites was labeled with the name of the main dialect to which it is supposed to belong. The labels are meant to facilitate the systematic examination of the groups. The main dialects are abbreviated in the labels, and we used the labels in the dendrograms presented in this section. The dialects are as follows:
Northeastern dialects: Moesian (M); Balkan (Ba); South Balkan (Pba)
Southeastern dialects: Thracian (Thr); West Rupskian (Zrup); Rodopian (Ro)
Northwestern dialects: West Moesian (ZM); Northwestern bordering (SzP)
Southwestern dialects: Central Southwestern (CYuz); Southwestern bordering (YuzP);
Standard (St)

We focus now in more detail on two interesting and distinct groups, namely the Eastern Balkan group and the Rodopian groups. We will in particular examine them via the dendrograms resulting from clustering.

The Eastern Balkan group around Varna and Shumen is a compact group that is distant from the surrounding dialects.[6] This group seems to behave differently with respect to one very decisive phonetic feature, namely open /e/ vs. /æ/, which is the realization of 'yat' before syllables with a closed vowel, of schwa in a closed syllable, of "small er", and of the big nosovka. This group is shown in Fig. 8. Note that the closest neighboring dialects are the so called 'erkech' dialects (Kazichino and Golitsa), and the Teteven dialects, which are subdialects of the Central Balkan dialect. Recall that this connection is present in the area map (Fig. 5).

Traditionally, the Southeastern group includes: Eastern Rupskian dialects (which are presented as Rodopian and Thracian here), Central Rupskian dialects (presented as Rodopian here), and Western Rupskian dialects (presented as Western Rupskian here as well). The maps show that Eastern Rupskian and

[5]We use the numbering of the maps also used in the atlas.

[6]Note that it is marked as Balkan in the traditional map, but we view it as a group of Moesian dialects, because the Moesian and Balkan dialects merge in this area allowing us to reserve the term 'Balkan' for the Central Balkan area around the Central Old mountains and the Sredna Gora mountain. We find this clearer, at least in drawing comparisons.

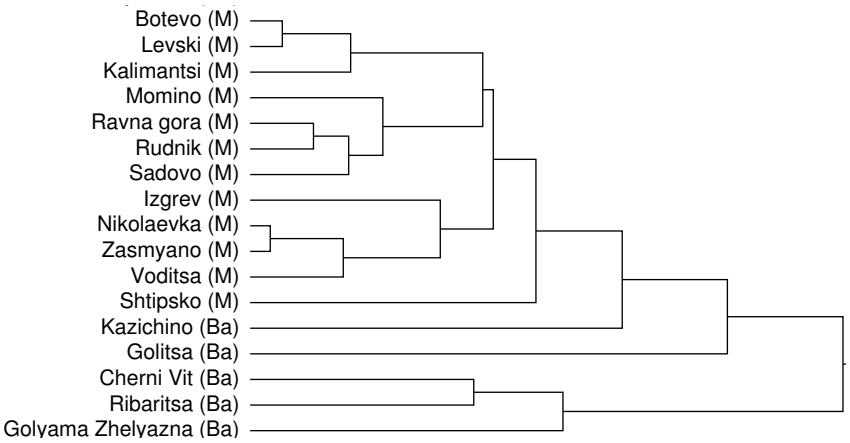

Figure 8: The Moesian dialects group very uniformly together. Recall that M stands for Moesian and Ba for Balkan.

Western Rupskian (with some small exceptions) are more closely unified than Central Rupskian (Rodopian). In this group, the differing features are more prominent than the unifying ones. The dendrograms show the distinct sub-dialect groups: dialects of Smolyan, of Zlatograd, of Velingrad, etc.

The first Rodopian cluster (from Ahryane to Tihomir in Fig. 9) is rather compact in the dendrogram. The main dialect is Smolyan. We can also distinguish some of the other neighboring sub-dialects, such as the 'Shiroka lyka' dialect (Shiroka lyka, Stoykite, etc.). Its closest non-Rodopian clusters are Balkan and Moesian.

The second cluster (see Fig. 10) presents the Chepinski Rodopian sub-dialect. It is closely related to West Rupskian dialects.

