Exploring dialect phonetic variation using PARAFAC

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

Tim Van de Cruys
University of Groningen
The Netherlands
t.van.de.cruys@rug.nl

Abstract

In this paper we apply the multi-way decomposition method PARAFAC in order to detect the most prominent sound changes in dialect variation. We investigate various phonetic patterns, both in stressed and unstressed syllables. We proceed from regular sound correspondences which are automatically extracted from the aligned transcriptions and analyzed using PARAFAC. This enables us to analyze simultaneously the co-occurrence patterns of all sound correspondences found in the data set and determine the most important factors of the variation. The first ten dimensions are examined in more detail by recovering the geographical distribution of the extracted correspondences. We also compare dialect divisions based on the extracted correspondences to the divisions based on the whole data set and to the traditional scholarship as well. The results show that PARAFAC can be successfully used to detect the linguistic basis of the automatically obtained dialect divisions.

1 Introduction

Dialectometry is a multidisciplinary field that uses quantitative methods in the analysis of dialect data. From the very beginning, most of the research in dialectometry has been focused on the identification of dialect groups and development of methods that would tell us how similar (or different) one variety is when compared to the neighboring varieties. Dialect data is usually analyzed on the aggregate level by summing up the differences between various language varieties into a single number. The main drawback of aggregate analyses is that it does not expose the underlying linguistic structure, i.e. the specific linguistic elements that contributed to the differences between the dialects. In recent years there have been several attempts to automatically extract linguistic basis from the aggregate analysis, i.e. to determine which linguistic features are responsible for which dialect divisions. Although interesting for dialectology itself, this kind of research is very important in the investigation of sound variation and change, both on the synchronic and diachronic level.

The paper is structured as follows. In the next section, we discuss a number of earlier approaches to the problem of identifying underlying linguistic structure in dialect divisions. In section 3, we give a description of the dialect data used in this research. Section 4 then describes the methodology of our method, explaining our data representation using tensors, our three-way factorization method, and the design of our data set. In section 5, the results of our method are discussed, examining the values that come out of our factorization method in a number of ways. Section 6, then, draws conclusions and gives some pointers for future work.

2 Previous work

In order to detect the linguistic basis of dialect variation Nerbonne (2006) applied factor analysis to the results of the dialectometric analysis of southern American dialects. The analysis is based on 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 variation. However, this approach is based only on vowel positions in specific words.

Prokić (2007) extracted the 10 most frequent non-identical sound correspondences from the aligned word transcriptions. Based on the relative frequency of each of these correspondences each site in the data set was assigned a *correspondence index*. Higher value of this index indicates sites

Proceedings of the 11th Meeting of the ACL-SIGMORPHON, ACL 2010, pages 46–53,
Uppsala, Sweden, 15 July 2010.

where the presence of a certain sound is dominant with respect to some sound alternation. Although successful in describing some important sound alternations in the dialect variation, it examines only the 10 most frequent sound alternations without testing patterns of variation between different sound correspondences.

Shackleton (2007) applies principal component analysis (PCA) to a group of self constructed articulation-based features. All segments found in the data are translated into vectors of numerical features and analyzed using PCA. Based on the component scores for features, different groups of varieties (in which a certain group of features is present) are identified. We note that the main drawback of this approach is the subjectivity of the feature selection and segment quantification.

Wieling and Nerbonne (2009) used a bipartite spectral graph partitioning method to simultaneously cluster dialect varieties and sound correspondences. Although promising, this method compares the pronunciation of every site only to the reference site, rather than comparing it to all other sites. Another drawback of this method is that it does not use any information on the frequencies of sound correspondences, but instead employs binary features to represent whether a certain correspondence is present at a certain site or not.

In this paper we present an approach that tries to overcome some of the problems described in the previous approaches. It proceeds from automatically aligned phonetic transcriptions, where pronunciations of every site are compared to the corresponding pronunciations for all other sites. Extracted sound correspondences are analyzed using the multi-way decomposition method PARAFAC. The method allows us to make generalizations over multi-way co-occurrence data, and to look simultaneously at the co-occurrence patterns of all sound correspondences found in the data set.

