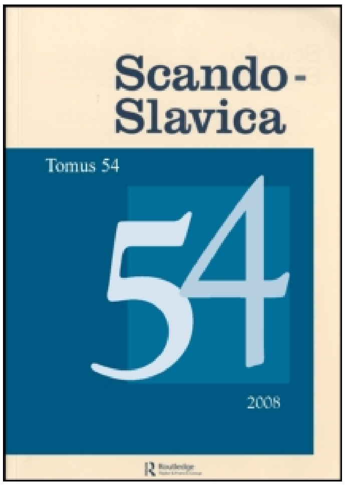


Scando-Slavica
Publication details, including instructions for authors and subscription information:
http://www.informaworld.com/smpp/title~content=t716100751

Quantitative and Traditional Classifications of Bulgarian Dialects Compared

Peter Houtzagers[a]; John Nerbonne[a]; Jelena Prokić[a]
[a] Faculty of Arts, University of Groningen, The Netherlands

Online publication date: 02 December 2010

To cite this Article Houtzagers, Peter , Nerbonne, John and Prokić, Jelena(2010) 'Quantitative and Traditional Classifications of Bulgarian Dialects Compared', Scando-Slavica, 56: 2, 163 — 188

To link to this Article: DOI: 10.1080/00806765.2010.530801

URL: http://dx.doi.org/10.1080/00806765.2010.530801

Scando-Slavica 56:2 (2010), 163–188.

Routledge
Taylor & Francis Group

Quantitative and Traditional Classifications of Bulgarian Dialects Compared

Peter Houtzagers, John Nerbonne, Jelena Prokić
Faculty of Arts, University of Groningen, P.O.B. 716, NL-9700 AS Groningen, The Netherlands. h.p.houtzagers@rug.nl, j.nerbonne@rug.nl, j.prokic@rug.nl

Abstract
Dialect classification is a classical problem in traditional dialectology. In the course of the last few decades, several quantitative approaches have been suggested as solutions for this problem, one of which uses "Levenshtein distance" for measuring linguistic distances between dialects. In the present paper we shall introduce the Levenshtein algorithm as well as two methods with which the results of the measuring can be analyzed, viz. multidimensional scaling and clustering. Then we shall apply these methods to the Bulgarian language area and present a quantitative classification of Bulgarian dialects. Finally, we shall compare the classification obtained to the most widely accepted traditional Bulgarian dialect map, analyze the similarities and differences and evaluate our method.

Keywords: Bulgarian, dialectology, dialect geography, linguistic geography, Levenshtein distance, multidimensional scaling, clustering, dialectometry.

1. Introduction

The division of language areas into dialect groups is a well-known problem in traditional dialectology. The most wide-spread way of doing this is by drawing dialect borders along isoglosses or bundles of isoglosses that are considered more important than others. The choice of which isogloss(es) to use is, of course, subjective and so are the decisions in cases where isoglosses do not coincide or do not provide neat bisections of the language area. Other traditional methods, such as division based on the phoneme inventory and on speakers' judgments, also have obvious flaws. The former often groups together large numbers of otherwise heterogeneous dialects, the latter can be argued to be subjective (albeit in another sense than the isogloss method) and to contain non-linguistic elements. Since the 1970s a growing number of scholars have introduced various quantitative approaches to the determina-

DOI: 10.1080/00806765.2010.530801

tion of dialect borders, all of which have to do with counting the differences between dialects and calculating "linguistic distances" between them. The approaches differ in many ways, e.g. the selection of the data, the definition of "difference", the way in which differences are counted and the way in which the linguistic distance is calculated. Most of the techniques used in quantitative dialectology examine pairs of varieties and count how many specific items are the same (or different). The fraction of different items is then interpreted as a linguistic distance between the two sites. See Nerbonne and Heeringa (2009) and the papers in Nerbonne and Kretzschmar (2006) for recent overviews of these quantitative approaches.

One of these computational methods, introduced in linguistics by Kessler (1995), uses the so-called "Levenshtein distance". Two of the present authors (Nerbonne and Prokić) have been working on applying Levenshtein distance to Bulgarian dialects since 2006.[1] Their work resulted in a number of possible classifications that show (mostly minor) differences according to the clustering technique and the number of classes chosen. Because the application of the Levenshtein distance to Bulgarian was partly a test of the method itself, it is relevant to compare the borders obtained with the borders on the dialect maps produced by traditional methods and to try to explain both the similarities and the differences. This was done in collaboration with the third author (Houtzagers). In this article we shall first present the data that were used for applying the Levenshtein method to Bulgarian dialects and briefly introduce the Levenshtein method itself. Then we shall present the reader with the classification of Bulgarian dialects obtained by using this method. Finally we shall compare this classification with the most authoritative map produced by traditional Bulgarian dialectology.

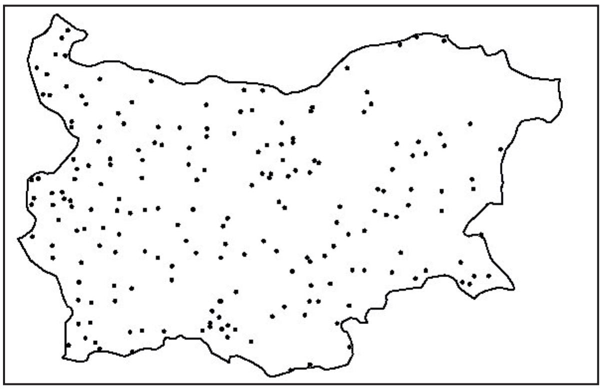
The project as a whole concentrates on phonetics and lexicon. In this article, we have restricted ourselves to phonetics. Only dialects spoken in the Republic of Bulgaria are taken into consideration.

2. The Data

The data set used in this paper consists of phonetic transcriptions of 156 words collected from 197 sites all over Bulgaria (see Fig. 1).

[1] This was part of a dialectometry project called *Buldialect – Measuring Linguistic Unity and Diversity in Europe*, in which the universities of Groningen (The Netherlands) and Tübingen (Germany) and the Bulgarian Academy of Sciences participated and that was financially made possible by the Volkswagen Stiftung.

Fig. 1. Distribution of the Selected 197 Sites



The main source of the data was the *Archive of the Ideographic Dialect Dictionary of Bulgarian* of the University of Sofia that was initiated during the 1950s by Professor Stojko Stojkov, at that time the leading expert in the field of Bulgarian dialectology.[2] Part of the project was the digitalization of the selected data.

The main criterion for word selection in the *Buldialect* project was availability: the words in the data set are frequent words that are collected from all, or almost all of the 197 sites. Only words which were expected to show some degree of variation were included. Another important criterion for word selection was the balance between various phonetic features present in the data set. For example, the reflexes of the Old Bulgarian[3] vowels that show dialectal variation are represented with the same or nearly the same number of words. Below we present a list of the 39 different dialectal features that are repre-

[2] The 197 sites were selected in such a way that the geographical distribution was as even as possible. From the dotless areas in Fig. 1 no data were available. These empty areas are essentially the same as those in the *Bălgarski dialekten atlas* (see the introductory part and the first map of each individual volume of BDA), which also appeared under Stojkov's leadership. Stojkov chose not to include villages with a population that was either dialectally heterogeneous or that had migrated to its present dwelling-place from other parts of Bulgaria. The data for the *Archive* were collected according to the same principles as those for BDA. For instance, the informants used were the oldest inhabitants of the village in question under the strict condition that they were born locally, with a preference for women. For more details concerning the data collection see BDA I, 8–9. A description of the Archive can be found in Stojkov and Mladenov 1969.