The third Rodopian cluster (see Fig. 11) is closer to Southwestern bordering dialects and Moesian ones.

The fourth Rodopian cluster (see Fig. 12) shows the Hvoyna Rodopian sub-dialect, which is interestingly separated from the other Rodopian sub-dialects present in the first dendrogram: Smolyan, Shiroka lyka, Chepintsi. It is closer to West Rupskian dialects and Balkan dialects.

The fifth Rodopian cluster (see Figure 13) shows the Zlatograd Rodopian sub-dialect and its relatedness to some of the dialects around Smolyan. It shows similarities to the Thracian dialects as well.

To sum up, these subclusters emphasize the integrity of smaller dialect areas even in those areas which show little overall cohesiveness. The traditional groups do not always cluster uniformly and exhaustively together. On the contrary, they occasionally merge, even while displaying many connections among sub-dialects.

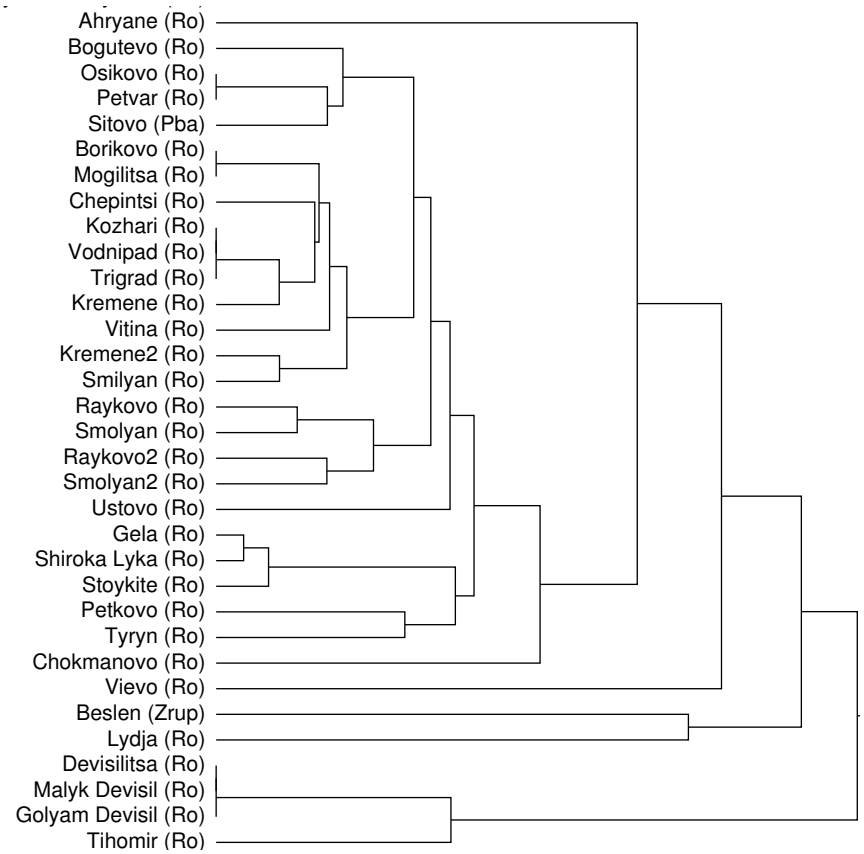

Figure 9: Smolyan Rodopian and related varieties.

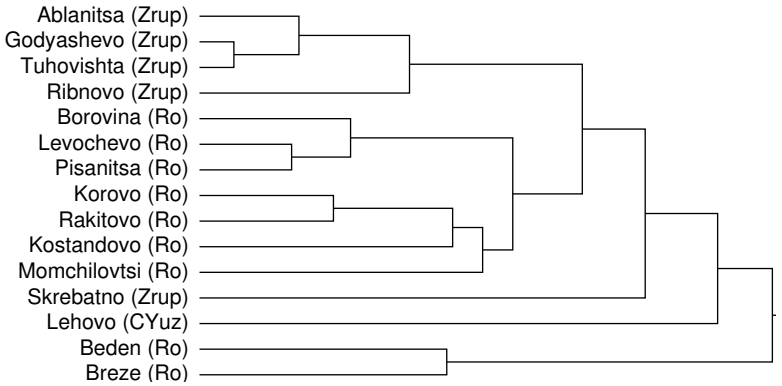

Figure 10: The Rodopian dialects are closest to the West Rupskian dialects.

