Identifying Linguistic Structure in a Quantitative Analysis of Dialect Pronunciation

Jelena Prokić
Alfa-Informatica
University of Groningen
The Netherlands
j.prokic@rug.nl

Abstract

The aim of this paper is to present a new method for identifying linguistic structure in the aggregate analysis of the language variation. The method consists of extracting the most frequent sound correspondences from the aligned transcriptions of words. Based on the extracted correspondences every site is compared to all other sites, and a correspondence index is calculated for each site. This method enables us to identify sound alternations responsible for dialect divisions and to measure the extent to which each alternation is responsible for the divisions obtained by the aggregate analysis.

1 Introduction

Computational dialectometry is a multidisciplinary field that uses quantitative methods in order to measure linguistic differences between the dialects. The distances between the dialects are measured at different levels (phonetic, lexical, syntactic) by aggregating over entire data sets. The aggregate analyses do not expose the underlying linguistic structure, i.e. the specific linguistic elements that contributed to the differences between the dialects. This is very often seen as one of the main drawbacks of the dialectometry techniques and dialectometry itself. Two attempts to overcome this drawback are presented in Nerbonne (2005) and Nerbonne (2006). In both of these papers the identification of linguistic structure in the aggregate analysis is based on the analysis of the pronunciation of the vowels found in the data set.

In work presented in this paper the identification of linguistic structure in the aggregate analysis is based on the automatic extraction of regular sound correspondences which are further quantified in order to characterize each site based on the frequency of a certain sound extracted from the pool of the site's pronunciation. The results show that identification of regular sound correspondences can be successfully applied to the task of identifying linguistic structure in the aggregate analysis of dialects based on word pronunciations.

The rest of the paper is structured as follows. Section 2 gives an overview of the work previously done in the areas covered in this paper. In Section 3 more information on the aggregate analysis of Bulgarian dialects is given. Work done on the identification of regular sound correspondences and their quantification is presented in Section 4. Conclusion and suggestions for future work are given in Section 5.

2 Previous Work

The work presented in this paper can be divided in two parts: the aggregate analysis of Bulgarian dialects on one hand, and the identification of linguistic structure in the aggregate analysis on the other. In this section the work closely related to the one presented in this paper will be described in more detail.

2.1 Aggregate Analysis of Bulgarian

Dialectometry produces aggregate analyses of the dialect variations and has been done for different languages. For several languages aggregate analyses have been successfully developed which distinguish various dialect areas within the language area. The

Proceedings of the ACL 2007 Student Research Workshop, pages 61–66,
Prague, June 2007.

most closely related to the work presented in this paper is quantitative analysis of Bulgarian dialect pronunciation reported in Osenova et al. (2007).

In work done by Osenova et al. (2007) aggregate analysis of pronunciation differences for Bulgarian was done on the data set that comprised 36 word pronunciations from 490 sites. The data was digitalized from the four-volume set of Atlases of Bulgarian Dialects (Stojkov and Bernstein, 1964; Stojkov, 1966; Stojkov et al., 1974; Stojkov et al., 1981). Pronunciations of the same words were aligned and compared using L04.[1] Results were analyzed using cluster analysis, composite clustering, and multidimensional scaling. The analyses showed that results obtained using aggregate analysis of word pronunciations mostly conform with the traditional phonetic classification of Bulgarian dialects as presented in Stojkov (2002).

2.2 Extraction of Linguistic Structure

Although techniques in dialectometry have shown to be successful in the analysis of the dialect variation, all of them aggregate over the entire available data, failing to extract linguistic structure from the aggregate analysis. Two attempts to overcome this withdraw are presented in Nerbonne (2005) and Nerbonne (2006).

Nerbonne (2005) suggests aggregating over a linguistically interesting subset of the data. Nerbonne compares aggregate analysis restricted to vowel differences to those using the complete data set. Results have shown that vowels are probably responsible for a great deal of aggregate differences, since there was high correlation between differences obtained only by using vowels and by using complete transcriptions (r = 0.936). Two ways of aggregate analysis also resulted in comparable maps. However, no other subset has been analyzed in this paper, making it impossible to conclude how successful other subsets would be if similar analysis was done.