3 Data description

The data set used in this paper consists of phonetic transcriptions of 152 words collected at 197 sites evenly distributed all over Bulgaria. It is part of the project *Buldialect – Measuring Linguistic unity and diversity in Europe.* Phonetic transcriptions include various diacritics and suprasegmentals, making the total number of unique phones in the data set 95: 43 vowels and 52 consonants.[1] The sign for primary stress is moved to a corresponding vowel, so that there is a distinction between stressed and unstressed vowels. Vowels are also marked for their length. Sonorants /r/ and /l/ have a mark for syllabicity and for stress in case they are syllabic. Here we list all phones present in the data set:
ˈɑ, e, i, ˈe, ə, ˈɛ, ɤ, ˈɒ, ɑ, ɪ, o, ˈo, u, ˈɑː, ʊ, ˈuː, ˈɤ, ˈə, ˈa, ˈi, ˈɪ, ˈeː, ɛ, ˈɔ, ˈʌ, ˈiː, ˈu, eː, ɨ, ˈɨ, ˈoː, ˈɛː, ˈɤː, uː, ɑː, y, ˈaː, a, oː, ɤː, ˈʊ, ˈy, ˈɪː, j, ɡ, n, nʲ, ɟ, r, w, x, rʲ, h, ɕ, f, s, v̠, ç, ɸ, p, t͡ʃ, m, k, t͡ɕ, pʲ, c, l, lʲ, t, tʲ, ʃ, d, dʲ, ˈr̩, vʲ, d͡ʑ, ʒ, ʑ, t͡s, r̩, cʲ, z, sʲ, b, gʲ, mʲ, l̩, zʲ, ˈl̩, kʲ, bʲ, d͡ʑ, d͡ʑ, fʲ, ɰ

Each of the 152 words in the data set shows phonetic variation, with some words displaying more than one change. There are in total 39 different dialectal features that are represented in the data set, with each of the features being present in a similar number of words. For example, the reflexes of Old Bulgarian vowels that show dialect variation are represented with the same or nearly the same number of words. A more detailed description of all features can be found in Prokić et al. (2009). For all villages only one speaker was recorded. In the data set, for some villages there were multiple pronunciations of the same word. In this reasearch we have randomly picked only one per every village.

4 Methodology

4.1 Tensors

Co-occurrence data (such as the sound correspondences used in this research) are usually represented in the form of a *matrix*. This form is perfectly suited to represent two-way co-occurrence data, but for co-occurrence data beyond two modes, we need a more general representation. The generalization of a matrix is called a *tensor*. A tensor is able to encode co-occurrence data of any n modes. Figure 1 shows a graphical comparison of a matrix and a tensor with three modes – although a tensor can easily be generalized to more than three modes.

Tensor operations come with their own algebraic machinery. We refer the interested reader to Kolda and Bader (2009) for a thorough and insightful introduction to the subject.

[1] The data is publicly available and can be downloaded from `http://www.bultreebank.org/BulDialects/index.html`

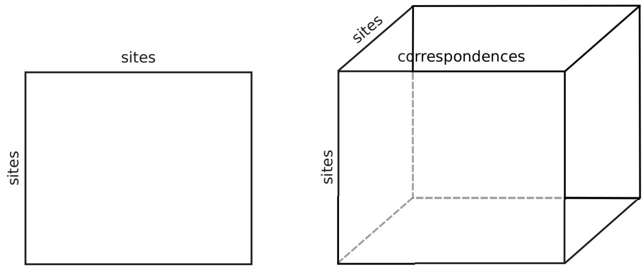

Figure 1: Matrix representation vs. tensor representation.

4.2 PARAFAC

In order to create a succinct and generalized model, the co-occurrence data are often analyzed with dimensionality reduction techniques. One of the best known dimensionality reduction techniques is principal component analysis (PCA, Pearson (1901)). PCA transforms the data into a new coordinate system, yielding the best possible fit in a least squares sense given a limited number of dimensions. Singular value decomposition (SVD) is the generalization of the eigenvalue decomposition used in PCA (Wall et al., 2003).