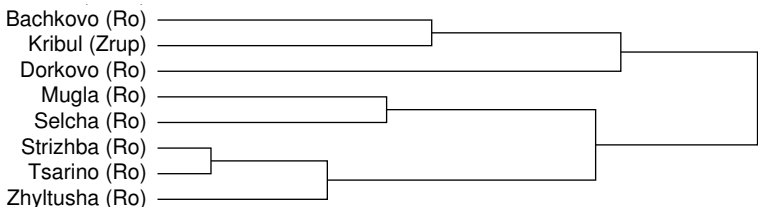

Figure 11: Rodopian dialects close to bordering Southwestern dialects.

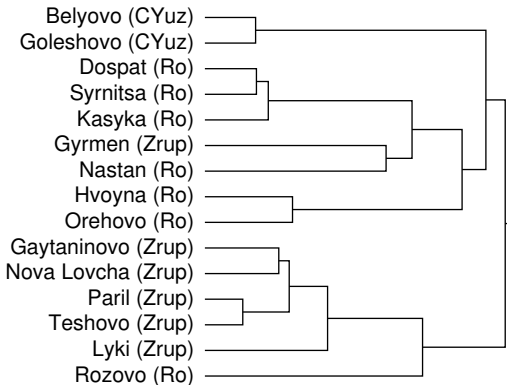

Figure 12: The Hvoyna Rodopian sub-dialect.

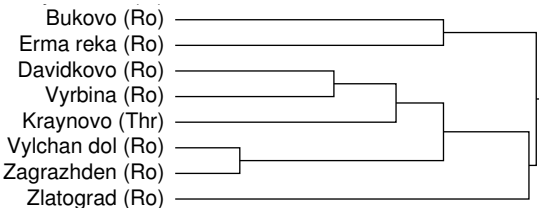

Figure 13: The Zlatograd Rodopian sub-dialect.

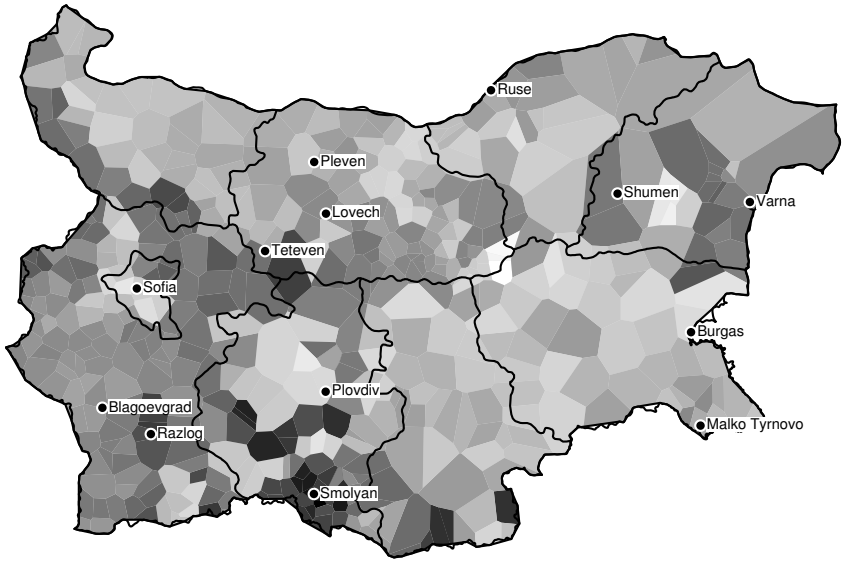

Figure 14: The lightest areas are most similar to Standard Bulgarian, while the darkest ones are least similar.

7 The Standard

The Bulgarian standard language is regarded as close to dialects in the Northeast (Central Balkan dialect), (Stoykov 2002, p. 108). At the same time, some similarity to the Southeastern sites would not be surprising either, first, because it would confirm the unity of Eastern dialects with respect to the important feature of 'yat' realizations, and second, because there are phonetic features that are typical for the Northeast, but not accepted in the standard language, including the reduction of open vowels and the existence of final palatal consonants.

