Serdica J. Computing **3** (2009), 269–298

Serdica
Journal of Computing

Bulgarian Academy of Sciences
Institute of Mathematics and Informatics

THE COMPUTATIONAL ANALYSIS OF BULGARIAN DIALECT PRONUNCIATION

Jelena Prokić, John Nerbonne, Vladimir Zhobov, Petya Osenova, Kiril Simov, Thomas Zastrow, Erhard Hinrichs

Abstract. The paper presents a computational analysis of Bulgarian dialect variation, concentrating on pronunciation differences. It describes the phonetic data set compiled during the project* 'Measuring Linguistic Unity and Diversity in Europe' that consists of the pronunciations of 157 words collected at 197 sites from all over Bulgaria. We also present the results of analyzing this data set using various quantitative methods and compare them to the traditional scholarship on Bulgarian dialects. The results have shown that various dialectometrical techniques clearly identify east-west division of the country along the 'jat' border, as well as the third group of varieties in the Rodopi area. The rest of the groups specified in the traditional atlases either were not confirmed or were confirmed with a low confidence.

1. Introduction. Computational dialectometry is a multidisciplinary field that uses various quantitative methods in the analysis of dialect data. Work

ACM Computing Classification System (1998): J.5.

Key words: Dialectology, Computational dialectometry, Phonetic similarity.

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

in a dialectometry began with [31] who invented the first technique for measuring the distances between dialects. He aggregated over the individual differences between sites by counting the overlapping features between any two sites. In this way he introduced an aggregate view of language variation, as opposed to the traditional division of sites based on the individual linguistic features. Further improvement in the development of dialectometry came with the work of Hans Goebl [4, 5] who introduced weighting features. Brett Kessler [15] was the first to use Levenshtein distance in order to calculate the linguistic distance between dialects. Levenshtein distance was later successfully applied to many other languages. For a detail overview of the development of dialectometry and recent trends in the filed see [25, 24].

In the project 'Buldialect–Measuring Linguistic Unity and Diversity' quantitative methods for measuring linguistic diversity were applied to the dialect pronunciation data created as a part of the project. The data was collected and digitalized as a joint work between the University of Sofia, and the Institute for Parallel Processing, Bulgarian Academy of Sciences. These machine-readable data were the basis for applying various methods taken from dialectometry and *information theory* in order to get new insights into Bulgarian dialect variation at one hand, and into further develop quantitative methods for the study of language on the other.

The structure of this paper is as follows. A detailed description of the data set is presented in the next section. In Section 3 we discuss the results of analyzing the data using multidimensional scaling and hierarchical clustering. An information theoretic approach to the same data is described in Section 4. In Section 5 we present discussion and conclusions.

2. Data description. The phonetic data set of the Buldialect project consists of the varying pronunciations of 157 words collected at 197 sites from all over Bulgaria (see Figure 1). The main source of the data was the large dialect archive at the University of Sofia. The word pronunciations started to be gathered in 1950s, and this work continues till now. For this purpose, expressly designed questionnaires were used. More than one person was interviewed at each village. For some missing concepts and/or sites, additional expeditions were organized on the spot.

Part of these data was selected and converted into X-SAMPA encoding for further computer processing and into IPA encoding for human usage. In the next two subsections we give a detailed description of the sources used for the data collection, as well as the main phonetic characteristics present in the data set.

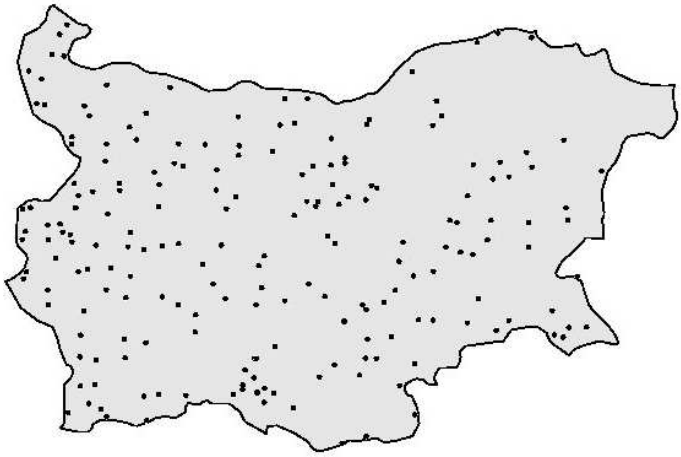

Fig. 1. Distribution of 197 sites

2.1. Sources for the pronunciation data. The sources for pronunciation data are of various types. These are supervised students' theses, published monographs, dictionaries, and the archive of the *Ideographic Dictionary of Bulgarian Dialects*. These are described in detail below.

2.1.1. Theses. The principal source for the pronunciation data are theses written by graduating students of Bulgarian language at the University of Sofia. Each thesis is a complete description of the dialect of a particular village (in almost all cases the native village of the student). The collection of these descriptions began in the end of 1950s and intensified significantly in the following decades. Most of the theses from the initial period were supervised by Prof. Stojko Stojkov ([39] and [33]) — the leading expert in the field of Bulgarian dialectology at the time, while others were supervised by two of his most distinguished students — Prof. Todor Bojadzhiev and Prof. Maksim Mladenov. The majority of the theses used for the pronunciation data were written in the period 1960–1985, very few of them earlier or later. 73 of them date from 1960s and 58 from 1970s, when tape-recorders became more available. It is important to note that Prof. Stojkov formed a working group (referred to as a 'circle') in Bulgarian dialectology, which was among the most popular extracurricular activities in the faculty. In this group the students received additional training in field work and phonetic transcription.

Stojkov's basic assumptions were that a dialect is a self-contained linguistic system and that a satisfactory dialect description should provide a thorough account of all levels of this system, contrary to the practice of collecting and

describing only exotic and rare words and features. The theses follow his assumptions: there are chapters on phonetics (including historical changes), morphology, notes on the syntax, a dictionary and transcribed dialect texts. The phonetic transcription system used in the descriptions was developed in its present form primarily by Stojkov and now is in general use in Bulgarian linguistic publications. It is based on the Cyrillic alphabet with some Latin letters and many diacritics added. This system allows a very detailed representation of phonetic variation. For example there are 8 basic symbols for vowels and diacritics for two degrees of vowel reduction, for raised or lowered pronunciation, rounding, and length. Stojkov recommended the introduction of new symbols and detailed descriptions for specific sounds. It is important to note that this system can be adequately, and biuniquely, translated into the symbols of IPA.

The basic methods for the collection of dialect material were the observation of natural dialect speech and work with questionnaires (the latter was primarily applied in collecting lexical data [35] and [42]). Direct questioning was greatly disfavored, if not downright prohibited. The informants were selected among the oldest inhabitants of the village under the strict condition that they were born locally. Work with only one informant per site was considered unacceptable. Preference was given to women because they were socially and otherwise less mobile at the time. The conversations were centered on traditional rural life — customs, religious practices, agricultural work, surrounding nature — and the field-workers were instructed to intrude as little as possible in order to obtain longer chunks of dialect texts.

2.1.2. Dialect descriptions and dictionaries. Published dialect descriptions and dictionaries are another important source. There are two series of such publications — *Bulgarian Dialectology. Investigations and Data* (comprising 10 volumes, 1962–1981), and *Studies in Bulgarian Dialectology* (comprising 10 volumes, 1965–1984), and also standalone books. Most of the villages for which such monograph-length descriptions are available are included in the following list (several dictionaries were published after the work on the project started and therefore could not be included): Dobroslavci [7], Gabare [28], Govedarci [36], Hvoyna [14], Momchilovci [13], Mugla [38], Nova Nadezhda [10], Pavelsko [14], Radovene [9], Vojnjagovo [30].

There are also book-length descriptions of larger areas, e.g. Godech [44], Ihtiman [22] and [23], Kjustendil [43], Silistra [16], Sofia [27], Strandzha mountain [6], Teteven [32], Troyan [17]. They provide complete descriptions of the dialect in the region and pay attention to internal variation. The above mentioned system for phonetic transcription is used consistently in all these books.

2.1.3. Ideographic Dictionary of Bulgarian Dialects.

Another important source is the archive of the *Ideographic Dictionary of Bulgarian Dialects*. This project was launched by Prof. Stojko Stojkov in the middle of the 50s. If in most dialect dictionaries the dialect words are arranged alphabetically and are explained or translated into the standard language, the *Ideographic Dictionary* reverses that order: the word of the standard language are alphabetically arranged and are followed by the corresponding dialect words. Thus all the dialect words meaning 'potato' are in a single entry.