To be able to make generalizations among the three-way co-occurrence data, we apply a statistical dimensionality reduction technique called parallel factor analysis (PARAFAC, Harshman (1970); Carroll and Chang (1970)), a technique that has been sucessfully applied in areas such as psychology and bio-chemistry. PARAFAC is a multilinear analogue of SVD. The key idea is to minimize the sum of squares between the original tensor and the factorized model of the tensor. For the three mode case of a tensor $T \in \mathbb{R}^{D_1 \times D_2 \times D_3}$ this gives the objective function in 1, where k is the number of dimensions in the factorized model and $\circ$ denotes the outer product.

$$\min_{x_i \in \mathbb{R}^{D1}, y_i \in \mathbb{R}^{D2}, z_i \in \mathbb{R}^{D3}} \parallel T - \sum_{i=1}^{k} x_i \circ y_i \circ z_i \parallel_F^2 \tag{1}$$

The algorithm results in three matrices, indicating the loadings of each mode on the factorized dimensions. The model is represented graphically in Figures 2 and 3. Figure 2 visualizes the fact that the PARAFAC decomposition consists of the summation over the outer products of n (in this case three) vectors. Figure 3 represents the three resulting matrices that come out of the factorization, indicating the loadings of each mode on the factorized dimensions. We will be using the latter representation in our research.

Computationally, the PARAFAC model is fitted by applying an alternating least-squares algorithm. In each iteration, two of the modes are fixed and the third one is fitted in a least squares sense. This process is repeated until convergence.[2]

4.3 Sound correspondences

In order to detect the most important sound variation within Bulgarian dialects, we proceed from extracting all sound correspondences from the automatically aligned word transcriptions. All transcriptions were pairwise aligned using the Levenshtein algorithm (Levenshtein, 1965) as implemented in the program L04.[3] The Levenshtein algorithm is a dynamic programming algorithm used to measure the differences between two strings. The distance between two strings 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, namely 1. The algorithm is also directly used to align two sequences. An example showing two aligned pronunciations of the word вълна /vɤlna/ 'wool' is given in Figure 4.[4]

v	'ɤ	-	n	ɑ
v	'ɑ	l	n	ə

Figure 4: Example of two pairwise aligned word transcriptions.

From the aligned transcriptions for all words and all villages in the data set we first extracted

[2] The algorithm has been implemented in MATLAB, using the Tensor Toolbox for sparse tensor calculations (Bader and Kolda, 2009).

[3] http://www.let.rug.nl/kleiweg/L04

[4] For some pairs of transcriptions there are two or more possible alignments, i.e. alignments that have the same cost. In these cases we have randomly picked only one of them.

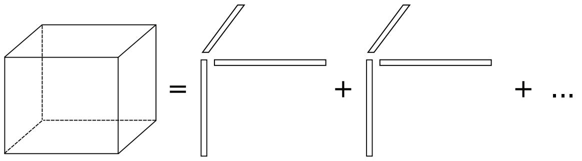

Figure 2: Graphical representation of PARAFAC as the sum of outer products.

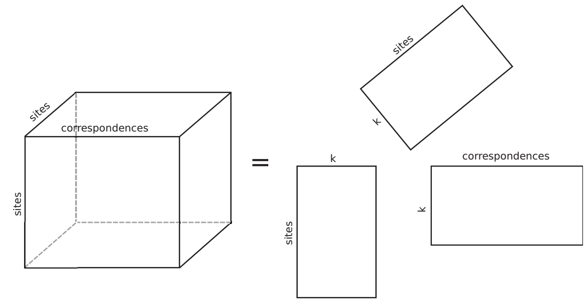

Figure 3: Graphical representation of the PARAFAC as three loadings matrices.

all corresponding non-identical sounds. For example, from the aligned transcriptions in Figure 4 we would extract the following sound pairs: [ˈɤ]-[ˈɑ], [-]-[l], [ɑ]-[ə]. The hyphen (‘-’) stands for a missing (i.e. inserted or deleted) sound, and in further analyses it is treated the same as any sound in the data set. For each pair of corresponding sounds from the data set we counted how often it appeared in the aligned transcriptions for each pair of villages separately. In total we extracted 907 sound correspondences and stored the information on each of them in a separate matrix. Every matrix records the distances between each two villages in the data set, measured as the number of times a certain phonetic alternation is recorded while comparing pronunciations from these sites.

Since we are interested in analyzing all sound correspondences simultaneously, we merged the information from all 907 two-mode matrices into a three-mode tensor $n \times n \times v$, where n represents the sites in the data set, and v represents the sound alternations. By arranging our data in a cube instead of a matrix, we are able to look into several sets of variables simultaneously. We are especially interested in the loadings for the third mode, that contains the values for the sound correspondences.