We measured the distance from the standard to each of the 490 varieties in our sample using the same techniques described above. We illustrate the results of these measurements in Fig. 14, displaying the distances of the varieties to the standard.

Several comments are in order. The general picture suggests that the most similar dialects are the Central, Southeastern and Northwestern ones. The Western and Rodopian dialects are least similar, which is not surprising bearing in mind the source of the Standard language. Note that the Rodopian area (Smolyan) as well the West Rupskian area (Razlog) are not homogeneously similar to the Standard. At the same time, it is interesting to notice that the most similar and the least similar areas are often geographically close. For example, the Teteven region shows dissimilarity near the 'yat' border and similarity in the East. At the same time the Shumen region and the entire Northeast seem to be dissimilar from the Standard. Perhaps these effects are caused by the transitional dialect zones (Pleven, Teteven, Razlog). It is said that the Shumen dialect is very close to the one in Thessaloniki, Greece.

8 Discussion of the Larger Sample

For the sake of completeness, and in order to put our claim of consistency to the test, we briefly discuss the similarities and differences between results based on the 54-word set and those based on the 36-word set (discussed above). No new maps need to be presented, but we will wish to refer to the ones above. First, we note that the correlation between distances assigned on the basis of the 36-word set and those assigned on the basis of the 54-word set is very high indeed ($r = 0.97$). The high correlation indicates that the 36 common words are representative of the data in the atlases. For example, the line map for 54-word set (similar to Fig. 3), the MDS results (similar to Fig. 7), and the composite cluster maps (similar to Fig. 6) all highly resemble the results for the 36-word set.

Nonetheless, the more unstable clustering gives somewhat differing results, which may be taken into consideration as well. We shall examine these results briefly, even at the risk of overemphasizing less stable aspects of the analysis. The dendrogram for 54 words and 5 groups highlights the West, East, Rodopian and Northeastern Balkan groups. The 36-word dendrogram seems to recognize more distinct areas in Bulgaria (Fig. 4). As a result, the classifying map for 54 words stresses the West-East distinction in a way that is closer to the traditional division. In the 54-word set some more dialect features are included (the ones not relevant throughout the four atlases) and these justify the East-West division more substantially. One such feature is, for example, the presence/omission of the fricative [x] in different positions in the word. Some of the West Moesian dialects are distinguished—in the area of Pleven and Lovech. This is again not very surprising as these dialects are at the border of the 'yat' division, and they exhibit similarities to Western as well as Eastern dialects. The other distinct area in 54-word set is the dialect of Pavlikjans, in the Plovdiv region. It contains a lot of archaic Rodopian features.

Furthermore, the clusters for Rodopi differ in number. For the 54-word set it shows only 3 compact groups of dialects: Smolyan, Zlatograd and the varieties closer to West Rupskian.

9 Conclusions and prospects

In this paper we applied dialectological techniques which had been successfully applied to Germanic and Romance language to the Slavic language, Bulgarian. The application involved including a novel set of sounds and a dialect area which is famous for its contact phenomena. Nonetheless, the techniques could be applied straightforwardly, indicating that they are more general than had been shown until now.

We analyzed the relations among Bulgarian dialects based on the pronunciation distances calculated between all pairs of 490 Bulgarian varieties, and including additionally the standard language. The data was selected and digitized from the four-volume set of Atlases of Bulgarian Dialects. The relations among Bulgarian dialects turned out to be quite complex. The comparison with

the earlier maps and their dialect divisions showed that the most significant border remains the 'yat' realization border. However, our research also confirms Balan's claims (Teodorov-Balan 1904) about the need to introduce a third major dialect area, namely: Rodopi. With respect to the four-way distinction in the atlases, our results suggest that the Western dialects are more cohesive than the Eastern ones, and that within the Western dialects, North and South are additionally distinguished. The East shows weaker cohesion among dialects, and its North and South areas are more uniform. One explanation could be that more migrations have taken place in the East than in the West. At the same time, several special subclusters are also distinguished: the Rodopian region, the Northeastern Balkan region and some interesting areas near the 'yat' border.