The material for the dictionary was collected from all possible sources: theses and term papers written on the bases of a questionnaire composed by Stojko Stojkov [34]; abundant material from field work expeditions, which were regularly organized in the summers; all published dialect descriptions and dictionaries; and the personal archives of other scholars. In addition, the archive of the dialectological section of the Institute for Bulgarian Language was also consulted (in the form in which it existed in 1969 when Stojko Stojkov was the head of the Institute). The material is in the form of index-cards (over two million), each containing a dialect word, its counterpart in the standard language, and its location. All the materials used for the compilation of the dictionary are transcribed with uniform phonetic transcription. A description of the archive was published in the journal 'Bylgarski ezik' in 1969, v. 2 [41]. The archive is in the process of transferring the material to computers (the work is completed up to the letter Д (D), also parts of the letters Е (E) and Ж (ZH)).

2.1.4. Tape recordings

Tape recordings of dialect speech are another important source. Since 1981 we have collected a phono-archive. We now have over 250 hours of recorded dialect speech from ca. 100 villages from all parts of the Bulgarian language territory. There is either a thesis or a published description for all but 3 villages (Garvan (Silistra), Huhla (Ivajlovgrad) and Drabishna (Ivajlovgrad)). The files for these villages were filled with material from tapes and the results are quite satisfactory. The inclusion of these villages was necessary in order to obtain a more adequate geographic network. Finally, the material from two villages (Vresovo (Ajtos) and Karanovo (Ajtos)) was collected in the field by Georgi Kolev. *The Bulgarian Dialect Atlas* is also regularly used to verify the accuracy of our material.

We ultimately analyzed the pronunciation differences in 157 words. The use of an unusually long list has the advantage that the signal of provenance emerges strongly and also that many variations are analyzed, some of which might be absent in a shorter list.

2.2. Criteria for the selection of words. It is evident that the first requirement in selecting words is their availability. The complete list of words in the data set can be found in Appendix A. The words included in the list are frequent and almost invariably show up in the theses (one quarter of the words are in the Swadesh list). Only words which are expected to show some degree of variation were included, which is why we did not use the entire Swadesh list. It is also evident that words displaying lexical variation were not included in the sample of words whose pronunciation differences were analyzed. For example the word дъб /dɤb/ 'oak' was replaced by a Turkish borrowing in a number of villages. We did, however, included a limited number of words where the variation is strictly speaking, morphological, rather than phonetic or phonological.

2.2.1. General remarks. There is a balance between the different segments represented: the reflexes of all important Old Bulgarian vowels are represented with the same (or nearly the same) number of words to avoid skewing of the results. For example there are three words with the reflex of the back nasalized vowel and three words with the reflex of the back jer in the root. In cases where we had to choose between two frequent words containing one and the same feature, preference was given to the word that displays more than one variation. Thus път /pɤt/ 'road' was preferred to зъб /zɤb/ 'tooth' because in addition to the variation of the root vowel there is also variation of the final consonant (the final consonant of зъб /zɤb/ 'tooth' is also subject to variation, viz. the preservation of final voicing, but it is much more limited and is represented by other words in the list). The word събота /sɤbota/ 'Saturday' was preferred to работа /rabota/ 'work' because apart from the frequent vowel elision съпта /sɤpta/ it contains the reflex of the back nasalized vowel.

Some words which were included primarily for one feature contain other features as a side effect, so to speak. Such additional features are also represented in the data. For example the word звезда /zvezda/ 'star' was included for the initial consonant (fricative or affricate), but it contains three other features: initial vs final stress, the reflex of 'jat' in the first syllable, and the generalized accusative case (if the stress is final). The word вежда /veʒda/ 'eyebrow' was included for the reflex of *dj, but also contains the same three additional features. The asterisk after the number of a feature in the list below signifies that the inclusion of the word in the list is subject to some condition. Thus word глава /glava/ 'head' will be considered for generalized accusative only if the stress is final.

2.2.2. Feature description. There are 39 different dialectal features which have been represented in our choice of 157 words. Below is a list of underlying linguistic features followed by a short description of each.

Reflexes of 'jat'

This is the well known 'jat' boundary (see, for example [21] and [37]), dividing the Bulgarian dialects into two large groups — eastern and western dialects. West of the boundary the reflex is always [e] (this is slightly simplified, as some western villages have a more open vowel, and in other the reflex is [a] after /r/ and /ʦ/) and east of the boundary the reflex is [ja] or [ɛ]. Examples: хляп /xlʲap/, хлеп /xlep/, хлеп /xlɛp/ 'bread'; горе /gore/, гори /gori/, гор'ъ /gorʲɤ/ 'upstairs'

Etymological 'ja'

The term etymological 'ja' refers to the vowel a preceded by the palatal approximant [j] or a post-alveolar consonant. Examples: офчар /oftʃar/, офчер /oftʃer/, офчер /oftʃɛr/ 'shepherd'; ядеш /jadeʃ/, едеш /edeʃ/ 'eat-you'

Initial prothetic j

Examples: агне /agne/, йагне /jagne/ 'lamb'; език /ezik/, йезик /jezik/ 'tongue'; утре /utre/, ютре /jutre/ 'tomorrow'

j before front vowels

In the standard language the palatal approximant is not allowed before front vowels (with the exception of a few rarely used borrowings) which leads to alternations like пея /pejɤ/ 'sing-I' — пееш /peeʃ/ 'sing-you'. In the dialects j may be kept before front vowels, especially e. Examples: кое /koe/, койе /koje/ (also кве /kve/) 'which'

Elision of j

Example: нея /neja/, неа /nea/ 'her-acc'

Reflexes of the back nasalized vowel

This is one of the most important dialect features in Bulgarian and is invariably used in dialect classifications. In fact, when groups of dialects are referred to as a-dialects, u-dialects and so on, the names of the dialects come from the reflex of the back nasalized vowel. The areas of the different reflexes differ in size (very large areas for /ɤ/ and /a/, very small for /o/ and /e/), but the sound change is remarkably consistent and there is very little, in fact negligible lexical conditioning. It is possible to predict with great certainty the pronunciation of other words on the basis of the three words included in the list. Examples: мъш /mɤʃ/, маш /maʃ/, муш /muʃ/, мош /mɔʃ/, мош /moʃ/, меш /mɛʃ/, мънч /mɤntʃ/ 'man'; каде /kade/, куде /kude/ 'where'; вътре /vɤtre/, унутре /unutre/, внетре /vnetre/, натре /natre/ 'inside'

Reflexes of the front nasalized vowel

This is also an important feature, though it is used less often in classifications. Some of the reflexes, such as зит /zit/ 'son-in-law' or 'brother-in-law', are quite rare. It is also very consistent and there is no lexical conditioning, except in the formation of secondary imperfective verbs, where the generalization of various vowel alternations is possible (наредя /naredjɤ/ — нарядам /narjadam/ 'arrange' analogously to седна /sedna/ — сядам /sjadam/ 'sit'). Examples: зет /zet/, зьот /zʲɔt/, з'ът /zʲɤt/, зит /zit/, зент /zent/ 'son-in-law' or 'brother-in-law'; десет /deset/, десит /desit/, дес'ът, /desʲɤt/, десат, /desat/ 'ten'

Reflexes of the back 'jer'

The development of the back 'jer' is consistent in the areas where the reflex is ъ /ɤ/ (and also ɒ or ɛ) but only in stressed root syllables. There is a great deal of lexical variation in the southwest and in fact the reflexes must be studied word by word. There is a peripheral area in the southwest where the reflex in the root is consistently /a/, and moving southwest one finds more and more /o/ reflexes. Examples: дъшт /dɤʃt/, дошт /doʃt/, дашт /daʃt/, дошт /dɔʃt/, дешт /dɛʃt/ 'rain'; такъф /takɤf/, такоф /takof/, такаф /takaf/, такоф /takɔf/, такеф /takɛf/ 'such'

Reflexes of the front 'jer'

The reflexes of the front 'jer' exhibit even more lexical variation than the reflexes of the back 'jer' in a broad area. Only the extreme southeast is relatively consistent in having the reflex e [e] and the extreme west is absolutely consistent in having the reflex ъ /ɤ/. In these cases, other words can be safely predicted on the basis of the words in the list, as there is no lexical variation with respect to this vowel. This may be termed conditioned predictability, since only if the reflex is ъ /ɤ/ and is common to all words, other words can be predicted. Examples: тънко /tɤŋko/, тенко /teŋko/, тьонко /tʲɔŋko/, тенко /tɛŋko/ 'thin-neuter'