5 Results

In order to detect the most prominent sound correspondences we analyzed the three-mode tensor described in the previous section using a PARAFAC factorization with $k = 10$ dimensions. In Table 5 we present only the first five dimensions extracted by the algorithm. The final model fits 44% of the original data. The contribution of the first extracted dimension (dim1) to the final fit of the model is the largest – 23.81 per cent – while the next four dimensions contribute to the final fit with similar percentages: dim2 with 10.63 per cent, dim3 with 9.50 per cent, dim4 with 9.26 per cent, and dim5 with 9.09 per cent. Dimensions six to ten contribute in the range from 8.66 per cent to 6.98 per cent.

For every dimension we extracted the twenty sound correspondences with the highest scores. In the first dimension we find 11 pairs involving vowels and 9 involving consonant variation. The three sound correspondences with the highest scores are the [ɑ]-[ə], [o]-[u], and [e]-[i] alternations. This finding corresponds well with the traditional scholarly views on Bulgarian phonetics (Wood and Pettersson, 1988; Barnes, 2006) where we find that in unstressed syllables mid vowels [e] and [o] raise to neutralize with the high vowels [i] and [u]. The low vowel [a] raises to merge with [ə].

For every sound alternation we also check their geographical distribution. We do so by applying the following procedure. From the aligned pairs of transcriptions we extract corresponding pairs of sounds for every alternation. We count how many times each of the two sounds appears in the transcriptions for every village. Thus, for every pair of sound correspondences, we can create two maps that show the distribution of each of the sounds separately. On the map of Bulgaria these values

Table 1: First five dimensions for the sound correspondences.

dim1	dim2	dim3	dim4	dim5
[ɑ]-[ə]	[ə]-[ɤ]	[u]-[o]	[ɑ]-[ə]	[e]-[i]
[u]-[o]	[e]-[i]	[ɑ]-[ɤ]	[ə]-[ɤ]	[i]-[ˈe]
[e]-[i]	[ˈe]-[ˈɛ]	[ɑ]-[ə]	[ʊ]-[o]	[e]-[ə]
[-]-[j]	[-]-[j]	[ɤ]-[e]	[e]-[ə]	[r]-[rʲ]
[e]-[ˈe]	[ʃ]-[ɕ]	[e]-[ˈe]	[d]-[dʲ]	[d]-[dʲ]
[ʃ]-[ɕ]	[t͡ʃ]-[t͡ɕ]	[ˈe]-[ˈɛ]	[v]-[vʲ]	[ˈe]-[ˈɑ]
[t͡ʃ]-[t͡ɕ]	[ˈɑ]-[ˈɛ]	[-]-[j]	[n]-[nʲ]	[-]-[j]
[ˈe]-[ˈɛ]	[r]-[rʲ]	[ˈe]-[ˈɑ]	[-]-[j]	[ˈo]-[u]
[n]-[nʲ]	[l]-[lʲ]	[e]-[i]	[ˈe]-[ˈɛ]	[l]-[lʲ]
[ɑ]-[ɤ]	[e]-[ə]	[n]-[nʲ]	[l]-[lʲ]	[v]-[vʲ]
[e]-[ə]	[d]-[dʲ]	[r]-[rʲ]	[t]-[tʲ]	[u]-[o]
[ˈɑ]-[ˈɛ]	[n]-[nʲ]	[t͡ʃ]-[t͡ɕ]	[ˈe]-[ˈɑ]	[n]-[nʲ]
[ˈe]-[ˈɑ]	[u]-[ʊ]	[ˈɤ]-[ˈɑ]	[e]-[ˈe]	[-]-[v]
[d]-[dʲ]	[ˈɤ]-[ˈɔ]	[-]-[r]	[ʃ]-[ɕ]	[ˈɤ]-[ə]
[ɤ]-[e]	[ə]-[ˈɑ]	[ʃ]-[ɕ]	[t͡ʃ]-[t͡ɕ]	[u]-[ʊ]
[l]-[lʲ]	[ɤ]-[e]	[l]-[lʲ]	[r]-[rʲ]	[t͡ʃ]-[t͡ɕ]
[v]-[vʲ]	[ˈo]-[u]	[u]-[e]	[p]-[pʲ]	[ˈɑ]-[ˈɛ]
[r]-[rʲ]	[ʒ]-[ʑ]	[-]-[ˈɤ]	[ʒ]-[ʑ]	[ɑ]-[ˈɤ]
[ʒ]-[ʑ]	[i]-[ə]	[v]-[-]	[ə]-[ˈɑ]	[ə]-[ˈɑ]
[ˈɤ]-[ˈɔ]	[v]-[vʲ]	[ɑ]-[ˈɤ]	[e]-[i]	[b]-[bʲ]