[3] Henceforth "OBg". "Standard Bulgarian" will be abbreviated as "StBg".

sented in the chosen 156 words. A more detailed description of the features can be found in Prokić et al. (2009).[4]

1. Reflexes of *jat*, e.g. [xlʲap][5] vs. [xlep] ‘bread’
2. Reflexes of the sequences **ja, *ča,* etc. Example: [ofˈtʃar] vs. [ofˈtʃer] ‘shepherd’
3. Presence or absence of initial prothetic [j], e.g. [ˈagne] vs. [ˈjagne] ‘lamb’
4. Presence or absence of [j] before front vowels, e.g. [koˈje] vs. [koˈe] ‘which’ (neuter singular)
5. Elision or no elision of [j], e.g. [ˈneja] vs. [ˈnea] ‘she’ (accusative)
6. Reflexes of the back nasalized vowel, e.g. [mɤʃ] vs. [maʃ] vs. [muʃ] ‘man’
7. Reflexes of the front nasalized vowel, e.g. [zet] vs. [zʲɔt] vs. [zʲɤt] ‘son-in-law’
8. Reflexes of back *jer*, e.g. [dɤʃt] vs. [doʃt] vs. [daʃt] ‘rain’
9. Reflexes of the front *jer,* e.g. [ˈtɤŋko] vs. [ˈteŋko] vs. [ˈtʲɔŋko] ‘thin’ (neuter)
10. Choice of the vowel inserted between the two last consonants in **vjatr* ‘wind’ and **ogn* ‘fire’, e.g. [ˈvjatɤr] vs. [ˈveter]
11. Presence vs. absence of vowel reduction in unstressed syllables results of vowel reduction if present, e.g. [ˈpepel] vs. [ˈpepil] vs. [ˈpepʲɤl] ‘ashes’
12. Reflexes of *jery* (OBg **y*), e.g. [eˈzik] vs. [eˈzɨk] ‘tongue’
13. Rounding or no rounding of vowels in e.g. [ʒif] vs. [ʒuf] ‘alive’
14. Unrounding or no unrounding of vowels in e.g. [klʲutʃ] vs. [klitʃ] ‘key’
15. Presence or absence of alternation [o]-[e] in e.g. [ˈselo] ‘village’ vs. [vɤˈʒe] ‘rope’
16. Presence or absence of vowel elision, e.g. [ˈrapta] vs. [ˈrabota] ‘work’
17. Presence or absence of change by analogy in such cases as [ˈdole] from [ˈdolu] ‘down’ (by analogy with [ˈgore] ‘up’)
18. Reflexes of syllabic liquids, e.g. [vɤlk] vs. [vlɤk] vs. [vl̩k], vs. [vuk] ‘wolf’

[4] Some of the characteristics mentioned in the list are binary, others are not. In the case of nonbinary characteristics sometimes only two examples have been given but the number of possible variants is mostly greater than two.

[5] Between square brackets we shall use the IPA. When referring to OBg forms we shall use a notation in Latin script that is widely accepted in Slavic historical linguistics. StBg forms will be spelled according to the accepted orthography and Scando-Slavica’s transliteration.

19. Reflexes of **tj*, **dj* e.g. [ˈleʃta] vs. [ˈleʃtʃa] vs. [ˈletʃa] vs. [ˈletsa] 'lentil'
20. The fate of the original initial cluster **čr* + following vowel **ь* or **ě* in words like [tʃerˈven] vs. [tsɤrˈven] 'red'
21. Presence or absence of epenthetic [l] in e.g. [zeˈmlʲa] vs. [zeˈmʲa] 'land'
22. Presence or absence of the two voiced affricates [dz] and [dʒ], e.g. [dzveˈzda] vs. [zveˈzda] 'star'
23. Palatalization or not of [l] and [n] in [poneˈdelʲnik] vs. [poneˈdelnik] 'Monday, [ˈagnʲe] vs. [ˈagne] 'lamb'
24. Results of palatalization of **st*, **zd* in words corresponding to StBg [ˈgosti] 'guests' [ˈgrozde] 'grapes', e.g. [ˈgosje], [ˈgojse]; [ˈgrosje], [ˈgrojze]
25. Presence or absence of simplification of the clusters **str*, **zdr*, e.g. [seˈsra] vs. [seˈstra] 'sister'
26. Presence or absence of epenthesis of [t], [d] in the clusters **sr*, **zr*, e.g. [ˈstrʲada] vs. [ˈsrʲada] 'Wednesday'
27. Existence or nonexistence in the dialect of a voiceless velar fricative, e.g. [xlʲap] vs. [lʲap] 'bread'
28. Existence or nonexistence in the dialect of a voiceless labiodental fricative, e.g. [ˈfurna] vs. [ˈxurna], [ˈvurna], etc. 'oven'
29. Preservation or loss of **v* before rounded vowels, e.g. [vol] vs. [ol] 'ox'
30. Presence or absence of prothetic [v] before rounded vowels, e.g. [ˈvogɤn] vs. [ˈogɤn] 'fire'
31. Devoicing or not of voiced obstruents in certain positions, e.g. [ofˈtsa] vs. [ovˈtsa] 'sheep', [dop] vs. [dʒob] 'pocket'
32. The form of the preposition **vъ* and the prefix **vъ-*, e.g. [ˈvlizam] vs. [uˈlizam] 'enter'
33. Various assimilations and dissimilations, e.g. [ofˈtsa] vs. [osˈtsa] 'sheep'
34. Nonsystematic changes in individual words, e.g. [ˈvetʃe] v.s [ˈvetse] 'already'
35. Morphophonemic alternations or suffixes connected with the formation of secondary imperfective verbs, e.g. [ˈvlizam] vs. [ˈvlʲazam] vs. [ˈvlʲavam] 'enter' (imperfectives corresponding with perfective [ˈvlʲaza])
36. Form of certain grammatical endings, such as that of the first person plural in all tenses, e.g. [ˈbʲaxme] vs. [ˈbexmo] 'be' (past tense first person plural)
37. Choice of the suffix in certain nouns that originally belonged to the *n*-stem paradigm, e.g. [ˈkamɤk] vs. [ˈkamik] vs. [ˈkamen] 'stone'

38. Form of certain words, among others the words for ‘you’ (plural or polite) and ‘then’, e.g. [ˈvie] vs. [vi] vs. [ve] ‘you’, [toˈgava] vs. [toˈgas] vs. [teˈgaj] ‘then’
39. Stress in certain bisyllabic forms, e.g. [ˈvino] vs. [viˈno] ‘wine’

A full list of the words used may be found in the appendix.

3. Levenshtein Distance

We use Levenshtein distance as a measure of pronunciation difference in this paper. It is a natural extension of the basic technique mentioned in the introduction, where one counts points of difference in a fixed inventory of linguistic items in order to gauge linguistic distance. Levenshtein distance is used in a variety of scientific and technical fields in order to measure the differences between two sequences, or strings (Levenshtein 1965). In its simplest version – which is applied in this paper – the distance between two strings is the smallest number of insertions, deletions, and substitutions needed to transform one string into the other. For example, in order to align the two word transcriptions presented in Fig. 2, we would need four operations: [r] would have to be replaced with [r^j], [e] would have to be deleted, [n^j] would have to be replaced with [n] and [e] with [i]. Every operation is assigned the same value, namely 1. This means that the distance between these two pronunciations is 4. Every sequence is represented as a series of phones which are not further defined. As a consequence, the pair [r - r^j] counts as different to the same degree as the pair [e - i]. Stress is represented not suprasegmentally, but rather as a feature on vowels so that a stressed [i] is regarded as different from an unstressed one. We noted in the introduction that most quantitative dialectology assays the linguistic distance between varieties by counting points of difference. Levenshtein distance generalizes on the simple counting of mis-matching segments by allowing for insertions and deletions (linguistic epentheses and elisions).