Epenthesis of 'jer'

The first word below ended in Old Bulgarian in back 'jer' and the second in front 'jer', and there was no vowel between the final two consonants in both words. The elision of the word-final, and therefore weak, 'jer' likely resulted in an inadmissible syllabic structure, more specifically, in a syllable-final combination of obstruent and sonorant, and a vowel was inserted between the two consonants. The vowel inserted in the first word is to a certain degree irregular, as many dialects have inserted e /e/ only in this word and ъ /ɤ/ in all other words under the same phonetic conditions. The vowel inserted is often specific for this word alone. Examples: вятър /vʲatɤr/, ветер /veter/ 'wind'; огън /ogɤn/, огин /oɟin/ 'fire'

Vowel reduction

Vowel reduction is by far more common in the eastern dialects. The vowel reduction in the standard language is interpreted as a purely phonetic rule, conditioned by missing stress. In the dialects, however, the vowel reduction is often lexicalized or conditioned by morphological factors. Especially unpredictable is the reduction of unstressed e /e/, which may depend on the consonantal environment. The word пепел /pepel/ 'ash' may have an additional variation in form of a back rounded vowel in the first syllable. Examples: пепел /pepel/, пепил /pepil/, пеп'ъл /pepʲɤl/ 'ash'

Reflexes of 'jery'

Except in two small areas, one of them outside the modern borders, the Old Bulgarian 'jery' merged with the old i. Examples: език /ezik/, език /eзik/ 'tongue'

Rounding of vowels

Rounding of front vowels occurs in consonantal environments of labial/labiodental or postalveolar consonants (as in many other languages, the articulation of Bulgarian postalveolar consonants involves rounding of the lips). The vowel /i/ is subject to rounding much more frequently. The rounding may be accompanied by retraction all the way to a back vowel, in which case the preceding consonant (with the possible exception of the postalveolars) is palatalized. The change is found almost exclusively in eastern dialects. Examples: жиф /ʒif/, жйф /ʒyf/, жуф /ʒuf/ 'alive'

Unrounding of vowels

This sound change, the opposite of the one in *Rounding of vowels*, is less common and found in fewer words, though some of them, like либе /libe/ 'sweetheart', have made their way into the standard language thanks to the fact that the sound change is found in the dialect of Koprivshtica, where several classical writers were born. Examples: ключ /klʲutʃ/, клич /klitʃ/ 'key'

Alternation o-e

This alternation is another example of the Proto-Slavic syllabic synharmonism. After soft and postalveolar consonants only front vowels were allowed. The alternation lost its phonetic regularity but was preserved in numerous morphophonemic alternations, e.g. the singular ending of the neuter nouns: село /selo/ 'village' but въже /vɤʒe/ 'rope'. The alternation is better preserved in western and southeastern dialects. Examples: джоп /ʤop/, джеп /ʤep/ 'pocket'

Vowel elision

Elision of unstressed vowels is best attested in southeastern and northeastern dialects. The elision may be conditioned by position: in a trisyllabic word with initial stress the middle vowel is likely to be lost (рапта /rapta/ < работа /rabota/ 'work', съпта /sɤpta/ < събота/sɤbota/ 'Saturday'). It may also be morphologically conditioned: the plural ending is lost before the definite article даскалте /daskalte/ < даскалите /daskalite/ 'the teachers'). Examples: неделя /nedelʲa/, нделя /ndelʲa/ 'Sunday'

Change by analogy

The only plausible explanation for some changes appears to be analogy. For example, we find долу /dolu/ as well as доле /dole/ 'down', presumeably due to analogy with горе /gore/ 'up'. There are other likely cases of analogy, even though alternation y-e /u-e/ is not otherwise attested. Examples: долу /dolu/, доле /dole/ 'down' (analogy with горе /gore/ 'up'); пека /pekɤ/ 'bake-I', пекат /pekɤt/ 'bake-they', печа /peʧɤ/ 'bake-I', печат /peʧɤt/ 'bake-they'

Syllabic liquids

The old syllabic liquids were preserved in many western (especially northwestern) dialects and were replaced by a combination of liquid consonant and vowel (most frequently ъ /ɤ/, but other vowels are possible) in other dialects. The sequence of the liquid and the vowel in dialects without syllabic liquids also differ. Some dialects favor a fixed order (ръ /rɤ/, лъ /lɤ/ or ър /ɤr/, ъл /ɤl/) and do not even permit the other sequence. More dialects have the alternation ър /ɤr/ — ръ /rɤ/ and ъл /ɤl/ — лъ /lɤ/. In monosyllabic words the sequence is usually unpredictable. In polysyllabic word the sequence is conditioned by the number of the following consonants (държа /dɤrʒa/ 'hold' — дръжка /drɤʒka/ 'handle'; гълтам /gɤltam/ 'swallow — imperfective aspect' — глътна /glɤtna/ 'swallow — perfective aspect'). The alternation is found in inflection as well as in word-formation. In many dialects the combinations of liquids and the back nasalized vowel merged with the old liquids (търся /tɤrsʲa/ < тр̨сити /tr̨siti/ 'search', кълбо /kɤlbo/ < кл̨бо /kl̨bo/ 'ball'), but not in the Rhodopes and in the westernmost dialects. The early Old Bulgarian contrast between syllabic liquids and liquids followed by a 'jer' left no traces in Bulgarian dialects. Examples:

Syllabic r: сърп /sɤrp/, сръп /srɤp/, срп /sr̩p/, сорп /sɔrp/, серп /sɛrp/ 'sickle'

Syllabic l: вълк /vɤlk/, влък /vlɤk/, влк /vl̩k, вък /vɤk/, вук /vuk/, волк /vɔlk/, велк /vɛlk/ 'wolf'

Reflexes of *tj, *dj

The most common reflexes of these Proto-Slavic clusters are шт /ʃt/ and жд /ʒd/, but other reflexes are found, of which the postalveolar affricates form a compact area in the west. There is some irregularity in the reflex of the *dj in вежда /veʒda/, but the other possibilities either display lexical variation, like межда /meʒda/ 'landmark', or are less available, like прежда /preʒda/ 'yarn'. Examples: леща /leʃta/, лешча /leʃʧa/, леча /leʧa/ 'lentils'; вежда /veʒda/, вежа /veʒa/, вежджа /veʒʤa/, веджа /veʤa/ 'eyebrow'

The clusters чрь, чрѣ

The variation in these words concerns the initial consonant (alveolar or postalveolar affricate), and also the vowel, which may also be replaced by a syllabic liquid. The area of the alveolar affricate in череша /ʧereʃa/ 'cherry' is smaller than the area of the same consonant in черен /ʧeren/ 'black' and червен /ʧerven/ 'red', found not only in western, but in many southeastern dialects. Examples: червен /ʧerven/, цървен /ʦɤrven/ 'red'; череша /ʧereʃa/, црешня /ʦreʃnʲa/ 'cherry'

Epenthetic л /l/

The palatal approximant j in Old Bulgarian affected the preceding consonant in a variety of ways, depending on its place of articulation. After the labial consonants p, b, m, and v (f did not exist in native vocabulary) an epenthetic palatal lateral consonant developed. The process can also be described as a change j > l. The other possibilities are j > n, coalescence of the labial consonant and the palatal approximant into a single consonant with secondary palatal articulation, or preservation of j. The reverse changes, l > j and n > j are also found in Bulgarian dialects, the first being more common. It is not clear whether the epenthetic l was lost or never existed in the first place in the dialects where it does not occur. Examples: земя /zemʲa/, земля /zemlʲa/, земня /zemnʲa/ 'land'

Voiced affricates

A group of dialects in the southeast lack the two voiced affricates. The voiced postalveolar affricate is more frequent than its alveolar counterpart, despite being found only in borrowed, primarily Turkish words. (It has been suggested that the oldest result of the so-called first palatalization of г /g/ was an affricate, which was later replaced by a fricative, while the result of the first palatalization of к /k/ remained an affricate.) The sound appears in native vocabulary as a reflex of the *dj in some dialects, and in other dialects may be found in the place of г /g/ before front vowels (the process was dubbed 'new first palatalization' by Stojko Stojkov). Examples: джоп /ʤop/, жоп /ʒop/ 'pocket'; звезда /zvezda/, дзвезда /ʣvezda/ 'star'