The second paper (Nerbonne, 2006) applies factor analysis to the result of the dialectometric analysis in order to extract linguistic structure. The study focuses on the pronunciation of vowels found in the data. Out of 1132 different vowels found in the data 204 vowel positions are investigated, where a vowel position is, e.g., the first vowel in the word 'Washington' or the second vowel in the word 'thirty'. Factor analysis has shown that 3 factors are most important, explaining 35% of the total amount of variance. The main drawback of applying this technique in dialectometry is that it is not directly related to the aggregate analysis, but is rather an independent step. Just as in Nerbonne (2005), only vowels were examined.

2.3 Sound Correspondences

In his PhD thesis Kondrak (Kondrak, 2002) presents techniques and algorithms for the reconstruction of the proto-languages from cognates. In Chapter 6 the focus is on the automatic determination of sound correspondences in bilingual word lists and the identification of cognates on the basis of extracted correspondences. Kondrak (2002) adopted Melamed's parameter estimation models (Melamed, 2000) used in statistical machine translation and successfully applied them to determination of sound correspondences, i.e. diachronic phonology. Kondrak induced a model of sound correspondence in bilingual word lists, where phoneme pairs with the highest scores represent the most likely correspondences. The more regular sound correspondences the two words share, the more likely it is that they are cognates and not borrowings.

In this paper the identification of sound correspondences will be used to extract linguistic elements (i.e. phones) responsible for the dialect divisions. The method presented in this study differs greatly from Kondrak's in that he uses regular sound correspondences to directly compare two words and determine if they are cognates. In this study extracted sound correspondences are further quantified in order to characterize each site in the data set by assigning it a unique index. This is the first time that this method has been applied in dialectometry.

3 Aggregate Analysis

In the first phase of this project L04 toolkit was used in order to make an aggregate analysis of Bulgarian dialects. In this section more information on the data set used in the project, as well as on the process of the aggregate analysis will be given.

[1] L04 is a freely available software used for dialectometry and cartography. It can be found at http://www.let.rug.nl/kleiweg/L04/

3.1 Data Set

The data used in this research, as well as the research itself, are part of the project Buldialect—Measuring linguistic unity and diversity in Europe.[2] The data set consisted of pronunciations of 117 words collected from 84 sites equally distributed all over Bulgaria. It comprises nouns, pronouns, adjectives, verbs, adverbs and prepositions which can be found in different word forms (singular and plural, 1st, 2nd, and 3rd person verb forms, etc.).

3.2 Measuring of Dialect Distances

Aggregate analysis of Bulgarian dialects done in this project was based on the phonetic distances between the various pronunciations of a set of words. No morphological, lexical, or syntactic variation was taken into account.

First, all word pronunciations were aligned based on the following principles: a) a vowel can match only with the vowel b) a consonant can match only with the consonant c) [j] can match both vowels and consonants.

An example of the alignment of two pronunciations is given in Figure 1.[3]

g	l	ˈɑ	v	ɑ
g	l	ə	v	ˈɤ
		1		1

Figure 1: Alignment of word pronunciation pair

The alignments were carried out using the Levenshtein algorithm,[4] which also results in the calculation of a distance between each pair of words. The distance is the smallest number of insertions, deletions, and substitutions needed to transform one string to the other. In this work all three operations were assigned the same value—1. All words are represented as series of phones which are not further defined. The result of comparing two phones can be 1 or 0; they either match or they don't. In Figure 1 the cheapest way to transform one pronunciation to the other would be by making two substitutions: [ˈɑ] should be replaced by [ə], and [ɑ] by [ˈɤ], meaning that the distance between these two pronunciations is 2. The distance between each pair of pronunciations was further normalized by the length of the longest alignment that gives the minimal cost.[5] After normalization, we get the final distance between two strings, which is 0.4 (2/5) in the example shown in Figure 1. If there are more plausible alignments with the minimal cost, the longest is preferred. Word pronunciations collected from all sites are aligned and compared in this fashion, allowing us to calculate the distance between each pair of sites. The difference between two locations is the mean of all differences between words collected from these two sites.