We also explored the relation of the dialects to the standard language. We found that the standard is most similar to South Balkan dialects from the Northeast, which was expected. We also observed interesting that most similar and least similar varieties need not be geographically near one another.

In our opinion, this work paints a faithful picture of Bulgarian dialects even though it is based on a limited number of word pronunciations. In addition, our more inclusive data set confirmed the results of the smaller one, even while hinting at potentially different dialect subgroups.

We see the future work in several directions. We should like to examine different dialect data, and in particular data collected from sites that were not selected for being purely Bulgarian. It would be important to identify the regular aspects of the distinctions at the base of the analysis here, i.e. the linguistic basis of the aggregate analysis. It would be interesting to include lexical variation (Nerbonne and Kleiweg, 2003) in a parallel analysis, and to examine the degree to which lexical differences correlate with differences in pronunciations. Finally, Bulgarian and the Balkan are most famous linguistically for the extensive language contact which has developed there, and it would be fascinating to modify and apply the quantitative techniques used here in order to explore and analyze language contact.

10 Acknowledgments

The authors would like to thank Kiril Simov for his kind help in the digitization of the data, Luchia Antonova for her valuable comments on the X-SAMPA conversion, for the selection of the Bulgarian sites and her recommendations on Bulgarian dialects in the process of our work, Christine Siedle for her kind help with the geographical coordinates and the maps, Peter Kleiweg for the software facilities and his quick reactions on software questions.

We also owe thanks to Renee van Bezooijen, Charlotte Gooskens, Marco Spruit and Bill Kretzschmar for the very useful discussion and suggestions on a preliminary version of this paper.

References

[Bolognesi and Heeringa 2002] R. Bolognesi and W. Heeringa. De invloed van dominante talen op het lexicon en de fonologie van Sardische dialecten. In: D. Bakker, T. Sanders, R. Schoonen and Per van der Wijst (eds.). In: *Gramma/TTT: tijdschrift voor taalwetenschap* , 9 (1), 2002, 45-84. Available at: http://www. let.rug.nl/˜heeringa/dialectology/ papers/.

[Gooskens and Heeringa 2004] Charlotte Gooskens and Wilbert Heeringa. Perceptive Evaluation of Levenshtein Dialect Distance Measurements using Norwegian Dialect Data. In: *Language variation and Change*, 16, 2004, pp. 189–207.

[Handbook of the International Phonetic Association 2003] *A Guide to the Use of the International Phonetic Alphabet*. Cambridge University Press. 2003.

[Heeringa 2004] Wilbert Heeringa *Measuring Dialect Pronunciation Differences using Levenshtein Distance* . Doctoral dissertation. University of Groningen, Groningen. 2004. Avail. at: http://www.let.rug.nl/˜heeringa/dialectology/thesis/.

[Hoppenbrouwers and Hoppenbrouwers 2001] C. Hoppenbrouwers and G. Hoppenbrouwers *De indeling van de Nederlandse streektalen. Dialekten van 156 steden en dorpen geklasseerd volgens de FFM* . Koninkljike Van Gorcum B.V., Assen. 2001.

[Jain and Dubes 1988] A.K. Jain and R.C. Dubes *Algorithms for Clustering Data*. Prentice Hall, Englewood Cliffs, New Jersey. 1988.

[Kleiweg et. al. 2004] Peter Kleiweg, John Nerbonne and Leonie Bosveld. Geographic Projection of Cluster Composites. In: Alan Blackwell, Kim Marriott and Atsushi Shimojima (eds.) *Diagrammatic Representation and Inference. Third International Conference, Diagrams 2004* . Cambridge, UK, March 2004 Lecture Notes on Artificial Intelligence 2980. Springer, Berlin 2004, pp. 392–394.

[Kruskal and Wish 1978] J.B. Kruskal and M. Wish *Multidimensional scaling*. In: Number 07-011 in Sage University Paper Series on Quantitative Applications in the Social Sciences. Sage Publications, Newbury Park. 1978.