Soft consonants

The impressionistic, actually synesthetic term 'soft' has the advantage of encompassing into a single category two phonetically different groups of consonant: palatal and palatalized consonants. The inclusion of these sounds in a single category is justified by their common phonotactic behavior and phonemic status. It is often claimed that the western dialects have only four soft consonants but that they are softer than the soft consonants in the eastern dialects. In somewhat stricter phonetic terms that means that the western soft consonants are palatal (к /k/, г /g/, л /l/, н /n/), while the eastern ones are palatalized, and each consonant except the postalveolars is paired with a palatalized counterpart. In fact the soft counterparts of k and g in the eastern dialects are also palatals, which leaves only the soft lateral and nasal consonants with different pronunciation in the dialects. The standard pronunciation of these soft consonants is ль /lʲ/ and нь /nʲ/ and the use of palatal consonants may sound regional, but not all speakers are sensitive to such a small phonetic difference. Examples: (й)агне /(j)agne/, (й)агн'е /(j)agnʲe/ 'lamb'; майка /majka/, мак'а /maca/ 'mother'; влк /vl̩k/, вльк /vl̩ʲk/ 'wolf'; понеделник /ponedelnik/, понедельник /ponedelʲnik/ 'Monday'; сирене /sirene/, сиренье /sirenʲe/ 'cheese'; фурна /furna/, фурн'а /furnʲa/ 'oven'; ябълка /jabɤlka/, ябълька /jabɤlʲka/ 'apple'; ябълци /jabɤlci/, ябъльци /jabɤlʲci/ 'apple'

Palatalization of т /t/, д /d/

The examples below differ morphologically — the first is a plural form of a masculine noun and the second is a singular form of a neuter noun, but they follow the same pattern almost invariably. The palatalization, where it occurs, is caused by a palatal approximant following the alveolar stop. Examples: гости /gosti/, гос'е /gosʲe/, гойсе /gojse/ 'guests'; грозде /grozde/, грозг'е /grozǰe/, гроз'е /grozʲe/, гройзе /grojze/ 'grapes'

Simplification of the clusters стр /str/, здр /zdr/

This simplification is a feature of some dialects in the southeast and is quite regular, as it occurs in all words containing the clusters. Examples: сестра /sestra/, сесра /sesra/ 'sister'

Epenthesis of т /t/, д /d/ in the clusters ср /sr/, зр /zr/

This phonetic change, the opposite of the one in the previous feature, is found in the southwest. Examples: сряда /srʲada/, стряда /strʲada/ 'Wednesday'

The voiceless velar fricative

In some dialects such a consonant does not exist and in others its use is restricted to certain positions. [x] is weakest in word-initial and intervocalic seem

to be the weakest positions. It may be replaced by another consonant (f, w, h) or by nothing at all. In some dialects the loss of x is compensated by lengthening of the preceding vowel. Examples: хляп /xlʲap/, ляп /lʲap/ 'bread'; страх /strax/, стра /stra/ 'fear'

The voiceless labiodental fricative

There is no such consonant in the native vocabulary of Bulgarian. It was introduced through borrowings, mostly from Greek. It is still absent in a number of dialects, westernmost and easternmost. The form фтурна /fturna/, recorded in one of the villages, is an interesting folk etymology — it was interpreted as coming from the verb туря /turja/ 'put' and the preposition в /v/ 'in'. In some southeastern dialects, the voiceless bilabial fricative is used, especially before the vowel у /u/. Examples: фурна /furna/, вурна /vurna/, хурна /xurna/, хурна /hurna/, фурна /ɸurna/ 'oven'

Loss of в /v/ before rounded vowels

This is a good example for the universal preference for combination of more contrasting rather than similar sounds. It is found consistently in the eastern part of the eastern dialects. The change occurs both in stressed and unstressed syllables. Examples: вол /vol/, ол /ol/ 'ox'

Prothetic v before rounded vowels

This change is opposite to the one in in the previous feature. Interestingly, there is at least one dialect (the Erkech dialect in the easternmost part of Stara planina mountain) in which the two lexical sets are completely reversed: all words beginning with v followed by rounded vowel are pronounced without the в /v/, and all words beginning with stressed о are pronounced with a prothetic в /v/. Examples: огън /ogɤn/, вогън /vogɤn/ 'fire'

Voicing of obstruents

In Bulgarian all voiceless obstruents but /x/ are paired with voiced obstruents. The phonemic contrast between voiceless and voiced consonants is possible before vowels, sonorants, and в /v/. The ambivalent position of v is worth noting: it belongs to the obstruents in being paired with a voiceless counterpart ф /f/ and in being subject to final devoicing and voicing assimilation to following voiceless obstruent. On the other hand, it is similar to the sonorants in allowing both voiceless and voiced consonants before it, or in other words, in not triggering voicing assimilation in the preceding voiceless obstruent. It was not paired with a voiceless consonant in Old Bulgarian and was in the group of the sonorants. Examples: джоп /ʤop/, джоб /ʤob/ 'pocket'; жиф /ʒif/, жив /ʒiv/ 'alive'; офца /oftsa/, овца /ovtsa/, осца /ostsa/ 'sheep'

The preposition and the prefix в /v/

In many dialects, western and northeastern, the preposition and the prefix в /v/ are replaced by у /u/. The preposition, on the other hand, may appear doubled. Examples: влизам /vlizam/, улизам /ulizam/ 'enter'; в /v/, ф /f/, въф /vɤf/, ъф /ɤf/ 'in'

Various assimilations and dissimilations

The word много /mnogo/ 'much, many' barely exists in this form in the dialects. Examples: офца /oftsa, осца /ostsa/ 'sheep'; едно /edno/, ено /eno/ 'one'; много /mnogo/, млого /mlogo/, мого /mogo/, ного /nogo/, фного /fnogo/ 'much, many'; тъмно /tɤmno/, тъвно /tɤvno/ 'dark'

Nonsystematic changes

These are changes found in individual words. The words below are old comparative degrees. Examples: бързо /bɤrzo/, бърже /bɤrʒe/ 'quickly'; вече /veʧe/, век'е /vece/ 'already'

Morphophonemic alternations

The formation of the so called secondary imperfective verbs in many cases involves vowel and consonant alternations. It seems that some dialects favor suffixes, while other dialects, western and southeastern, favor alternations, but a lot of further investigation is needed. In the word плащам /plaʃtam/ 'pay' the reflex of *tj is found but the alternation is suspended in some dialects. Examples: влизам /vlizam/, влазам /vlazam/, влявам /vlʲavam/ 'enter'; връщам /vrɤʃtam/, вращам /vraʃtam/ 'give back'

Different verbal ending

The verbal ending -мо /-mo/ for all tenses is found in the dialects close to the western border. Examples: бяхме /bʲaxme/, бехмо /bexmo/ 'were-we'

Different suffixes

The words камък /kamɤk/ 'stone', ечемик /eʧemik/ 'barley, ремък /remɤk/ 'strap' and пламък /plamɤk/ 'flame' belonged to the n-stem nouns in Old Bulgarian and developed in three different ways in the dialects. All forms have large and well defined areas. The word ечемик /eʧemik/ may differ from the rest. Камък /kamɤk/ was selected because it is the most available word. Examples: камък /kamɤk/, камик /kamik/, камен /kamen/ 'stone'

Various forms

The variants of each of these words are derived from a common Old Bulgarian form, so in spite of the seemingly great phonetic differences they cannot be interpreted as lexical variation. Examples: вие /vie/, ви /vi/, ве /ve/ 'you'; тогава /togava/, тогас /togas/, тегай /tegaj/ 'then'

Stress

The stress in most Bulgarian dialects is free (it may fall on any syllable in polysyllabic words) and movable (it may be moved to other syllables in inflection and word-formation). Examples: вино /ˈvino/, вино /viˈno/ 'wine'

3. Linguistic analysis. In this section we present the results of the aggregate analysis of the data described in the previous section. We first calculated the distances between each pair of corresponding words using a modified Levenshtein algorithm, which also resulted in the calculation of the distances between the sites. After that, the distances obtained between the sites were analyzed using multidimensional scaling and hierarchical clustering.

3.1. Levenshtein algorithm. 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. For example, the distance between two word transcriptions in Figure 2 is 2: [e] has to be inserted between [b] and [r] and [ə] has to be replaced by [ɑ]. The algorithm is also directly used to align two sequences, as can be seen in Figure 2.

b	-	r	ə	n	e
b	e	r	ɑ	n	e
	1		1		

Fig. 2. Levenshtein distance between these two strings is 2

The Levenshtein algorithm thus results in the calculation of the distance between each pair of strings. The distance between two sites is the mean of all word distances calculated for those two sites. We note that using the mean Levenshtein distance over a large sample of pronunciations effectively aggregates over a large number of individual segment differences, the basis of most isoglosses. The final result is a distance matrix which contains the distances between each

two sites in the data set. Brett Kessler [15] was the first to use Levenshtein distance in order to calculate the linguistic distance between dialects. Later it was successfully applied to many other languages. An overview of the application of the Levenshtein algorithm in dialectology can be found in [24].