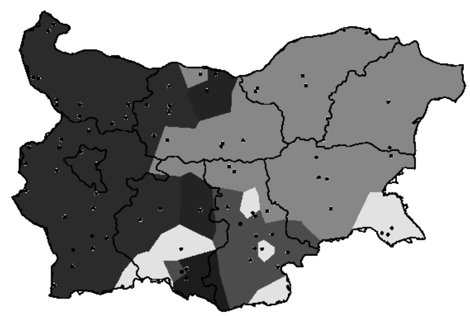

Figure 2: Classification map

The results were analyzed using clustering (Figure 2) and multidimensional scaling (Figure 3). Clustering is a common technique in a statistical data analysis based on a partition of a set of objects into groups or clusters (Manning and Schütze, 1999). Multidimensional scaling is data analysis technique that provides a spatial display of the data revealing relationships between the instances in the data set (Davison, 1992). On both the maps the biggest division is between East and West. The border between these two areas goes around Pleven and Teteven, and it is the border of "yat" realization as presented in the traditional dialectological atlases (Stojkov, 2002). The most incoherent area is the

[2]The project is sponsored by Volkswagen Stiftung. More information can be found at http://www.sfs.uni-tuebingen.de/dialectometry

[3]For technical reasons primary stress is indicated by a high vertical line before the syllable's vowel.

[4]Detailed explanation of Levensthein algorithm can be found in Heeringa (2004).

[5]An interesting discussion on the normalization by length can be found in Heeringa et al. (2006). In this paper the authors report that contrary to results from previous work (Heeringa, 2004) non-normalized string distance measures are superior to normalized ones.

area of Rodopi mountain, and the dialects present in this area show the greatest similarity with the dialects found in the Southeastern part around Malko Tyrnovo. On the map in Figure 3 it is also possible to distinguish the area around Golica and Kozichino on the East, which conforms to the maps found in Stojkov (2002). Results of the aggregate analysis conform both to the traditional maps presented in Stojkov (2002), and to the work reported in Osenova et al. (2007).

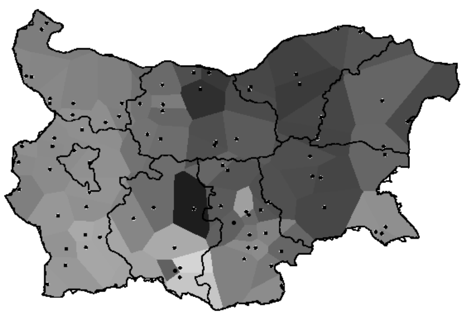

Figure 3: MDS map

4 Regular Sound Correspondences

The same data used for the aggregate analysis was reused to extract sound correspondences and to identify underlying linguistic structure in the aggregate analysis. The method and the obtained results will be presented in more detail.

4.1 Method

From the aligned pairs of word pronunciations all non-matching segments were extracted and sorted according to their frequency. In the entire data set there were 683 different pairs of sound correspondences that appeared 955199 times.

e	i	36565	j	-	21361
ə	ɤ	26398	ɑ	ə	20515
o	u	26108	e	ˈe	19934
ˈɒ	ˈe	23689	r	rʲ	19787
v	-	22100	ˈɤ	-	18867

Table 1: Most frequent sound correspondences

The most frequent correspondences were taken to be the most important sound alternations responsible for dialect variation. The method was tested on the 10 most frequent correspondences which were responsible for the 25% of sound alternations in the whole data set.

In order to determine which of the extracted sound correspondences is responsible for which of the divisions present in the aggregate analysis, each site was compared to all other sites with respect to the 10 most frequent sound correspondences. For each pair of sites all sound correspondences were extracted, including both matching and non-matching segments. For further analysis it was important to distinguish which sound comes from which place.

For each pair of the sound correspondences from Table 1 a correspondence index is calculated for each site using the following formula:

$$\frac{1}{n-1}\sum_{i=1,j\neq i}^{n} s_i {\longrightarrow} s'_j \tag{1}$$

where n represents the number of sites, and $s_i {\longrightarrow} s'_j$ the comparison of each two sites (i, j) with respect to the sound correspondence s/s'. $s_i {\longrightarrow} s'_j$ is calculated applying the following formula:

$$\frac{|s_i, s'_j|}{|s_i, s'_j| + |s_i, s_j|} \tag{2}$$

In the above formula s_i and s'_j stand for the pair of sounds involved in one of the most frequent sound correspondences from Table 1. $|s_i, s'_j|$ represents the number of times s is seen in the word pronunciations collected at site i, aligned with the s' in word pronunciations collected at site j. $|s_i, s_j|$ is the number of times s stayed unchanged. For each pair of sound correspondences a correspondence index was calculated for the s, s' correspondence, as well as for the s', s correspondence. For example, for the pair of correspondences [e] and [i], the relation of [e] corresponding to [i] is separated from the relation of [i] corresponding to [e].[6]