The transcriptions shown in Fig. 2 were aligned based on the following principles: (a) a vowel can be aligned only with another vowel; (b) a consonant can be aligned with another consonant, a sonorant or one of the semivowels [j] and [w].

Fig. 2. Alignment of Two Word Transcriptions

s	ˈi	r	e	n^j	e
s	ˈi	r^j	—	n	i
		1	1	1	1

The Levenshtein distance between the two words is 1+1+1+1=4.

The procedure is admittedly rough, but it has been evaluated and shown to work well given a large amount of data (more than 60 words). The evaluations have concerned comparisons of results with expert opinion, meta-analysis demonstrating consistency, and a comparison to dialect speakers judgments of similarity (for an overview of evaluation, see Nerbonne and Heeringa 2009, 561–563). It is naturally possible to introduce more phonetic sensitivity to the procedure by employing a segment distance table, which allows the linguist to specify variable costs that may be incurred per operation depending on the phonetic or phonological segments involved, and this is important when seeking to detect sound correspondences. Heeringa (2004, 27–120) experiments extensively with several feature systems, both phonetically and phonologically inspired and even with spectrograms (acoustics). However, he concludes (p. 186) that using phonetically more sensitive segment representations does not improve the results obtained using simple (phone) representations. It is important note that Heeringa's evaluation concerned the aggregate analysis of dialect differences, in which one attempts to measure how dissimilar one entire variety is to another (but crucially without attempting to ascertain which differences are most important). This leads us to prefer the simpler comparison wherever the focus is on the properties of entire varieties, such as their classification (Heeringa et al. 2006). The simple phone representations have the further advantage of making fewer assumptions about the nature of phonetic similarity, e.g. the assumptions implicit in different feature systems.[6]

[6] One anonymous reviewer correctly noted that some differences are "systemically more significant than others," but note that our goal is to characterize how similar (or dissimilar) two varieties are phonetically, and so we ignore the systemic, or phonological perspective. Maguire (2008) develops a quantitative procedure that is sensitive to systemic status, but his procedure requires a substantial set of phonemic comparisons which were not at our disposal. It would be interesting to explore this as well.

The advice of statisticians is to infer properties of populations based on large representative samples. Since we compare 156 words with an average length of a bit more than four segments, we compare about 600 segments per site pair. This is a large sample compared to non-computational studies. Since we begin from dialect atlas material, we cannot claim to avoid subjective choices entirely. But we argue in Section 2 (above) that the data contains many examples of the geographically variable sounds which earlier scholarship discusses, and therefore reflects a range of scholarly views on Bulgarian dialects, and not merely our own. Note, too, that many of the 600 comparisons involve material that happens to be in words chosen for other reasons. So while бели /beli/ 'white-PL' was chosen because it illustrates the [e/æ/ etc.] variation (see section 5.1 under (b), below), the sounds left and right also form part of the comparison, adding an element of randomness to the sample that should improve its representativeness. We claim this improves on the manual selection of a few isoglosses.

We align all the word transcriptions in the way described above, calculating the distances between each pair of related words in each pair of sites.[7] This calculation yields distances between each pair of sites in the following way. The distance between two sites is the mean of all word distances calculated for those two sites. The final result is a distance matrix that contains the distances between each two sites in the data set. This matrix is further analyzed using multi-dimensional scaling and hierarchical clustering, which will be described in the next section.

4. Classification of Bulgarian Dialects Using Levenshtein Distance

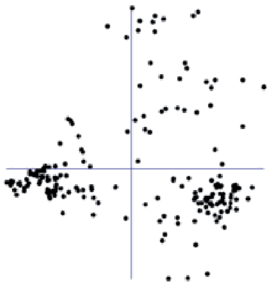
The result of measuring the pronunciation distance of each pair of words in the 156-word sample is a set of more than 19,000 varietal distances – one for each pair of varieties. We organize these in a site × site table, but one which is too large to be appreciated via visual inspection. We turn therefore to statistical techniques for analyzing the distances. Multidimensional scaling

[7] We note that certain imperfectly matching correspondences will occur often, e.g. [ε] and [æ]. Wherever these mismatching correspondences occur within the set of 156 words in the sample, they will contribute to the aggregate difference. Thus frequent mismatching correspondences contribute disproportionately in comparison to infrequent ones. This results in an implicit "weighting" reflecting the frequency with which segments are encountered, but we note that it is an open question whether this in turn reflects dialect differences well (as perceived by dialect speakers).

(henceforth also "MDS") is a dimension-reduction technique which takes as input a set of points and (abstract, e.g. linguistic) distances among them (Kruskal and Wish 1978). MDS then proposes a small set of dimensions and coordinates for each input point so that the (abstract) distances are approximated as nearly as possible.

Unlike clustering (see below), MDS results are stable so they form an interesting standard against which we may examine the results of clustering. Since in the Bulgarian data, 92.4% of the variation is explained by the first two dimensions, we may conveniently examine the two-dimensional MDS reconstruals (scatterplots) in order to visualize the data, in particular looking for separation of the clusters (Legendre and Legendre 1998). In MDS plots, dissimilar objects are plotted far from each other and similar objects are close. In Fig. 3 an example of MDS is given: each dot represents one of the 197 sites studied in this project. The figure shows clusters of dots alternating with relatively "white" areas, which means that the linguistic distances from a given dialect to its linguistically closest neighbours vary greatly.

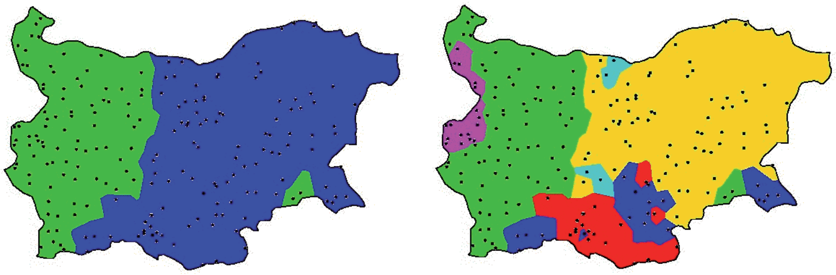
Fig. 3. Example of MDS



Clustering is the process of partitioning a set of objects into groups (Manning and Schütze 1999). The goal of cluster analysis is to find structure in the data by detecting and grouping together similar objects. In dialectometry, cluster analysis is frequently used to analyze the distances between sites and detect dialect regions by grouping the sites that share linguistic features. In this research, we proceed from the distance matrix obtained using Levenshtein algorithm described in the previous section and analyze it with a hierarchical clustering algorithm called WPGMA (Weighted Pair Group Method using Arithmetic Averages). A detailed description of this algorithm and some alternative clustering techniques, tested on the same data set used in this research, can be found in Prokić and Nerbonne 2009. All hierarchical clustering algorithms proceed from a distance matrix, repeatedly choosing the two closest elements and fusing them. They differ in the way in which distances are recalculated from the newly fused elements to the others. In this research we present the results of the two-way and six-way divisions of the data done by WPGMA algorithm. We chose a two- and a six-way division in order to be able to compare the results obtained with those of the traditional division of the Bulgarian dialect area. As we shall see below (section 5), the traditional division distinguishes two main dialect groups, that can be further divided into six subgroups.

The results of the WPGMA algorithm can be seen in Fig. 4. The main division of the sites is into eastern and western varieties. The six-way division of the sites shows four main dialect areas: eastern, western, southern (in the area of Rhodope Mountains) and a transitional zone at the border with Serbia. The southern group of dialects is further divided into smaller groups, which indicates that this group of dialects is more heterogeneous than the other three.