3.2. Data processing. Before applying the Levenshtein algorithm, all word transcriptions were preprocessed in the following way:

- First, all diacritics and suprasegmentals were removed from word transcriptions. In order to process diacritics and suprasegmentals, they should be assigned certain weights appropriate for the specific language that is being analyzed. Since no study of this kind was available for Bulgarian, diacritics and suprasegmentals were removed, which resulted in the simplification of data representation. For example, [u], [u:], [ˈu], and [ˈuː] counted as the same phone. Thus, all words were represented as series of phones which are not further defined. The result of comparing two phones can be 1 or 0; either they match or they do not. For example, pair [e, ɛ] counts as different to the same degree as pair [e, i]. Although it is linguistically counterintuitive to use less sensitive measures, [8] has shown that in the aggregate analysis of dialect differences a more detailed feature representation of segments does not improve the results obtained by using simple phone representation.

- All transcriptions were aligned based on the following principles: a) vowels may align with vowels; b) consonants may align with consonants, semivowels [j], [w] and sonorants. No other alignments are allowed. The alignments were carried out using the Levenshtein algorithm described in the previous subsection.

The final result is a distance matrix which contains the distances between each two sites in the data set. This distance matrix was further analyzed using multidimensional scaling (MDS) and the clustering algorithm weighted pair group method using arithmetic averages (WPGMA) that are explained below.

3.3. Multidimensional scaling. Multidimensional scaling is a dimension-reducing method used in exploratory data analysis and a data visualization method, often used to look for separation of the clusters [18]. The goal of the analysis is to detect meaningful underlying dimensions that allow the researcher to explain observed similarities or dissimilarities between the investigated objects. It displays the structure of distance-like data geometrically by attempting to arrange "objects" in a space within a certain small number of dimensions, which, however, accord with the observed distances. As a result, dissimilar objects are plotted

far apart from each other, while similar objects are close to one another. This enables us to "explain" the distances in terms of underlying dimensions. It has been used in linguistics and dialectology since [1].

3.4. Hierarchical clustering algorithms. Cluster analysis is the process of partitioning a set of objects into groups or clusters [20]. The goal of clustering is to find structures in data by finding objects that are similar enough to be put in the same group and by identifying distinctions between the groups. The data in each subset share some common trait — often proximity according to some defined distance measure. Clustering methods can be divided into hierarchical and partitional clustering. Hierarchical clustering algorithms produce a set of nested partitions of the data by finding successive clusters using previously established clusters. This kind of hierarchy is represented with a dendrogram — a tree in which more similar elements are grouped together (see below). Hierarchical clustering algorithms can be described by the following scheme formalized by [12]:

- Estimate pairwise distances
- Put information on distances into matrix
- Find the shortest distance in the matrix
- Fuse the two closest points
- Calculate the distance between the newly formed node and the rest of the nodes (matrix updating algorithms)
- Repeat until there are no more nodes to be fused

Based on the way in which the distances between a newly formed node and the rest of the nodes are calculated, [11] identify seven different algorithms. In this study we applied WPGMA in order to find grouping in the data. See [29] for a discussion of alternatives.

WPGMA calculates the distance between the two clusters, i.e. between a newly formed node and the rest of the nodes, as the average of distances between all members of two clusters. The clusters that fuse receive equal weight regardless of the number of members in each cluster.

$$d_{k[ij]} = \left(\frac{1}{2} \times d_{ki}\right) + \left(\frac{1}{2} \times d_{kj}\right)$$

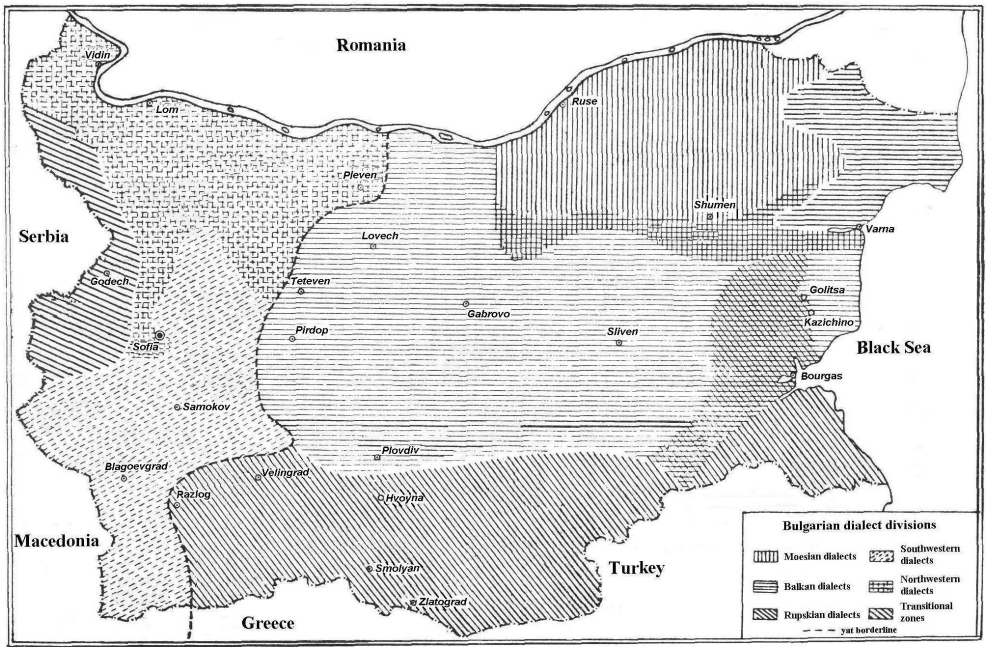

Fig. 3. The traditional map of Bulgarian dialects

In this formula i and j are the two closest points that have just been fused into one cluster$[i, j]$, and k represents all the remaining points (clusters). Because all clusters receive equal weights, objects in smaller clusters are more heavily weighted than those in the big clusters. As a result there is no distortion during the fusion of a large group of objects with the small group of objects. This enables us to detect dialect areas that contain a small number of sites, unlike with some other hierarchical clustering algorithms.

3.5. Traditional scholarship. Traditional scholarship [40] divides the Bulgarian language area into two main groups: western and eastern. The border between these two areas is so-called 'jat' border that reflects different pronunciations of the old Slavic vowel 'jat'. It goes from Nikopol to the north, near Pleven and Teteven down to Petrich in the south (the bold dashed line in Figure 3). Stojkov divides each of these two areas further into three smaller dialect zones, which can also be seen on the map in Figure 3. In the west, he distinguishes Southwestern dialects, Northwestern dialects and the (Serbian) transitional zone. In the east, according to [40], there are Moesian, Balkan and Rupskian dialects.

This 6-fold division is based on the variation of different phonetic features. No lexical or syntactic differences were taken into account.

3.6. Results.

3.6.1. MDS. The results of applying multidimensional scaling to the data analyzed using the Levenshtein algorithm can be seen in the MDS plot in Figure 4. Here, the first two extracted dimensions are plotted against axes x and y. On the right, all three extracted dimensions are represented by different shades of red, green and blue.

The first three dimensions represented in Figure 4 explain 98% of the variation found in the data. The first extracted dimension explains 80% of the variation, and the second extracted dimension an additional 16% of the variation. The MDS plot reveals that there are two separate groups in the data. This division of sites follows the x axes. By putting all sites on the MDS map, we can see that this division of sites corresponds to the division of the country to the east and west (dark green and red on the MDS map). This division explains 80% of the variation of the data, making it the most important division of the Bulgarian dialect area. The second dimension, which explains 16% of the variation, divides the sites along the y axis. This division of sites corresponds to the separation of the Rodopi area in the south from the rest of the country (light green color on the MDS map). It should also be noted that the southern group of varieties is much more heterogeneous than the rest of the data. It lies between the two much more homogeneous groups without clear separation to any of the two. No other dialect areas were detected using this method.