For example, the indices for the sites Aldomirovci and Borisovo with respect to the sound correspondence [e]-[i] were calculated in the following way. In the file with the sound correspondences extracted from all aligned word pronunciations collected at

[6]It would also be possible to modify this formula and calculate the ratio of s to s corresponding to any other sound. In this case the result would be a very small number of sites with the very high correspondence index.

these two sites, the algorithm searches for pairs represented in Table 2:

Aldomirovci	e	i	e
Borisovo	i	e	e
no. of correspondences	24	0	3

Table 2: How often [e] corresponds to [i] and [e]

For each of the sites the indices were calculated using the above formula. The index for site i (Aldomirovci) was:

$$\frac{|e,i|}{|e,i|+|e,e|} = \frac{24}{24+3} = 0.89 \quad (3)$$

The index for site j (Borisovo) was calculated in the similar fashion from the Table 2:

$$\frac{|e,i|}{|e,i|+|e,e|} = \frac{0}{0+3} = 0.00 \quad (4)$$

Each of these two sites was compared to all other sites with respect to the [e]-[i] correspondence resulting in 83 indices for each site. The general correspondence index for each site represents the mean of all 83 indices. For the site i (Aldomirovci) general index was 0.40, and for the site j (Borisovo) 0.21. Sites with the higher values of the general correspondence index represent the sites where sound [e] tends to be present, with respect to the [e]-[i] correspondence (see Figure 4). In the same fashion general correspondence indices were calculated for every site with respect to each pair of the most frequent correspondences (Table 1).

4.2 Results

The methods described in the previous section were applied to all phone pairs from the Table 1, resulting in 17 different divisions of the sites.[7]

Data obtained by the analysis of sound correspondences, i.e. indices of correspondences for sites was used to draw maps in which every site is set off by Voronoi tessellation from all other sites, and shaded based on the value of the general correspondence index. Light polygons on the map represent areas with the higher values of the correspondence index, i.e. areas where the first sound in the examined alternation tends to be present. This technique enables us to visualize the geographical distribution of the examined sounds. For example, map in Figure 4 represents geographical distribution of sound [e] with respect to the [e]-[i] correspondence, while map in Figure 5 reveals the presence of the sound [i] with respect to the [i]-[e] correspondence.

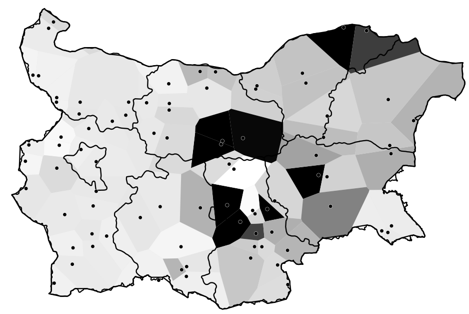

Figure 4: Distribution of [e] sound

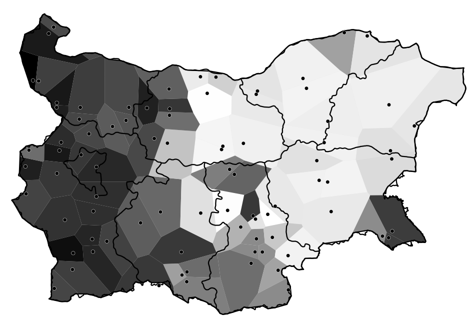

Figure 5: Distribution of [i] sound

In order to compare the dialect divisions obtained by the aggregate analysis, and those based on the general correspondence index for a certain phone pair, correlation coefficient was calculated for these 2 sets of distances. The results are shown in Table 3. Dialect divisions based on the [r]-[r^j] and [i]-[e] alternations have the highest correlation with the distances obtained by the aggregate analysis. The square of the Pearson correlation coefficient presented in column 3 enables us to see that 39.0% and 30.7% of the variance in the aggregate analysis can be explained by these two sound alternations.

[7]For three pairs where one sound doesn't have a corresponding one (when there was an insertion or deletion) it is not possible to calculate an index. Formulas for comparing two sites from the previous section would always give value 1 for the index.