Fig. 4. Two- and Six-Way Division Using the WPGMA Algorithm



Previous study (Prokić and Nerbonne 2009) has shown that if other clustering and statistical techniques are applied on the same data set, the same eastern, western and southern groups are identified that we see in Fig. 4. However, some well established hierarchical clustering techniques do not distinguish the transitional zone at the border with Serbia. An example is the technique called UPGMA (Unweighted Pair Group Method using Arithmetic Averages).

In this paper we will investigate the reasons for the different results obtained by quantitative techniques when compared to traditional scholarship.

If such differences occur, it is of course relevant to check whether the distribution of the phonetic features present in our data set corresponds well with the phonetic features responsible for the traditional dialect divisions. Another potential explanation for differences in classification is the possibility that some of the dialect areas defined in traditional atlases are not strongly founded in the linguistic data, but rather reflect the knowledge of cultural and historical differences. There is also a possibility that differences are due to the techniques we use. Because similarities between traditional and quantitative classifications should not be taken for granted, we shall also pay attention to notable similarities between the two.

5. Comparison with Traditional Classification

5.1. Stojkov's Classification

By far the most widely known and most authoritative classification of Bulgarian dialects is the one published by Stojkov (1968, 291)[8] and reproduced in Fig. 5.[9]

The starting point for Stojkov's classification is the reflex of the OBg phoneme **ě* (*jat*).

(a) In a non-palatal environment, i.e. not followed by a palatal or palatalized consonant or a syllable containing a front vowel, the reflex of *jat* west of the thick line is [e]. East of the thick line it is [a] or a low variant of [e]. If the reflex of *jat* is [a] or a very low variant of [e], a preceding consonant is usually palatalized. Example: [bel] 'white' (masculine singular) vs. [bʲal], [bʲæl] or [bɛl]. This isogloss more or less neatly splits up the Bulgarian language area in two continuous parts and it almost fully coincides with the dialect border shown on the map.

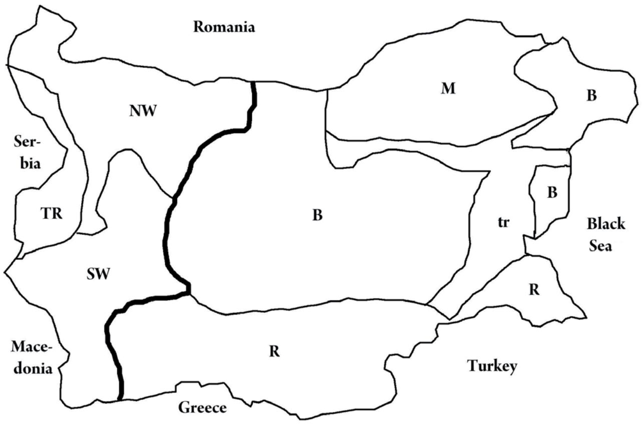
(b) East of the thick line there is a division based on the reflex of *jat* in a palatal environment: in the south and southeast (abbreviated R from *Rupski govori* 'Rupian dialects') the reflex is [a], [æ] or [ɛ], in the north

[8] The map in this second edition of *Bălgarska dialektologija* is essentially the same as the one in the first edition (Stojkov 1962) with some refinements as to the subdivision of the eastern Bulgarian dialects. It appears unaltered in the 1993 and 2002 editions.

[9] Stojkov was well aware how problematic it was to classify dialects in a satisfactory way and even called it "an extraordinarily difficult and almost impossible task" (1993, 81).

it is [e]. Example: [ˈbʲali], [ˈbʲæli] or [ˈbɛli] 'white' (plural) vs. [ˈbeli]. This second *jat* isogloss divides the Bulgarian language area into several discontinuous parts. As a matter of fact, the northeast corner has the same reflex of *jat* as the south. The boundary between the southern and the non-southern eastern dialects shown on the map therefore represents a simplification of the actual linguistic facts.

Fig. 5. Traditional Division of Bulgarian Dialects (after Stojkov)



Abbreviations: NW – northwest, SW – southwest, TR – transitional (between Bulgarian and Serbian), B – Balkan dialects, M – Moesian, R – Rupian, tr – transitional (between Balkan, Moesian and Rupian dialects).

The western dialects are subdivided into northwest and southwest (NW and SW), according to the reflex of the OBg back nasal *ǫ > [ɤ] or [a], e.g. [zɤp] vs. [zap] 'tooth' (StBg *zăb*).[10] There is also a transitional (TR) area with *ǫ > [u] and several other traits that make it transitional to Serbian.

The northern part of the east Bulgarian area is subdivided into Balkan and Moesian dialects (B and M), the former of which are discontinuous. About the distinction between Balkan and Moesian dialects see section 5.2.3 below. There are also transitional dialects between Balkan and Moesian and between Balkan and Rupian dialects (tr). The southern part of the east Bulgarian area is formed by the Rupian dialects (R; see above).

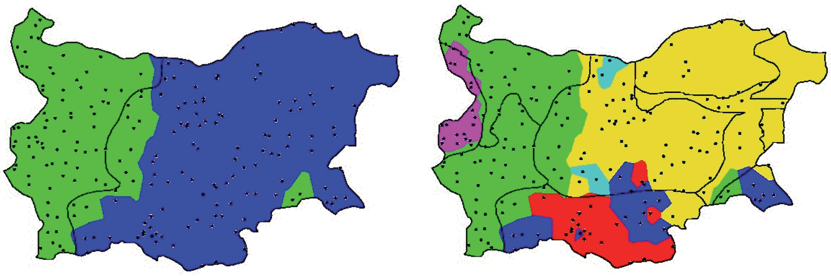
[10] The vowel spelled *ă* is pronounced [ə]. The word-final consonant (spelled *b*) is unvoiced.

5.2. Comparison

If one does not count the transitional dialects in the east, Stojkov's classification distinguishes six dialect groups. Moreover, it attaches greater importance to the split between east and west than to the other divisions of the Bulgarian language area. Therefore it can easily be compared to the two- and six-way quantitative divisions presented above. In Fig. 6 we projected Stojkov's boundaries (Fig. 5) on the quantitative maps (Fig. 4). The areas obtained by using the Levenshtein distance can be recognized by the colours (in the printed version of this journal: different types of shading).

Fig. 6 shows both similarities and differences between Stojkov's map and the quantitative maps. Differences are present wherever the transitions between different colours/types of shading do not coincide with the black lines. In the remainder of section 5 we shall discuss and try to explain the most important similarities and differences.[11]

Fig. 6. Quantitative and Traditional Boundaries



Quantitative 2- and 6-way classifications are shown as colours (as shading in the paper version of this article) with traditional boundaries projected on them as lines.

5.2.1. West vs. East

The division between west and east, which was the starting point for Stojkov's classification, roughly agrees on both maps. Moreover, if we compare the WPGMA maps given in Fig. 4 with the other quantitative maps, we see that the east-west boundary constantly appears at the highest (two-way) clustering level and stays identical on every map, independently of the type of

[11] We have restricted ourselves to differences involving more than three sites.

clustering. However, the middle of the east-west boundary on the quantitative maps runs east of Stojkov's. Two questions arise:

(a) How can we explain that a boundary so similar to Stojkov's _jat_-boundary, which in principle is based on a single isogloss, shows so much stability on the quantitative maps?

(b) How can we explain the difference between the boundary on the quantitative maps and Stojkov's?