3.6.2. WPGMA. The result of the WPGMA analysis is a dendrogram that can be seen in Figure 5.

The dendrogram shows that at the highest level of hierarchy there is a very short branch that separates a 2- and a 3-way split of the data. The two-way split of the data follows the 'jat' border, while in the 3-way division varieties from the Rodopi area form a separate group. In the maps in Figure 6 we can see the 2- and 3-way split produced by the WPGMA algorithm. These findings conform both with the traditional division of the sites as given by [40] and with the results obtained by MDS. In order to confirm the results obtained by WPGMA, we have also analyzed the data using two other hierarchical clustering algorithms, namely the unweighted pair group method using arithmetic averages and Ward's method. These two methods gave exactly the same 2- and 3-way divisions of the sites as WPGMA. The findings of the MDS, as well as the perfect agreement of three different hierarchical clustering algorithms, confirm that the main division of the sites goes along the 'jat' border dividing the Bulgarian dialect area into the east

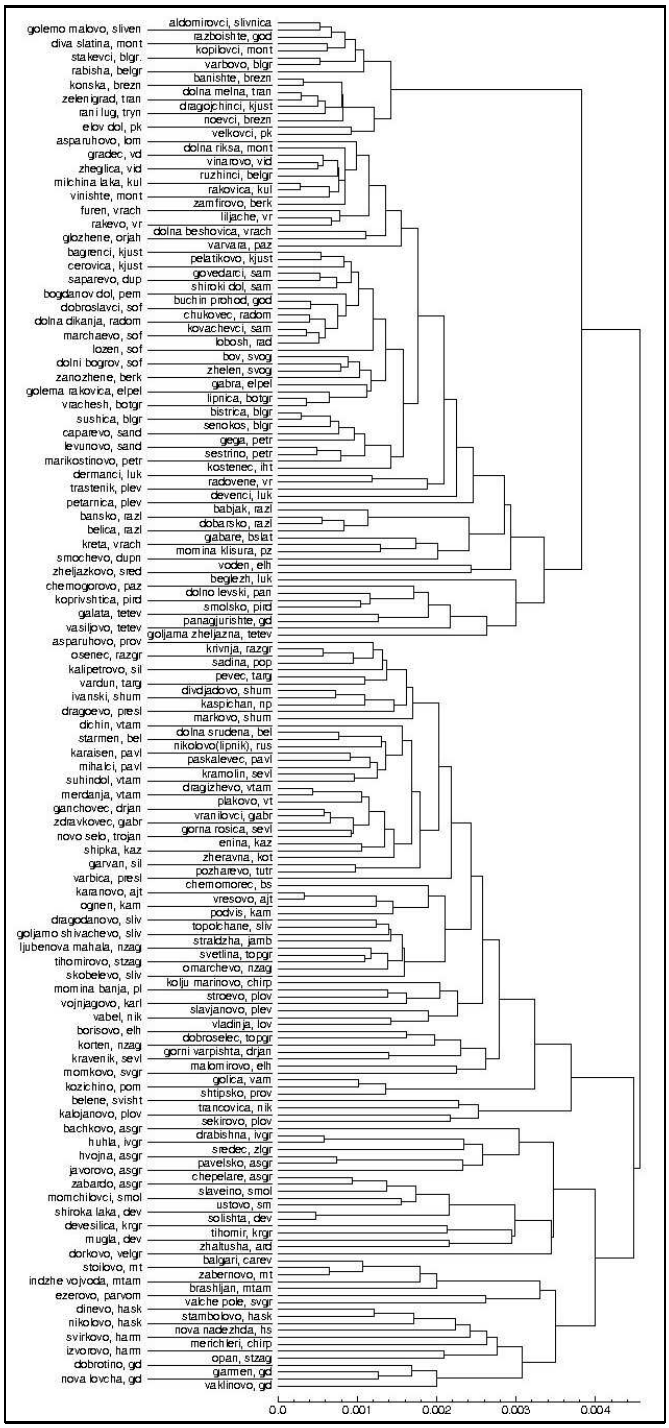

Fig. 5. The dendrogram produced by WPGMA

and west. The third dialect area that can be asserted with high confidence is the Rodopi area in the south. The varieties in this area are much more heterogeneous than the varieties found in the east or west. According to all three hierarchical clustering algorithms the Rodopi area is grouped with the eastern varieties, although the MDS plot, as well as the dendrogram in Figure 5, show that this dialect area lies between the east and west without clear separation from the two.

Since traditional scholarship defines six dialect areas, we have also performed a 6-way clustering of the data. The results can be seen in Figure 7. Apart from the already mentioned 2- and 3-way division of the sites, we can also see that a group of sites around the border with Serbia forms a separate group. No other groups were found in the data. The reason for this could be due either to the simplified representation of the data described in Subsection 3.2, or to the skewed feature distribution present in our data set. They may also point to shortcomings in the traditional studies. At the moment we are investigating the distribution of the features responsible for the traditional division of sites in our data set. However, 2- and 3-fold divisions of sites can be asserted with high confidence, which was also found in our previous study of the same data set [29].

3.6.3. Noisy clustering. Since hierarchical clustering algorithms are known for their instability [11], we have also performed noisy clustering in order to check the stability of the WPGMA results. Noisy clustering is a procedure in which small amounts of random noise are added to matrices during repeated clustering. The main purpose of this procedure is to reduce the influence of outliers on the regular clusters and to identify stable clusters. As shown in [26] it gives results that nearly perfectly correlate with the results obtained by bootstrapping—a statistical method for measuring the support of a given edge in a tree [3]. The advantage of the noisy clustering, compared to bootstrapping, is that it can be applied on a single distance matrix. The result of noisy clustering is a dendrogram which shows the confidence of every branch in the dendrogram. It ranges between 50 and 100 per cent, since we recognize only groups recognized 50 per cent or more of the times.

Applied to our data set, noisy clustering has confirmed that there are two relatively stable groups in the data: eastern and western. However, the dendrogram obtained by applying noisy clustering to the whole data set shows low confidence for the two-way split of the data, between 52 and 60 per cent. After removing the southern (Rodopi) villages from the data set, we obtained dendrograms that confirm a two-way split of the data along the 'jat' border with much higher confidence, ranging around 70 per cent. These values are still not very high. In order to check the reason for the influence of the southern varieties

on the noisy clustering we examine an MDS plot in two dimensions with cluster groups marked by colours. In Figure 8 we can see the MDS plot of 6 groups produced by the WPGMA algorithm. The MDS plot reveals two homogeneous groups (the green and red dots vs. dark blue and magenta dots) and a third, more diffuse, group that lies at a remove from them. The third group of sites represents the southern group of varieties, colored light blue and yellow, and is much more heterogeneous than the rest of the data. Closer inspection of the MDS plot in Figure 4 also shows that this group of dialects has a particularly unclear border to the eastern dialects, which could explain the results of the noisy clustering applied to the whole data set. More detailed discussion of the instability of our data set can be found in [29].

4. An information theoretic perspective. In this section we present the results of applying *information theory* to the dialect pronuncuation data in order to calculate the distances between language variaties.

The term *information* is used in a wide range of scientific fields. In *information theory* [2], it is defined on the basis of the *probability* of an element in a given data set. The probability of an element is estimated as the proportion of occurrences of that element vs. the whole number of elements:

$$p(x) = \frac{\text{number of occurences of x}}{\text{number of elements in the data set}} \tag{1}$$

If the elements of a data set are not distributed in a uniform way, the probability of the element will vary. Rarer elements carry more information than more frequent ones.

Based on this observation, it is possible to calculate the amount of information of an element z [19]:

$$I(p(z)) = -log_2 p(z) \tag{2}$$

where $p(z)$ is the probability of element z.

Obtaining the logarithm base two of the probability of an element converts the result into the binary system. Other logarithms are possible, but would not change the scale of the relations between the information amounts of the elements. Because the probability is always ≤ 1 the result has to be multiplied with -1 to get a positive value.

From the formula in 2 it follows that the information of an element *decreases* when the probability of the element *increases*. A rarer element carries more information than a more frequently occurring one. If an element has the

probability of 1, its information is 0 because if a data set contains only one different kind of elements, there is no *surprise* and with that also no information ($log_2 1 = 0$).

By summing up the information of every element in a data set X the absolute amount of information $I(X)$ in that data set is obtained:

$$I(X) = -\sum_{i=1}^{n} log_2 p(z_i) \tag{3}$$

where n is the number of elements in the data set. The entropy of the data set is the weighted average of the information of the individual elements.

On the basis of the probabilities of single elements, the amount of information in a whole dialect data set can be calculated. In the next step, for every site in the data set the amount of information is calculated. These values can be mapped to a symmetric matrix which can be used for different analyses and visualizations. Note that the information matrix represents the complete data set and not only a specific element.

Linguistically, we note that the amount of information is larger in varieties with larger segment inventories which are more uniformly distributed.