are represented using a gradual color, which enables us to see not only the geographic distribution of a certain sound but also how regular it is in a given sound alternation. The highest scoring sites are coloured black and the lowest scoring sites are coloured white.

In Figure 5 we see the geographical distribution of the first three extracted correspondences. The first two alternations [ɑ]-[ə] and [o]-[u] have almost the same geographical distribution and divide the country into west and east. While in the west there is a clear presence of vowels [ɑ] and [o], in the east those vowels would be pronounced as [ə] and [u]. The division into east and west corresponds well with the so-called *jat* line, which is, according to traditional dialectologists (Stojkov, 2002) the main dialect border in Bulgaria. On the maps in Figure 5 we represent it with the black line that roughly divides Bulgaria into east and west. The third correspondence follows a slightly different pattern: mid vowel [e] is present not only west of the *jat* line, but also in the southern part of the country, in the region of Rodopi mountains. In the central and northeastern areas this sound is pronounced as high vowel [i]. For all three sound correspondences we see a clear two-way division of the country, with almost all sites being characterized by one of the two pronunciations, which, as we shall see later, is not always the case due to multiple reflections of some sounds at certain positions.

We also note that the distribution of the sound correspondences that involve soft consonants and their counterparts have the same east-west distribution (see Figure 6). In the first dimension we find the following consonants and their palatal counterparts [n], [d], [l], [v] and [r], but because of space limitations we show maps only for three correspondences. The east-west division also emerges with respect to the distribution of the [ɑ]-[ɤ] and [ˈe]-[ˈɑ] sounds.

Unlike the correspondences mentioned before, the [ʃ]-[ɕ], [t͡ʃ]-[t͡ɕ], and [ʒ]-[ʑ] pairs are defining the south part of the country as a separate zone. As shown on the maps in Figure 7, the southern part of the country (the region of Rodopi mountains) is characterized by a soft pronunciation of [ʃ], [t͡ʃ] and [ʒ]. In traditonal literature on Bulgarain dialectology (Stojkov, 2002), we also find that soft pronunciation of [ʃ], [t͡ʃ] and [ʒ] is one of the most important phonetic features of the varieties in the Rodopi zone. Based on the correspondences extracted in the first dimension, this area is also defined by the presence of the vowel [ˈɛ] in stressed syllables ([ˈe]-[ˈɛ] and [ˈɑ]-[ˈɛ] correspondences).

In some extracted correspondences, only one of the sounds has a geographically coherent distribution, like in the case of the [ɤ]-[e] pair where [e] is found in the west and south, while the [ɤ] sound is only sporadically present in the central region. This kind of asymmetrical distribution is also found with respect to the pair [ɑ]-[ɤ].

Most of the sound correspondences in the first dimension either divide the country along the *jat* line or separate the Rodopi area from the rest of the varieties. The only two exceptions are the [-]-[j] and [ˈɤ]-[ˈɔ] pairs. They both define the southwest area as a separate zone, while the northwest shares its pronunciation of the sound in question with the eastern part of the country.

We use the first 20 correspondences from the first dimension and perform *k-means* clustering in order to check which dialect areas would emerge based on this limited set of sound correspond-