Correspondence	Correlation	r^2x100(%)
[e]-[i]	0.19	3.7
[i]-[e]	0.55	30.7
[ə]-[ɤ]	0.26	6.7
[ɤ]-[ə]	0.23	5.3
[o]-[u]	0.49	24.4
[u]-[o]	0.43	18.9
[ˈɑ]-[ˈe]	0.49	24.3
[ˈe]-[ˈɑ]	0.38	14.2
[v]- -	0.14	2.0
[j]- -	0.20	4.0
[ɑ]-[ə]	0.51	26.5
[ə]-[ɑ]	0.26	7.0
[e]-[ˈe]	0.18	3.2
[ˈe]-[e]	0.23	5.2
[r]-[rʲ]	0.62	39.0
[rʲ]-[r]	0.53	28.1
[ˈɤ]- -	0.17	2.9

Table 3: Correlation coefficient

5 Conclusion and Future Work

The dialect division of Bulgaria based on the aggregate analysis presented in this paper conforms both to traditional maps (Stojkov, 2002) and to the work reported in Osenova et al. (2007), suggesting that the novel data used in this project is representative. The method of quantification of regular sound correspondences described in the second part of the paper was successful in the identification of the underlying linguistic structure of the dialect divisions. It is an important step towards more general investigation of the role of the regular sound changes in the language dialect variation. The main drawback of the method is that it analyzes one sound alternation at the time, while in the real data it is often the case that one sound corresponds to several other sounds and that sound correspondences involve series of segments.

In future work some kind of a feature representation of segments should be included in the analysis in order to deal with the drawbacks noted. It would also be very important to analyze the context in which examined sounds appear, since we can talk about regular sound changes only with respect to the certain phonological environments.

References

Mark L. Davison. 1992. *Multidimensional scaling*. Melbourne, Fl. CA: Krieger Publishing Company.

Wilbert Heeringa, Peter Kleiweg, Charlotte Gooskens, and John Nerbonne. 2006. Evaluation of String Distance Algorithms for Dialectology. In John Nerbonne and Erhard Hinrichs, editors, *Linguistic Distances*. Workshop at the joint conference of International Committee on Computational Linguistics and the Association for Computational Linguistics, Sydney.

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

Grzegorz Kondrak. 2002. *Algorithms for Language Reconstruction*. PhD Thesis, University of Toronto.

Chris Manning and Hinrich Schütze. 1999. *Foundations of Statistical Natural Language Processing*. MIT Press. Cambridge, MA.

I. Dan Melamed. 2000. Models of translational equivalence among words. *Computational Linguistics*, 26(2):221–249.

John Nerbonne. 2005. Various Variation Aggregates in the LAMSAS South. In Catherine Davis and Michael Picone, editors, *Language Variety in the South III*. University of Alabama Press, Tuscaloosa.

John Nerbonne. 2006. Identifying Linguistic Structure in Aggregate Comparison. *Literary and Linguistic Computing*, 21(4).

Petya Osenova, Wilbert Heeringa, and John Nerbonne. 2007. A Quantitive Analysis of Bulgarian Dialect Pronunciation. Accepted to appear in *Zeitschrift für slavische Philologie*.

Stojko Stojkov and Samuil B. Bernstein. 1964. *Atlas of Bulgarian Dialects: Southeastern Bulgaria*. Publishing House of Bulgarian Academy of Science, volume I, Sofia, Bulgaria.

Stojko Stojkov, Kiril Mirchev, Ivan Kochev, and Maksim Mladenov. 1974. *Atlas of Bulgarian Dialects: Southwestern Bulgaria*. Publishing House of Bulgarian Academy of Science, volume III, Sofia, Bulgaria.

Stojko Stojkov, Ivan Kochev, and Maksim Mladenov. 1981. *Atlas of Bulgarian Dialects: Northwestern Bulgaria*. Publishing House of Bulgarian Academy of Science, volume IV, Sofia, Bulgaria.

Stojko Stojkov. 1966. *Atlas of Bulgarian Dialects: Northeastern Bulgaria*. Publishing House of Bulgarian Academy of Science, volume II, Sofia, Bulgaria.

Stojko Stojkov. 2002. *Bulgarska dialektologiya*. Sofia, 4th ed.