4.1. Analysis and visualization. The methods shown here result in *similarity matrices*. These can be analyzed and visualized in many ways: clustering and multidimensional scaling are common methods (see above). The following maps show another method, the interval algorithm. For more information on the interval algorithm, see [5, p. 93 ff.]. Both maps were created with the VDM software,[1] using the same interval algorithm (MinMWMax) with the same parameters (site 1 as reference point, 12 classes).

On the map in Figure 9 there is a clear distinction between the eastern and the western part of Bulgaria, on the borders to Serbia and Turkey are transitional dialects and the mountains in the south, the Rodopi, show a heterogeneous distribution of dialects.

5. Conclusions. Different quantitative techniques show that the main split in Bulgarian dialect area follow the 'jat' border and divides the country into the western and eastern language areas. These findings conform to the traditional dialect division as presented in [40]. Multidimensional scaling and WPGMA also

[1]For a detailed description of the VDM software, see http://ald.sbg.ac.at/dm/Engl/default.htm

reveal a third group of varieties—the Rodopi area in the south of the country. This area is much more heterogeneous compared to the rest of the country. These varieties are not clearly separated from the western and eastern varieties as shown by noisy clustering and MDS plot of clusters (Figure 8). WPGMA analysis has also revealed the fourth cluster at the border with Serbia. This group is marked as the Transitional zone in the map given in [40]. However, no other methods have confirmed this as a separate group in the data. Unlike in the traditional atlases, we did not find evidence of the separation of the western dialects into the Northwestern and Southwestern groups. The same holds for the division of the eastern dialects into Moesian and Balkan groups. The reasons for this could be in the simplified representation of the data where all diacritics and suprasegmentals were removed. It is also possible that some of the features responsible for the traditional divisions of sites are not present in our data set. This issue is being investigated at the moment. A third possibility is that some of the dialect divisions present in the traditional atlases do not have strong basis in the linguistic features, but were rather result of more general consideration on the part of the dialectologists. By closely examining the distribution of the features responsible for the traditional divisions in our data set and by applying other quantitative techniques to the data we hope to answer this question.

6. Acknowledgments. The work presented in the paper is supported by a grant from the Volkswagen Foundation awarded jointly to the University of Tübingen, the University of Groningen, the University of Sofia, and the Institute for Parallel Processing, Bulgarian Academy of Sciences.

REFERENCES

[1] Black P. Multidimensional scaling applied to linguistic relationships. In: Cahiers de l'Institut de Linguistique Louvain, **3** (1973), Expanded version of a paper presented at the Conference on Lexicostatistics, Montreal, University of Montreal.

[2] Cover T., J. Thomas. Elements of Information Theory. Wiley-Interscience, 2006.

[3] Felsenstein J. Inferring Phylogenies. Sinauer Associates, Inc., 2004.

[4] Goebl H. Dialektometrie: Prinzipien und Methoden des Einsatzes der Numerischen Taxonomie in Bereich der Dialektgeographie. Wien: Osterreichischen Akademie der Wissenschaften, 1982.

[5] Goebl H. Dialektometrische Studien. Anhand italoromanischer, rätoromanischer und galloromanischer Sprachmaterialien aus AIS und ALF, 1984.

[6] Gorov G. Strandzhanskijat govor. *Bylgarska dialektologija. Prouchvanija i materiali*, **1** (1962), 13–164.

[7] Gylybov L. Govoryt na s. Dobroslavci, Sofijsko. *Bylgarska dialektologija. Prouchvanija i materiali*, **2** (1965), 3–118.

[8] Heeringa W. Measuring Dialect Pronunciation Differences using Levenshtein Distance. PhD thesis, University of Groningen, Groningen, 2004.

[9] Hitov H. Rechnik na govora na s. Radovene, Vrachansko. *Bylgarska dialektologija. Prouchvanija i materiali*, **9** (1979), 223–342.

[10] Hristov G. Govoryt na s. Nova Nadezhda, Haskovsko. *Izvestija na Instituta za bylgarski ezik*, **4** (1955), 177–253.

[11] Jain A., R. Dubes. Algorithms for Clustering Data. Prentice Hall, Englewood Cliffs, New Yersey, 1988.

[12] Johnson S. Hierarchical clustering algorithms. *Psychometrika*, **32** (1967), No. 3, 241–254.

[13] Kabasanov S. Govoryt na s. Momchilovci, Smoljansko. *Izvestija na Instituta za bylgarski ezik*, **4** (1955), 5–101.

[14] Keremidchieva S. Govoryt na Ropkata (rodopska gramatika). Microprint, 1993.

[15] Kessler B. Computational dialectology in Irish Gaelic. In: Proceedings of the seventh conference on European chapter of the Association for Computational Linguistics, San Francisco, CA, USA, 1995, Morgan Kaufmann Publishers Inc., 60–66.

[16] Kochev I. Grebenskijat govor v Silistrensko. Izdatelstvo na BAN, 1966.

[17] Kovachev S. Trojanskijat govor. *Bylgarska dialektologija. Prouchvanija i materiali*, **4** (1968), 161–235.

[18] Legendre P., L. Legendre. Numerical Ecology. Elsevier, Amsterdam, second edition, 1998.

[19] Lyre H. Informationstheorie. Eine philosophisch-naturwissenschaftliche Einführung, UTB, 2002.

[20] Manning C., H. Schütze. Foundations of Statistical Natural Language Processing. MIT Press, Cambridge, MA, 1999.

[21] Mladenov M. Jatovata granica v svetlinata na novi danni. Slavistichen sbornik, Sofija, 1973, 241–256.

[22] Mladenov M. Ihtimanskijat govor, Izdatelstvo na BAN, 1966.

[23] Mladenov M. Leksikata na ihtimanskija govor. *Bylgarska dialektologija. Prouchvanija i materiali*, **2** (1967), 3–196.

[24] Nerbonne J. Data-driven dialectology. *Language and Linguistics Compass*, **3** (2009), No. 1, 175–198.

[25] Nerbonne J., Jr. Kretzschmar. Progress in dialectometry: Toward explanation. *Literary and Linguistic Computing*, **21** (2006), No. 4, 387–398.

[26] Nerbonne J., P. Kleiweg, F. Manni, W. Heeringa. Projecting dialect distances to geography: Bootstrap clustering vs. noisy clustering. In: Proc. of the 31st Annual Meeting of the German Classification Society.(Eds. Christine Preisach, Lars Schmidt Thieme, Hans Burkhardt, Reinhold Decker) Data Analysis, Machine Learning, and Applications. Berlin: Springer, 2008, 647–654.

[27] Popivanov G. Sofijskijat govor. *Sbornik na Bylgarskata akademija na naukite*, **XXXIV** (1940), 209–326.

[28] Popov K. Govoryt na s. Gabare, Beloslatinsko. *Izvestija na Instituta za bylgarski ezik*, **4** (1955), 103–176.

[29] Prokić J., J. Nerbonne. Recognizing groups among dialects. *International Journal of Humanities and Arts Computing*, 2009.

[30] Ralev L. Govoryt na s. Vojnjagovo, Karlovsko. *Bylgarska dialektologija. Prouchvanija i materiali*, **8** (1977), 3–198.

[31] Séguy J. La relation entre la distance spatiale et la distance lexicale. *Revue de Linguistique Romane*, **35** (1971), 335–357.

[32] Stojchev K. Tetevemskijat govor. Sbornik za narodni umotvorenija, **XXXI** (1915).

[33] Stojkov S. Izbrani ezikovedski trudove. Universitetsko izdatelstvo Sv. Kliment Ohridski, 2008.

[34] Stojkov S. Kratyk osvedomitelen vyprosnik za prouchvane na bylgarskite govori. Sofija, 1954.

[35] Stojkov S. Programa za sybirane na rechnikovi materiali ot bylgarskite narodni govori. *Ezik i literatura*, **2–3** (1955), 92–96.

[36] Stojkov S. Govoryt na s. Govedarci. *Izvestija na Instituta za bylgarski ezik*, **4** (1955), 255–320.

[37] Stojkov S. Osnovnoto dialektno delenie na bylgarski ezik. Slavjanska filologija, t. III, Sofija, 1963, 105–119.

[38] Stojkov S. Govoryt na s. Mugla, Devinsko. *Izvestija na Instituta za bylgarski ezik*, **20** (1971), 3–90.

[39] Stojkov S. Bylgarska dialektologia. Izdatelstvo na BAN, 3rd edition, 1993.

[40] Stojkov S. Bulgarska dialektologiya. Sofia, 4th ed., 2002.