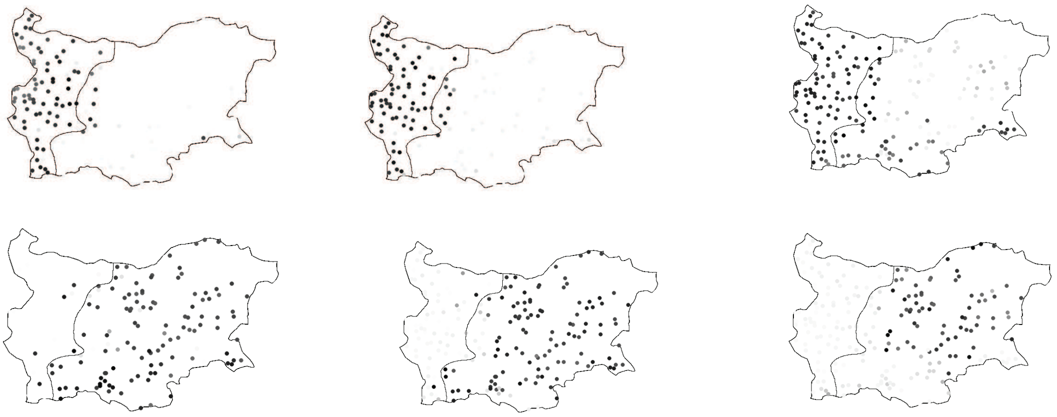

Figure 5: [ɑ]-[ə] (left), [o]-[u] (middle), [e]-[i] (right) sound correspondences.

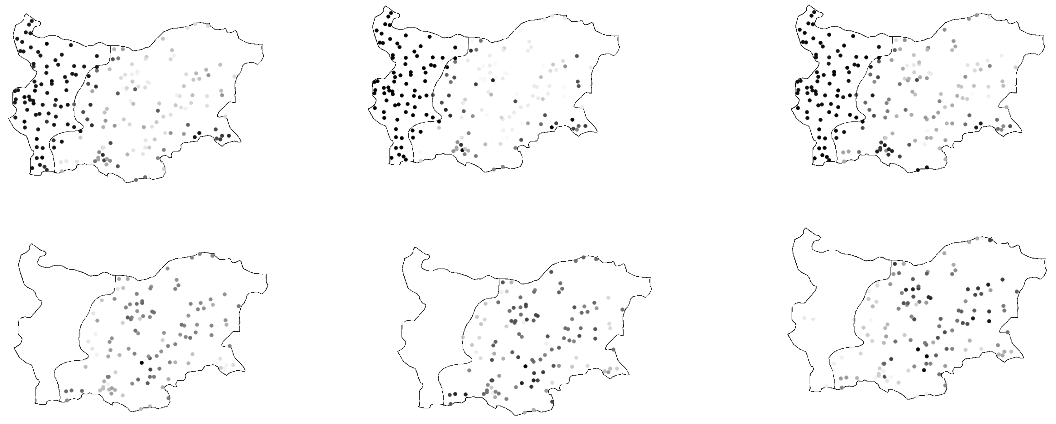

Figure 6: [d]-[d^j] (left), [v]-[v^j] (middle), [r]-[r^j] (right) sound correspondences.

ences. The results of the 2-way, 3-way and 4-way clustering are given in Figure 8.

In two-way clustering the algorithm detects an east-west split approximately along the *jat* line, slightly moved to the east. This fully corresponds to the traditional dialectology but also to the results obtained using Levenshtein algorithm on the whole data set where only east, west and south varieties could be asserted with great confidence (Prokić and Nerbonne, 2008). In Figure 9 we present the dialect divisions that we get if the distances between the sites are calculated using whole word transcriptions instead of only the 20 most prominent sound correspondences. We notice a high correspondence between the two analyses at the two- and three-level division. On the level of four and more groups, the two analyses start detecting different groups. In the analysis based on 20 sound correspondences, southern dialects are divided into smaller and smaller groups, while in the analysis based on the whole data set, the area in the west – near the Serbian border – emerges as the fourth group. This is no surprise, as the first 20 extracted correspondences do not contain any sounds typical only for this western area.

In order to compare two divisions of sites, we calculated the adjusted Rand index (Hubert and Arabie, 1985). The adjusted Rand index (ARI) is used in classification for comparing two different partitions of a finite set of objects. It is based on the Rand index (Rand, 1971), one of the most popular measures for comparing the degree to which partitions agree (in classification). Value 1 of the ARI indicates that two classifications match perfectly, while value 0 means that two partitions do not agree on any pair of points. For both two-level and three-level divisions of the sites the ARI for two classifications is 0.84. We also compared