Both questions can be answered relatively simply if we examine the phonetic maps of OT.[12] From these maps it becomes clear that Stojkov's _jat_-boundary forms part of a bundle of 48 isoglosses (this is the number of relevant maps in OT). The isoglosses reflect a great variety of phonetic characteristics, represented in 101 words in the data. Examples:

(1) Reflexes of _jat_ in specific positions

(2) Presence vs. absence of mixture of the reflexes of the two _jers_ and the two nasal vowels

(3) Vowel reduction phenomena

(4) Presence vs. absence of epenthetic _l_

(5) Change of *_d’_, *_t’_ into _g’_, _k’_

(5) Reflexes of *_lь_, *_lъ_ and syllabic *_l_

(6) Presence vs. absence of the changes *_a_ > [e] in certain positions

(7) Presence vs. absence of the change *_dn_ > [nn]

This bundle runs from north to south and occupies a broad strip of the Bulgarian dialect map. It contains more isoglosses to the east than to the west of the _jat_-boundary, reflected in 37 and 27 words, respectively.[13] The number of isoglosses accounts for the stability of the _jat_-boundary on the quantitative maps. The difference in geographic location of the boundary on the quantitative maps as compared to Stojkov's is in all probability due to the fact that on average the isoglosses in this bundle run slightly east of Stojkov's _jat_-boundary.

[12] As far as the dialects within the Bulgarian state borders are concerned, this atlas is for the most part a condensed edition of BDA. When referring to maps in OT, we shall use the letters "F" and "A" for maps that regard phonetics and accentology, respectively.

[13] East and west are divided by 48 isoglosses. Twenty-four of these run more to the east than the jat-boundary, nine more to the west. But more important for this discussion is the number of words in the data set that show the relevant characteristics. Some isoglosses are not represented in the data and others are represented several times.

5.2.2. Southeast vs. Northeast

Another similarity between the traditional map and the quantitative ones is the split between northeast and southeast, that is between north and south, east of the border discussed in the previous section (in terms of Fig. 5: the split between the Rupian dialects on the one hand and the Moesian and Balkan dialects on the other). This split appears in the three-way clustering for all three clustering techniques used. From the four-way clustering onward, two of the three clustering techniques show a further splitting up of the southeastern area into subareas. This confirms the impression we get from Fig. 6, viz. that the southeast is much less of a unity than the northeast.

If we examine the phonetic and accentual maps in OT, we see the same picture. There are many maps on which southeastern dialects differ from surrounding dialects, but more often than not this applies only to part of the southeast. Moreover, if one compares these maps among themselves, the parts of the southeast that distinguish themselves from their surroundings are not constant in any way and very often the relevant characteristic is shared by (sometimes considerable) areas outside the Rupian territory, especially by varying noncontingent areas in the northeast. For instance, on maps OT F 40–46 – which show reflexes of *jat* in the word *dve* (StBg) 'two' and in certain verbal endings (whether or not contracted with following **a* or **aa*) – we see a geographically nonconstant central area within Rupian that differs from its immediate surroundings but shows linguistic similarities with varying subareas elsewhere, mostly to the east and northeast. But there are also maps on which a larger part of the southeast or even the whole southeastern area distinguishes itself from the northeast. We shall give two examples:

(1) OT F 9: presence of epenthetic [ə][14] in such *l*-participles as StBg *pekla* (feminine) 'bake' (['pekəla]). This characteristic is shared by most (but not all) of the southeast and two noncontingent areas in the northeast.
(2) OT F 19: absence of a vowel in the verbal root **tъk-* (OBg) 'weave'. The whole southeast is opposed to the northeast here, but it shares its characteristic with the entire west.

We summarize that the impression one gets from the righthand map in Fig. 6 (a heterogeneous southeastern area opposed to a much less heterogeneous

[14] Notated *ъ* in OT.

northeastern one) corresponds with the impression one gets when comparing the phonetic and accentual maps in OT.

5.2.3. Stojkov's Moesian Dialect

On the rightmost map of Fig. 6 we also see a major difference between the quantitative and the traditional map: on the former the north-eastern dialects form one large group, whereas Stojkov distinguishes a Moesian group in the north. This difference between the two maps is not difficult to explain. Stojkov (1968, 69) mentions four phonetic features that he considers characteristic for the Moesian dialects:

> (1) *jat* is reflected as [a], [æ] or [ɛ] according to the phonetic environment, e.g. [bʲal] 'white' (masculine singular) vs. [ˈbɛli] (plural), cf. section 5.1 above.
> (2) The reflex of the OBg back *jer* has a 'velarized' phonetic realization.
> (3) Change of the consonant cluster **dn* into [nn], e.g. [ˈglanna] 'hungry' (feminine singular), cf. StBg *gladna*.
> (4) Nonexistence of the consonants *f* and *x*, e.g. [ˈodi] 'he goes', [sɤu̯ˈsem] 'entirely', cf. StBg *chodi*, *săvsem* (the third segment is realized as [f]).

However, if we consult OT and BDA we find that the first three of these distinguishing characteristics for the Moesian dialects are not supported by the maps. The reflexes of *jat* mentioned above are far from being typically Moesian (see OT F 34–35) and the same holds for the change **dn* > *nn* (OT F 166). As for the "velarized" articulation of OBg *jer*: such an articulation is distinguished neither in OT nor in BDA.

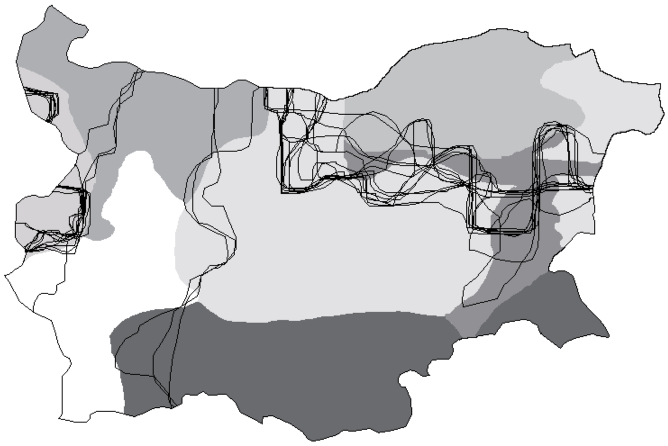
With respect to the nonexistence of *f* and *x*, we sometimes do find a map on which an area is visible that remotely resembles that of the Moesian dialects (OT F 135–141). In almost all cases the relevant characteristic is shared with significant areas to the east, west or south. The data set of the project contains 23 potentially relevant words. If we limit ourselves to these words and to the relevant segments of the words, we see that 15 words show an isogloss that runs more or less along the boundary of Stojkov's Moesian area. These 15 isoglosses are combined in Fig. 7.[15] The remaining 8 words do not

[15] The northeast corner of Bulgaria is not distinguished from the Moesian area as it is on Stojkov's map, but this is not surprising: there are no data from that area.

show such an isogloss. As we see on Fig. 7, even if we focus on the relevant segments of these 15 selected words, the isoglosses do not only delineate Stojkov's Moesian area but other parts of Bulgarian as well.

We conclude that as far as phonetics is concerned there is not enough evidence for distinguishing a Moesian dialect area. Three out of four phonetic characteristics are not visible on the traditional maps either, the fourth is sporadically present on the traditional maps and shows on the quantitative maps if one focuses on the relevant segments of the relevant words (15 out of 156). Apparently this signal is not strong enough to surface when the data as a whole is taken into account.

Fig. 7. Isoglosses of the Relevant Segments of the 15 Selected Words Discussed in Section 5.2.3



5.2.4. Northwest vs. Southwest; Stojkov's TR Zone

As we see in Fig. 5 above, Stojkov's map divides the western dialects into a thin Serbian transition region (TR) in the northwest, and then further subdivides the rest into a northern (NW) and a southern (SW) regions. The quantitative maps, however, do not reliably recognize the TR zone,[16] and never recognize the NW-SW division.