[41] Stojkov S, M. Mladenov. Proekt za ideografski dialekten rechnik na bylgarskija ezik. *Bylgarski ezik*, **2** (1969), 155–170.

[42] Stojkov S., M. Mladenov. Upytvane za prouchvane leksikata na mesten govor. Izdatelstvo na BAN, Sofija, 1971.

[43] Umlenski I. Kjustendilskijat govor. Izdatelstvo na BAN, 1965.

[44] Videnov M. Godechkiat govor. Izdatelstvo na BAN, 1978.

A. List of words

аз /az/ ‘I’	агне /agne/ ‘lamb’	бели /beli/ ‘white-plural’
берът /berɤt/ ‘pick up-they’	беше /beʃe/ ‘was, were’	бране /brane/ ‘picking up’
брашно /braʃno/ ‘flour’	бързо /bɤrzo/ ‘quickly’	бяхме /bʲaxme/ ‘were-we’
вежда /veʒda/ ‘eyebrow’	вече /veʧe/ ‘already’	вечер /veʧer/ ‘evening’;
видях /vidʲax/ ‘saw-I’	вие /vie/ ‘you-plural’	вино /vino/ ‘wine’
влизам /vlizam/ ‘enter’	вода /voda/ ‘water’	вол /vol/ ‘ox’
време /vreme/ ‘time’	връх /vrɤx/ ‘peak’	връщам /vrɤʃtam/ ‘give back-I’
вчера /vʧera/ ‘yesterday’	във /vɤv/ ‘in’	вълк /vɤlk/ ‘wolf’
вълна /vɤlna/ ‘wool’	вънка /vɤnka/ ‘outside’	вътре /vɤtre/ ‘inside’
вятър /vʲatɤr/ ‘wind’	глава /glava/ ‘head’	гладен /gladen/ ‘hungry’
говедо /govedo/ ‘bovine animal’	горе /gore/ ‘upstairs’	гости /gosti/ ‘guests’
градът /gradɤt/ ‘the city’	грозде /grozde/ ‘grapes’	дадоха /dadoxa/ ‘gave-they’
две /dve/ ‘two’	двор /dvor/ ‘yard’	ден /den/ ‘day’
дера /dera/ ‘flay-I’	пера /pera/ ‘launder-I’	десет /deset/ ‘ten’
дете /dete/ ‘child’	джоб /ʤob/ ‘pocket’	днес /dnes/ ‘today’
добре /dobre/ ‘well-adverb’	долу /dolu/ ‘downstairs’	дошъл /doʃɤl/ ‘has come-he’
дъжд /dɤʒd/ ‘rain’	дълбок /dɤlbok/ ‘deep’	дъно /dɤno/ ‘bottom’
дърво /dɤrvo/ ‘tree’	един /edin/ ‘one-masculine’	едно /edno/ ‘one-neutrum’
език /ezik/ ‘tongue’	ечемик /eʧemik/ ‘barley’	желязо /ʒelʲazo/ ‘iron’
жена /ʒena/ ‘woman’	жив /ʒiv/ ‘alive’	живели /ʒiveli/ ‘lived-they’
жълт /ʒɤlt/ ‘yellow’	жътва /ʒɤtva/ ‘harvest’	звезда /zvezda/ ‘star’
здрав /zdrav/ ‘healthy’	земя /zemʲa/ ‘Earth’	зет /zet/ ‘son/brother-in-law’
ѝ /i/ ‘her-dative’	им /im/ ‘them-dative’	име /ime/ ‘name’
камък /kamɤk/ ‘stone’	ключ /kljuʧ/ ‘key’	кое /koe/ ‘which-neuter’
кон /kon/ ‘horse’	кръв /krɤv/ ‘blood’	къде /kɤde/ ‘where’
лесно /lesno/ ‘easily’	леща /leʃta/ ‘lentils’	майка /majka/ ‘mother’
месец /mesets/ ‘month’	месо /meso/ ‘meat’	млякото /mlʲakoto/ ‘the milk’
много /mnogo/ ‘much, many’	мъж /mɤʒ/ ‘man’	мъже /mɤʒe/ ‘men’
мъжът /mɤʒɤt/ ‘the man’	наше /naʃe/ ‘ours’	неделя /nedelʲa/ ‘Sunday’

неще /neʃte/ ‘does not want’	нещо /neʃto/ ‘something’	нея /neja/ ‘her-accusative’
ние /nie/ ‘we’	носят /nosʲat/ ‘carry-they’	нощ /noʃt/ ‘night’
няма /nʲama/ ‘there is no’	овца /ovtsa/ ‘sheep-singular’	овце /ovtse/ ‘sheep-plural’
овчар /ovʧar/ ‘shepherd’	овчари /ovʧari/ ‘shepherds’	огън /ogɤn/ ‘fire’
онези /onezi/ ‘those’	орех /orex/ ‘walnut’	пека /peka/ ‘bake-I’
сека /seka/ ‘chop-I’	пепел /pepel/ ‘ash’	петел /petel/ ‘rooster’
петък /petɤk/ ‘Friday’	плащам /plaʃtam/ ‘pay-I’	понеделник /ponedelnik/ ‘Monday’
пръч /prɤʧ/ ‘billy-goat’	първият /pɤrvijat/ ‘the first’	път /pɤt/ ‘road’
пясък /pʲasɤk/ ‘sand’	река /reka/ ‘river’	ръка /rɤka/ ‘hand’
ръце /rɤtse/ ‘hands’	се /se/ ‘one self’	сега /sega/ ‘now’
седя /sedʲa/ ‘sit-I’	сестра /sestra/ ‘sister’	сирене /sirene/ ‘cheese’
сол /sol/ ‘salt’	средата /sredata/ ‘the middle’	сряда /srʲada/ ‘Wednesday’
старец /starets/ ‘old man’	страх /strax/ ‘fear’	сух /sux/ ‘dry’
събота /sɤbota/ ‘Saturday’	сърп /sɤrp/ ‘sickle’	със /sɤs/ ‘with’
такъв /takɤv/ ‘such’	твой /tvoj/ ‘yours’	това /tova/ ‘this’
тогава /togava/ ‘then’	тъмно /tɤmno/ ‘dark’	тънко /tɤnko/ ‘thin’
трева /treva/ ‘grass’	утре /utre/ ‘tomorrow’	ухо /uxo/ ‘ear’
фурна /furna/ ‘oven’	хляб /xlʲab/ ‘bread’	хоро /xoro/ ‘chain dance’
хубав /xubav/ ‘beautiful-m’	хубаво /xubavo/ ‘beautiful-n’	цял /tsʲal/ ‘whole’
чакат /ʧakat/ ‘wait-they’	червен /ʧerven/ ‘red’	черен /ʧeren/ ‘black’
череша /ʧereʃa/ ‘cherry’	чета /ʧeta/ ‘read-I’	чешма /ʧeʃma/ ‘fountain’
човек /ʧovek/ ‘human’	ще /ʃte/ ‘will’	я /ja/ ‘her-accusative’
ябълка /jabɤlka/ ‘apple’	ябълки /jabɤlki/ ‘apples’	яйце /jajtse/ ‘egg’
яйца /jajtsa/ ‘eggs’	ям /jam/ ‘eat-I’	ядеш /jadeʃ/ ‘eat-you’

Jelena Prokić, John Nerbonne
Information Science/
Humanities Computing
Faculty of Arts
University of Groningen
Oude Kijk in 't Jatstr. 26
Postbus 716
NL 9700 AS Groningen
The Netherlands
e-mail: j.prokic@rug.nl
e-mail: j.nerbonne@rug.nl

Vladimir Zhobov
Faculty of Slavic Studies
Sofia University "St. Kliment Ohridski"
15 Tsar Osvoboditel Blvd
1504 Sofia, Bulgaria
e-mail: zhobov@slav.uni-sofia.bg

Kiril Simov, Petya Osenova
Linguistic Modelling Laboratory
Institute for Parallel Processing
Bulgarian Academy of Sciences
Acad. G. Bonchev Str., Bl. 25A
1113 Sofia, Bulgaria
e-mail: kivs@bultreebank.org
e-mail: petya@bultreebank.org

Thomas Zastrow, Erhard Hinrichs
Seminar for Sprachwissenschaft
Faculty of Linguistics
University of Tübingen
Wilhelmstr. 19–23
72074 Tübingen, Germany
e-mail: zastrow@sfs.uni-tuebingen.de
e-mail: eh@sfs.uni-tuebingen.de

Received June 8, 2009
Final Accepted August 26, 2009