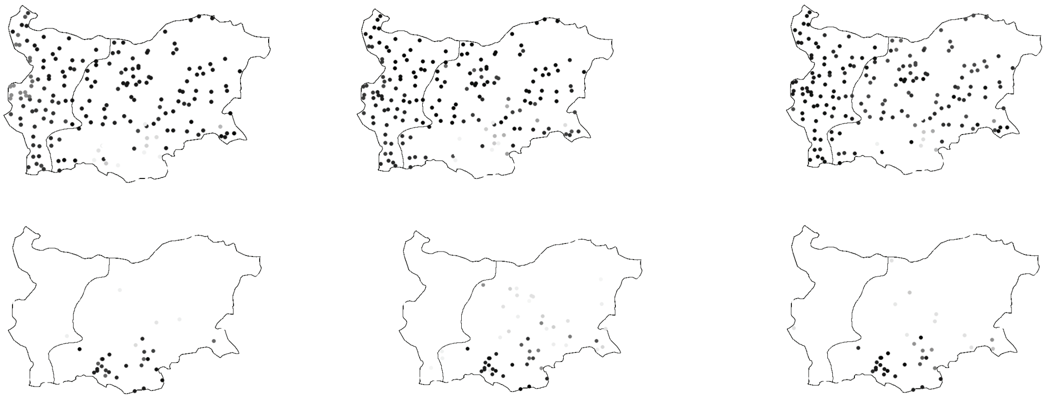

Figure 7: [ʃ]-[ɕ] (left), [t͡ʃ]-[t͡ɕ] (middle), [ʒ]-[ʑ] (right) sound correspondences.

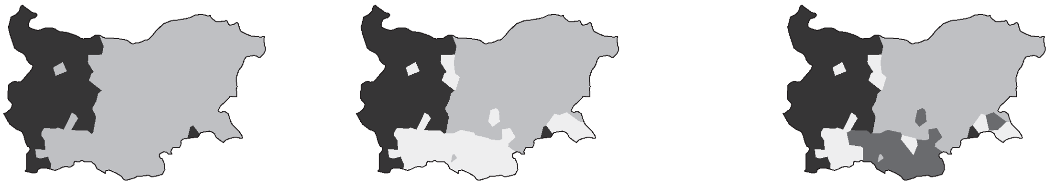

Figure 8: Dialect varieties detected by k-means clustering algorithm based on the first 20 sound correspondences in the first dimension.

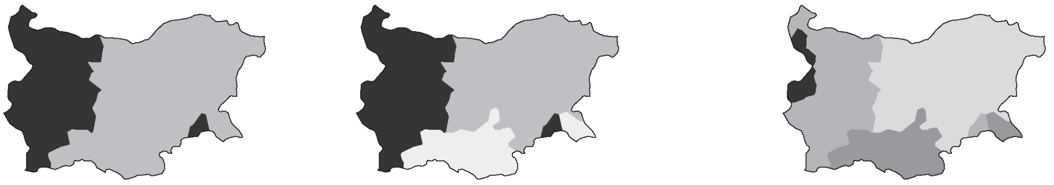

Figure 9: Dialect varieties detected by k-means clustering algorithm based on all word transcriptions.

both of the classifications to the classification of the sites done by Stojkov (2002). For the classification based on the first dimension extracted by PARAFAC, ARI is 0.73 for two-way and 0.64 for the three-way division. ARI score for the classification based on whole word transcriptions is 0.69 for two-way and 0.62 for three-way. As indicated by ARI the two classifications correspond with a high degree to each other, but to the traditional classification as well. We note that two-way classification based on the extracted sound correspondences corresponds higher to the traditional classification than classification that takes all sounds into account.

We conclude that the sound correspondences detected by PARAFAC form the linguistic basis of the two-way and three-way divisions of Bulgarian dialect area. Using the PARAFAC method we are able to detect that the most important sound changes on which two-way division is based are [o]-[u], [ɑ]-[ə] and palatal pronunciation of consonants. In the three-way division of sites done by *k-means*, the area in the south of the country appears as the third most important dialect zone. In the twenty investigated sound correspondences we find that the soft pronunciation of [ʃ],[t͡ʃ] and [ʒ] sounds is typical only for the varieties in this area. Apart from divisions that divide the country into west and east, including the southern varieties, we also detect sound correspondences whose distribution groups together western and southern areas.

We also analyzed in more depth sound correspondences extracted in other dimensions by the PARAFAC algorithm. Most of the correspondences found in the first dimension, also reappear in the following nine dimensions. Closer inspection of the language groups obtained using information from these dimensions show that eastern, western and southern varieties are the only three that are identified. No other dialect areas were detected based on the sound correspondences from these nine dimensions.