[16] WPGMA maps do (see Fig. 6 in section 5.2), but maps based on other clustering algorithms such as UPGMA do not.

These divergences between the quantitative and the traditional maps are more surprising than the others discussed above, as Stojkov justifies both divisions under discussion by referring to a considerable number of phonetic isoglosses.

With respect to the split NW vs. SW the phonetic characteristics concern:

(1) The reflexes of back *jer* in specific phonetic environments or in specific words
(2) The reflexes of front *jer*
(3) The reflexes of the back nasal
(4) Presence or absence of mixture of reflexes of back and front nasal
(5) Reflex of *jat* in *cjal* (StBg) 'whole' (masculine singular) and *celi* (plural)
(6) Final *o* or *e* in such words as *naše* (StBg) 'our' (neuter singular)
(7) Presence or absence of the second *j* in *jajce* (StBg) 'egg'

These characteristics are presented on 21 different maps in OT and are present in 21 words in the data.[17]

Stojkov distinguishes his TR dialects on the basis of the following characteristics:

(a) The reflexes of back and front *jer* in specific phonetic environments or in specific words
(b) Reduction or not of front *jer* in the suffix of such words as *žaden* (StBg) 'thirsty'
(c) The reflexes of the back nasal in specific words
(d) Reflexes of OBg **tj*, **ktj* and **dj* in general and in specific words
(e) palatalized or nonpalatalized *l* in such words as *bolna* (StBg) 'ill' (feminine singular)
(f) Labialization or not of *e* in certain phonetic environments

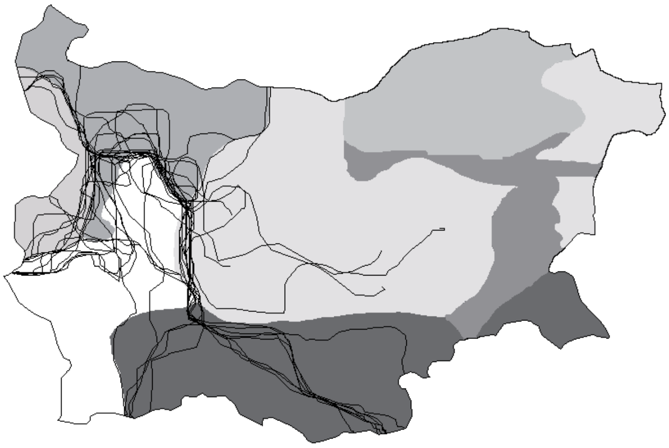
The characteristics given here are presented on 16 different maps in OT and are present in 22 words in our data.

In Fig. 8 below we see the isoglosses for the 21 words from the data set that show the relevant characteristics for the split NW vs. SW (nos. 1–7 above

[17] On these maps it is often the case that the NW (more rarely the SW) shares its characteristic with part of the dialects to the east (east of the jat-boundary), but this is not a problem since the split between east and west is undisputed.

in this section). For each word only the relevant phoneme is taken into consideration.[18] We could show a similar isogloss bundle for the distinguishing characteristics of the TR dialects (nos. a–f above in this section) but we shall not do so for reasons of space.

Fig. 8. Isoglosses of the 21 Words that Show a Northwest-Southwest Split



The existence in our own data of an isogloss bundle between the northwest and the southwest leads us to ask why the feature differences are not reflected in consistent differences in the final analyses.

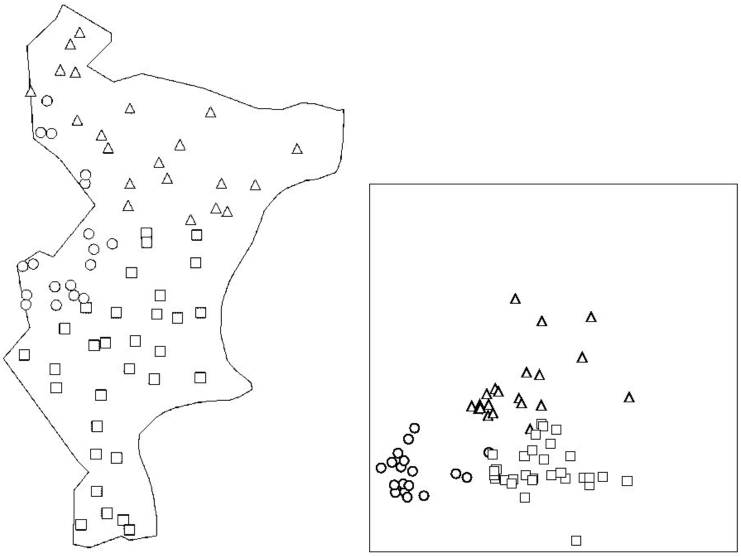
We shall try to shed some light on this matter by showing and analyzing some MDS plots. We shall examine the entire western region together, i.e., addressing both the question of Stojkov's TR zone and his proposed NW-SW split.

The MDS plot in Fig. 9 (below) clarifies that the TR group has a relatively distinct core, but that there are varieties intermediate between the TR varieties and the other western varieties. The region in the MDS plot between the TR group and the rest of the western dialects is not empty, as it would be if there were a substantial categorical division between the TR varieties on the one hand and the rest of the western varieties. In this case we are inclined to accept Stojkov's division, but note that is not a matter of very distinct subsets,

[18] The map not only confirms Stojkov's division into northwest and southwest, but some of the isoglosses also delineate Stojkov's TR dialect area. In addition, the same features confirm the east-west jat-boundary in the center of the country. In the north, the jat-boundary is strengthened by only two of the 21 features.

but rather two groups with some intermediate cases. We identify the problem in the clustering procedures, which are easily confused when the space between clusters is occupied as it is here.

Fig. 9. Stojkov's NW, SW and TR Dialects

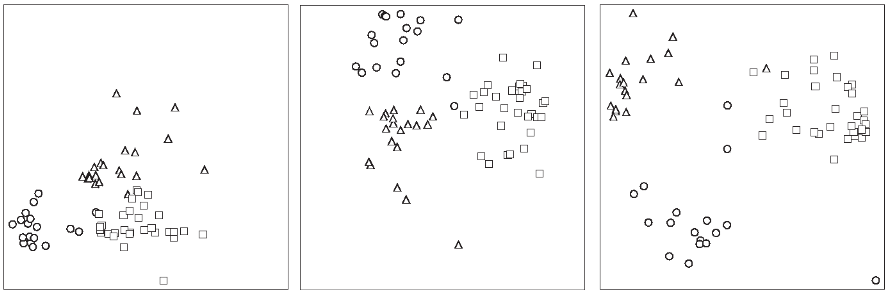


Left: Geographical location of Stojkov's NW, SW and TR dialects. Right: MDS plot of the linguistic distances between those dialects, legend: triangle – NW, square – SW, circle – TR.

If we now turn to the remaining western dialects in the MDS plots, again examining Fig. 9, we note that, while the SW varieties (squares) occupy a fairly compact section of the linguistic plane, the NW varieties (triangles) are comparatively diverse. This means that while it is possible to distinguish north and south, the decision of where to separate them will necessarily be arbitrary. These two groups are less clearly separated in the MDS plot. Our tentative conclusion is that the distinction is less clearly reflected in the data. Let us examine this more closely.

We attempt to focus on this issue by examining first, the aggregate distances based on just the words in which the relevant features appear, and second, the aggregate distances based on just the single segments themselves. The MDS plots are found in Fig. 10.

Fig. 10. MDS Plot of Western Dialects



Left: based on all 156 words. Center: focusing on the 21 words showing the relevant features. Right: focusing on the 21 segments themselves. Legend: triangle – NW, square – SW, circle – TR.