[Nerbonne 2006] John Nerbonne. Various Variation Aggregates in the LAMSAS South. Accepted to appear in: Catherine Davis and Michael Picone (eds.) *Language Variety in the South III*. University of Alabama Press, Tuscaloosa. 2006.

[Nerbonne and Heeringa 2001] John Nerbonne and Wilbert Heeringa. Computational Comparison and Classification of Dialects. In: *Dialectologia et Geolinguistica* 9, 2001, pp. 69–83.

[Nerbonne et. al. 1996] John Nerbonne, Wilbert Heeringa, Erik van den Hout, Peter van der Kooi, Simone Otten, Willem van de Vis Phonetic Distance

between Dutch Dialects. In: G. Durieux, W. Daelemans, and S. Gillis (eds.). *CLIN VI, Papers from the sixth CLIN meeting*. University of Antwerp, Center for Dutch Language and Speech, Antwerpen, 1996, 185-202. Available at: http://www.let.rug. nl/~heeringa/dialectology/ papers/.

[Nerbonne and Kleiweg 2003] John Nerbonne and Peter Kleiweg Lexical Distance in LAMSAS. In: John Nerbonne and William Kretzschmar(eds.) *Computational Methods in Dialectometry*. Special issue of *Computers and the Humanities*, 37(3), 2003, pp. 339–357.

[Nerbonne and Siedle 2005] John Nerbonne and Christine Siedle. Dialektklassifikation auf der Grundlage aggregierter Ausspracheunterschiede. In: *Zeitschrift fur Dialektologie und Linguistik* 72(2), 2005, pp. 129-147. Available at: http://www.let.rug.nl/ ~nerbonne/paper.html

[Kochev 1988] Kochev Iv. *Atlas of Bulgarian Dialects*, Generalized volume, Publishing House of the Bulgarian Academy of Sciences, 1988, volume II, Sofia, Bulgaria. 1988. (In Bulgarian)

[Popov et. al. 1998] Popov D., Simov K. and Vidinska Sv. *A Dictionary of Writing, Pronunciation and Punctuation of Bulgarian Language*. Atlantis LK, Sofia, Bulgaria. 1998. (In Bulgarian)

[Simov et. al. 2004] Simov K., Simov A., Ganev H., Ivanova K., Grigorov I. The CLaRK System: XML-based Corpora Development System for Rapid Prototyping. In: *Proceedings of LREC 2004*, Lisbon, Portugal, 2004. pp. 235–238

[Stoykov 2002] Stoykov St. *Bulgarska dialektologiya*. Sofia, 4th ed. 2002. (In Bulgarian)

[Stoykov 1966] Stoykov St. *Atlas of Bulgarian Dialects: Northeastern Bulgaria*, Publishing House of the Bulgarian Academy of Sciences, volume II, Sofia, Bulgaria. 1966. (In Bulgarian)

[Stoykov and Bernshteyn 1964] Stoykov St. and Bernshteyn S. B. *Atlas of Bulgarian Dialects: Southeastern Bulgaria*, Publishing House of the Bulgarian Academy of Sciences, volume I, Sofia, Bulgaria. 1964. (In Bulgarian)

[Stoykov et. al. 1981] Stoykov St., Kochev Iv. and Mladenov M. *Atlas of Bulgarian Dialects: Northwestern Bulgaria*, Publishing House of the Bulgaria Academy of Sciences, volume IV, Sofia, Bulgaria. 1981. (In Bulgarian)

[Stoykov et. al. 1975] Stoykov St., Mirchev K., Kochev Iv. and Mladenov M. *Atlas of Bulgarian Dialects: Southwestern Bulgaria*, Publishing House of the Bulgarian Academy of Sciences, volume III, Sofia, Bulgaria. 1975. (In Bulgarian)

[Teodorov-Balan 1904] Teodorov-Balan Al. *Rodopskoto narechie*. Sbornik statij po slavyanovedeniyu, posveshtennyich prof. M. S. Drinovu. Harkov, pp. 111-127. 1904. (In Bulgarian)

[Wells] Wells John *Computer-coding the IPA: a proposed extension of SAMPA* . Available at: http://www.phon.ucl.ac.uk/home/sampa/x-sampa.htm