6 Conclusion

In this paper we have applied PARAFAC in the task of detecting the linguistic basis of dialect phonetic variation. The distances between varieties were expressed as a numerical vector that records information on all sound correspondences found in the data set. Using PARAFAC we were able to extract the most important sound correspondences. Based on the 20 most important sound correspondences we performed clustering of all sites in the data set and were able to detect three groups of sites. As found in traditional literature on Bulgarian dialects, these three dialects are the main dialect groups in Bulgaria. Using the aggregate approach on the same data set, the same three dialects were the only groups in the data that could be asserted with high confidence. We conclude that this approach is successful in extracting underlying linguistic structure in dialect variation, while at the same time overcoming some of the problems found in the earlier approaches to this problem.

In future work sounds in the data set could be defined in a more sophisticated way, using some kind of feature representation. Also, the role of stress should be examined in more depth, since there are different patterns of change in stressed in unstressed syllables. We would also like to extend the method and examine more than just two sound correspondences at a time.

References

Brett W. Bader and Tamara G. Kolda. 2009. Matlab tensor toolbox version 2.3. `http://csmr.ca.sandia.gov/~tgkolda/TensorToolbox/`, July.

Jonathan Barnes. 2006. *Strength and Weakness at the Interface: Positional Neutralization in Phonetics and Phonology*. Walter de Gruyter GmbH, Berlin.

J. Douglas Carroll and Jih-Jie Chang. 1970. Analysis of individual differences in multidimensional scaling via an n-way generalization of ”eckart-young” decomposition. *Psychometrika*, 35:283–319.

Richard A. Harshman. 1970. Foundations of the parafac procedure: models and conditions for an ”explanatory” multi-mode factor analysis. In *UCLA Working Papers in Phonetics*, volume 16, pages 1–84, Los Angeles. University of California.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of Classification*, 2:193–218.

Tamara G. Kolda and Brett W. Bader. 2009. Tensor decompositions and applications. *SIAM Review*, 51(3), September.

Vladimir Levenshtein. 1965. Binary codes capable of correcting deletions, insertions and reversals. *Doklady Akademii Nauk SSSR*, 163:845–848.

John Nerbonne. 2006. Identifying linguistic structure in aggregate comparison. *Literary and Linguistic Computing*, 21(4):463–476.

Karl Pearson. 1901. On lines and planes of closest fit to systems of points in space. *Philosophical Magazine*, 2(6):559–572.

Jelena Prokić and John Nerbonne. 2008. Recognizing groups among dialects. *International Journal of Humanities and Arts Computing*, Special Issue on *Language Variation* ed. by John Nerbonne, Charlotte Gooskens, Sebastian Kürschner, and Renée van Bezooijen:153–172.

Jelena Prokić, John Nerbonne, Vladimir Zhobov, Petya Osenova, Krili Simov, Thomas Zastrow, and Erhard Hinrichs. 2009. The Computational Analysis of Bulgarian Dialect Pronunciation. *Serdica Journal of Computing*, 3:269–298.

Jelena Prokić. 2007. Identifying linguistic structure in a quantitative analysis of dialect pronunciation. In *Proceedings of the ACL 2007 Student Research Workshop*, pages 61–66.

William M. Rand. 1971. Objective criteria for the evaluation of clustering methods. *Journal of American Statistical Association*, 66(336):846–850, December.

Robert G. Shackleton. 2007. Phonetic variation in the traditional English dialects. *Journal of English Linguistics*, 35(1):30–102.

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

Michael E. Wall, Andreas Rechtsteiner, and Luis M. Rocha, 2003. *Singular Value Decomposition and Principal Component Analysis*, chapter 5, pages 91–109. Kluwer, Norwell, MA, Mar.

Martijn Wieling and John Nerbonne. 2009. Bipartite spectral graph partitioning to co-cluster varieties and sound correspondences in dialectology. In *Text Graphs 4, Workshop at the 47th Meeting of the Association for Computational Linguistics*, pages 14–22.

Sidney A. J. Wood and Thore Pettersson. 1988. Vowel reduction in Bulgarian: the phonetic data and model experiments. *Folia Linguistica*, 22(3-4):239–262.