The MDS plot on the left in Fig. 10 is repeated from Fig. 9. We concluded in discussing it there that a core of the TR dialects could be identified even though some dialects in the area were not easily distinguished from the other western varieties. When we focus on just the 21 sounds Stojkov uses as a basis, we obtain a situation shown on the right, where all three regional varieties are clearly distinct. We note that there are borderline cases even in this very focused view, shown by the single triangle within the group of squares (village of Kreta, Vraca province), and the two circles which are closer to the squares than to the other circles. These are the villages Bučin prochod (Sofia province) and Velkovci (Pernik province). The MDS plot in the middle reflects the Levenshtein distances between the 21 words in which Stojkov's features appear, without focusing on the relevant segments. We include it here to show how the other information in the words tends to cloud the neater classification on the right.

With respect to our central question concerning the reason for the difference between the quantitative and the traditional divisions of Bulgarian dialects, we conclude first that the TR varieties are largely, but not consistently distinct from the other western varieties. Our clustering algorithms are not up to the task of consistently identifying the TR varieties as a distinct group. Second, we may discern a distinction between the NW and SW varieties along traditional lines in our data, but the distinction is clouded by a large number of features that are not distributed according to the north-south division.

6. Conclusion

Our goal in this paper was to compare traditional and quantitative classifications of Bulgarian dialects. We drew on Stojkov's authoritative work for our views on traditional classification, and we used a simple version of Levenshtein distance to provide a base for a quantitative view. The general lines of the two views of the Bulgarian dialect landscape are similar. Both see the language area dominated by an east-west division – Stojkov's *jat* line, and both identify the Rupian south as a third most significant area. The quantitative work located the *jat* line slightly to the east of where Stojkov had drawn it, and it failed to identify anything like his Moesian area. In both of these cases we find for the quantitative work, and tend to conclude that it improves on Stojkov's.[19]

Assuming that Levenshtein distance is a probative measure of aggregate pronunciation differences, we relied on multidimensional scaling (MDS) to visualize the more than 19,000 distances between the pairs in our 197-site sample, encouraged by the fact that over 92% of the variation is captured in the first two dimensions. This allowed us to see that the Rupian area is much more diverse than either the east or the west in the north. We noted that while it is possible to distinguish the Rupian varieties in the two-dimensional MDS plots, there is essentially no clear margin distinguishing them, which means that the exact demarcation will have to be somewhat arbitrary. We likewise inspected the results of various clustering algorithms, but these reliably distinguished only the east from the west along the *jat* line – all of Stojkov's further divisions escaped the dull eye of the clustering algorithms.

The situation in the west is similar to that in the Rupian area. The MDS plot demonstrates that the Serbian transition zone, the northern and southern parts of the west, all of which Stojkov postulated, may indeed be distinguished when using aggregate pronunciation distance, but the borders are not linguistically prominent. It is not surprising that clustering fails to distinguish these areas reliably, even if one algorithm, WPGMA, was able to distinguish the Serbian transition zone Stojkov had postulated.

We noted above that most of the work presented here proceeds from the assumption that Levenshtein distance is a valid measure of the pronunciation

[19] Of course it should not be forgotten that the quantitative classifications discussed in this paper were based exclusively on phonetic data and Stojkov's are not. Therefore we cannot be conclusive about this yet.

differences found in dialects. Naturally this assumption may be questioned: for example, the built-in sensitivity to segment frequency in Levenshtein distance (see footnote 6) may be inappropriate. Ultimately, we think such questions must be settled by testing dialect speakers on their sensitivity to pronunciation differences. Computational measures of pronunciation differences may be modified in myriad and complicated ways. We have theoretical reasons for preferring the simple version of the measure we have applied here, but ultimately, we need to test our ideas against the social sensitivity of dialect speakers.

References

Abbreviations

BDA (*Bălgarski dialekten atlas*) – Stojkov, S. et al. (ed.) 1964–1981.
OT (*Obobštavašt tom)* – Kočev, I. et al. (ed.) 2001.

Heeringa, W. 2004. *Measuring Dialect Pronunciation Differences using Levenshtein Distance*. Groningen: PhD thesis, University of Groningen. Available at http://irs.ub.rug.nl/ppn/258438452.

Kessler, B. 1995. "Computational Dialectology in Irish Gaelic". *Proceedings of the Seventh Conference of the European Chapter of the Association for Computational Linguistics*:60–67. Dublin: EACL.

Kočev, I. et al. (eds.). 2001. *Bălgarski dialekten atlas, obobštavašt tom I–III. Fonetika, akcentologija, leksika*. Sofia: Trud.

Kruskal, J. and M. Wish. 1978. *Multidimensional Scaling*. London: Sage.

Legendre, P. and L. Legendre. 1998. *Numerical Ecology*. Second edition. Amsterdam: Elsevier.

Levenštejn, V. I. (= Levenshtein, V. I.). 1965. "Dvoičnye kody s ispravleniem vypadenij, vstavok i zameščenij simvolov". *Doklady Akademii Nauk SSSR* 163 (4): 845–848 (English translation: Levenshtein, V. I. 1966. "Binary Codes Capable of Correcting Deletions, Insertions and Reversals". *Soviet Physics Doklady* 10: 707–710).

Maguire, W. 2008. "Quantifying Dialect Similarity by Comparison of the Lexical Distribution of Phonemes". *International Journal of Humanities and Arts Computing* 2(1–2), 261–277.

Manning, C. and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge: MIT Press.

Nerbonne, J. and W. J. Heeringa. 2010. "Measuring Dialect Differences". In P. Auer and J.E. Schmidt (eds.), *Language and Space. An International Handbook of Linguistic Variation. Vol. 1: Theories and Methods*, Berlin/New York: de Gruyter/Mouton, 550–567.

Nerbonne, J. and W. Kretzschmar (eds.). 2006. *Progress in Dialectometry*. Special issue of *Literary and Linguistic Computing* 21(4).

Prokić, J. and J. Nerbonne. 2009. "Recognizing Groups among Dialects". *International Journal of Humanities and Arts Computing, Special Issue on Language Variation*, edited by John Nerbonne, Charlotte Gooskens, Sebastian Kurschner, and Renée van Bezooijen.

Prokić, J., J. Nerbonne, V. Zhobov, P. Osenova, K. Simov, T. Zastrow, and E. Hinrichs. 2009. "The Computational Analysis of Bulgarian Dialect Pronunciation". *Serdica Journal of Computing* 3 (3): 269–298.

Stojkov, S. 1962. *Bălgarska dialektologija*. Sofia: Nauka i izkustvo.

———. 1968. *Bălgarska dialektologija*. Second edition. Sofia: Nauka i izkustvo.

———. 1993. *Bălgarska dialektologija*. Third edition. Sofija: Izdatelstvo na Bălgarskata akademija na naukite.

———. 2002. *Bălgarska dialektologija*. Fourth edition. Sofia: Prof. Marin Drinov.

Stojkov, S. and M. Mladenov. 1969. "Proekt za ideografski dialekten rečnik na bălgarskija ezik". *Bălgarski ezik* 2:155–170.

Stojkov, S. et al. (ed.). 1964–1981. *Bălgarski dialekten atlas I–IV*. Sofia: Izdatelstvo na Bălgarskata akademija na naukite.

Appendix

Here we present the 156 words that form the data set of the project. The forms in Cyrillic script represent the StBg words in their accepted orthography. The phonological notation used also refers to StBg. Phonetic details such as final devoicing or the effects of vowel reduction are not represented. Unless indicated otherwise, verb forms are in the present and in the first person singular; nouns, adjectives and participles are in the singular masculine (if applicable). If a verb form has an indication for person, it is a present form unless indicated otherwise.

аз /az/ 'I', агне /ˈagne/ 'lamb'

бели /ˈbeli/ 'white-PL', берат /beˈrɤt/ 'pick-3PL', беше /ˈbeʃe/ 'be-PAST.3SG', бране /braˈne/ 'pick-VERB.NOUN', брашно /braʃˈno/ 'flour', бързо /ˈbɤrzo/ 'quickly', бяхме /ˈbjaxme/ 'be-PAST.1PL'

вежда /ˈveʒda/ 'eyebrow', вече /ˈveʧe/ 'already', вечер /ˈveʧer/ 'evening', видях /viˈdjax/ 'see-AOR', вие /ˈvie/ 'you-2PL', вино /ˈvino/ 'wine', влизам /ˈvlizam/ 'enter', вода

/voˈda/ ‘water’, вол /vol/ ‘ox’, време /ˈvreme/ ‘time’, връх /vrɤx/ ‘peak’, връщам /ˈvrɤʃtam/ ‘give back’, вчера / ˈ vʧera/ ‘yesterday’, във /vɤv/ ‘in’, вълк /vɤlk/ ‘wolf’, вълна /ˈvɤlna/ ‘wool’, вънка /ˈvɤnka/ ‘outside’, вътре /ˈvɤtre/ ‘inside’, вятър /ˈvjatɤr/ ‘wind’

глава /glaˈva/ ‘head’, гладен /ˈgladen/ ‘hungry’, говедо /goˈvedo/ ‘bovine animal’, горе /ˈgore/ ‘upstairs’, гости /ˈgosti/ ‘guest-PL’, градът /graˈdɤt/ ‘the city’, грозде /ˈgrozde/ ‘grapes’

дадоха /ˈdadoxa/ ‘give-AOR.3PL’, две /dve/ ‘two’, двор /dvor/ ‘yard’, ден /den/ ‘day’, дера /deˈrɤ/ ‘flay’, десет /ˈdeset/ ‘ten’, дете /deˈte/ ‘child’, джоб /ʤob/ ‘pocket’, днес /dnes/ ‘today’, добре /doˈbre/ ‘well-ADV’, долу /ˈdolu/ ‘downstairs’, дошъл /doˈʃɤl/ ‘come-AOR.PART’, дъжд /dɤʒd/ ‘rain’, дълбок /dɤlˈbok/ ‘deep’, дъно /ˈdɤno/ ‘bottom’, дърво /dɤrˈvo/ ‘tree’

един /eˈdin/ ‘one-MASC’, едно /edˈno/ ‘one-NEUT’, език /eˈzik/ ‘tongue’, ечемик /eʧeˈmik/ ‘barley’

желязо /ʒeˈljazo/ ‘iron’, жена /ʒeˈna/ ‘woman’, жив /ʒiv/ ‘alive’, жълт /ʒɤlt/ ‘yellow’, жътва /ˈʒɤtva/ ‘harvest’

звезда /zveˈzda/ ‘star’, здрав /zdrav/ ‘healthy’, земя /zeˈmja/ ‘land’, зет /zet/ ‘son-/brother-in-law’

ѝ /i/ ‘she-DAT’, им /im/ ‘they-DAT’, име /ˈime/ ‘name’

камък /ˈkamɤk/ ‘stone’, ключ /kljuʧ/ ‘key’, кое /koˈe/ ‘which-NEUT’, кон /kon/ ‘horse’, кръв /krɤv/ ‘blood’, къде /kɤˈde/ ‘where’

лесно /ˈlesno/ ‘easily’, леща /ˈleʃta/ ‘lentils’

майка /ˈmajka/ ‘mother’, месец /ˈmeseʦ/ ‘month’, месо /meˈso/ ‘meat’, млякото /ˈmljakoto/ ‘the milk’, много /ˈmnogo/ ‘much, many’, мъж /mɤʒ/ ‘man’, мъже /mɤˈʒe/ ‘men’, мъжът /mɤˈʒɤt/ ‘the man’

наше /ˈnaʃe/ ‘our-NEUT’, неделя /neˈdelja/ ‘Sunday’, неще /ˈneʃte/ ‘not want-3SG’, нещо /ˈneʃto/ ‘something’, нея /ˈneja/ ‘she-ACC’, ние /ˈnie/ ‘we’, носят /ˈnosjɤt/ ‘carry-3PL’, нощ /noʃt/ ‘night’, няма /ˈnjama/ ‘there is no’

овца /ovˈʦa/ ‘sheep’, овце /ovˈʦe/ ‘sheep-PL’, овчар /ovˈʧar/ ‘shepherd’, овчари /ovˈʧari/ ‘shepherd-PL’, огън /ˈogɤn/ ‘fire’, онези /oˈnezi/ ‘those’, орех /ˈorex/ ‘walnut’

пека /peˈkɤ/ ‘bake’, пепел /ˈpepel/ ‘ash’, петел /peˈtel/ ‘rooster’, петък /ˈpetɤk/ ‘Friday’, плащам /ˈplaʃtam/ ‘pay’, понеделник /poneˈdelnik/ ‘Monday’, пръч /prɤʧ/ ‘billy-goat’, първият /ˈpɤrvijɤt/ ‘the first’, път /pɤt/ ‘road’, пясък /ˈpjasɤk/ ‘sand’

река /reˈka/ ‘river’, ръка /rɤˈka/ ‘hand’, ръце /rɤˈʦe/ ‘hand-PL’

се /se/ ‘oneself’, сега /seˈga/ ‘now’, седя /seˈdjɤ/ ‘sit’, сестра /seˈstra/ ‘sister’, сирене /ˈsirene/ ‘cheese’, сол /sol/ ‘salt’, средата /sreˈdata/ ‘the middle’, сряда /ˈsrjada/ ‘Wednesday’, старец /ˈstareʦ/ ‘old man’, страх /strax/ ‘fear’, сух /sux/ ‘dry’, събота /ˈsɤbota/ ‘Saturday’, сърп /sɤrp/ ‘sickle’, със /sɤs/ ‘with’

такъв /taˈkɤv/ ‘such’, твой /tvoj/ ‘yours’, това /toˈva/ ‘this-NEUT’, тогава /toˈgava/ ‘then’, тъмно /ˈtɤmno/ ‘dark-NEUT’, тънко /ˈtɤnko/ ‘thin-NEUT’, трева /treˈva/ ‘grass’

утре /ˈutre/ ‘tomorrow’, ухо /uˈxo/ ‘ear’

фурна /ˈfurna/ ‘oven’

хляб /xljab/ ‘bread’, хоро /xoˈro/ ‘chain dance’, хубав /ˈxubav/ ‘beautiful’, хубаво / ˈxubavo/ ‘beautiful-NEUT’

цял /ʦjal/ ‘whole’

чакат /ˈʧakat/ ‘wait-3PL’, червен /ʧerˈven/ ‘red’, черен /ˈʧeren/ ‘black’, череша /ʧeˈreʃa/ ‘cherry’, чета /ʧeˈtɤ/ ‘read’, чешма /ˈʧeʃma/ ‘fountain’, човек /ʧoˈvek/ ‘human’

ще /ʃte/ ‘will (all persons)’

я /ja/ ‘she-ACC’, ябълка /ˈjabɤlka/ ‘apple’, ябълки /ˈjabɤlki/ ‘apple-PL’, яйце /jajˈʦe/ ‘egg’, яйца /jaˈjʦa/ ‘egg-PL’, ям /jam/ ‘eat’, ядеш /jaˈdeʃ/ ‘eat-2SG’