BULGARIAN DIALECTS IN THRACE: A DIGITAL APPROACH

Ronelle Alexander

University of California, Berkeley, USA
ralex@berkeley.edu

Περίληψη

Το άρθρο περιγράφει μια συσχετιστική βάση δεδομένων που περιέχει δείγματα ομιλίας από βουλγαρικές διαλέκτους, εκ των οποίων ένας μεγάλος αριθμός προέρχεται από τις νότιες περιοχές που συνορεύουν με την Ελλάδα. Η δομή της βάσης δεδομένων επιτρέπει στους ερευνητές να έχουν άμεση πρόσβαση σε τμήματα των ηχογραφήσεων πεδίων, οι οποίες έγιναν από μια βουλγαρο-αμερικανική ομάδα, καθώς και σε μεταγραφές και μεταφράσεις αυτών των τμημάτων. Τα τμήματα είναι σχολιασμένα, ώστε να επιτρέπουν στους ερευνητές να πραγματοποιούν αναζήτηση με διάφορους τρόπους. Παρόλο που αυτή η συγκεκριμένη βάση δεδομένων περιορίζεται σε δείγματα ομιλίας από τη Βουλγαρία, η δομή της μπορεί να προσαρμοστεί και να χρησιμοποιηθεί γενικότερα για την ανάλυση σλαβικών διαλέκτων. Παρουσιάζεται μια σύντομη περιγραφή της δομής της βάσης δεδομένων και εκφράζεται η ελπίδα ότι οι ερευνητές σε όλη την εθνογλωσσική περιοχή της Θράκης θα καταβάλουν προσπάθειες να χρησιμοποιήσουν τη συγκεκριμένη δομή με σκοπό να παρουσιάσουν και αναλύσουν δείγματα διαλεκτικού υλικού από τις δικές τους περιοχές.

Λέξεις-κλειδιά: Πομακική, βουλγαρική, διάλεκτος, ψηφιακή βάση δεδομένων.

The geographical region of Thrace encompasses southeastern Bulgaria, far northeastern Greece, and all of present-day European Turkey. Speakers of essentially the same Slavic-based dialect group inhabit all of these regions. Both they and their speech, however, are referred to by different terms, depending on the region, and on their religious affiliation. The term Pomak is widely used to refer to Muslim speakers of these dialects, in all three countries. In the last few years, it has been argued that the speech of Pomaks in Greece and Turkey should be elevated to the status of a minority language within these states (no reference is made to non-Muslim Slavic speakers in these regions).[1] Within Bulgaria, however, the

1. Bojadžiev 1991 (and especially the maps, 212-274) gives a thorough survey of the distribution of these dialects in the Greek and Turkish regions of Thrace (referred to by Bojadžiev as western [Aegean] and eastern [Adrianopolian] Thrace, respectively) in the period before the population redistribution of the 1920s. Although many of these dialects were spoken by non-Pomaks, there is no way to infer from his description how many of these dialects are still spoken in these regions today.

term Pomak refers to religion alone, and the speech of Bulgarian Pomaks is viewed within the complex of the dialectal range of the Bulgarian language. This view is justified by the fact that the dialects of Muslim (Pomak) Bulgarians do not differ in any significant way from those of their immediate Christian neighbors.

The question of whether Pomak is a language of its own or a dialect of Bulgarian (spoken either within Bulgaria or in diaspora) is not a linguistic one but is rather dependent on numerous cultural, historical and political factors; and it is not my intent to comment on any of these. I wish rather to focus on instances of speech itself as it functions within communities of these speakers, and on an innovative means I have developed to make natural speech samples more available to the broader public. Dialectology as a discipline is concerned with cataloguing such instances of speech over broad regions and analyzing these speech samples in different ways. As in any science, one of the most basic goals is establishing parameters in order to make the appropriate classification of types and subtypes. These regional speech samples are also analyzed with respect to differentiation on the level of any of a number of linguistic traits, and the results of these analyses are valuable both for historical and typological studies. Within Bulgaria, the discipline of Bulgarian dialectology has produced a great wealth of scholarship on the basis of Bulgarian dialectal materials.[2] Nearly all of this material is in print form, and available only in Bulgarian; in addition, the highly analytic focus of these materials means that the connection with actual speech in its communicative context is often lost.

In recent years I have worked together with a team of colleagues to create a means by which interested scholars and laymen can have access to actual dialect speech as recorded in the village context. Because of advances in digital techniques, and the ubiquitous presence of the internet, it is now possible to present dialectal materials in this way. The result of our work is a relational database bearing the title *Bulgarian Dialectology as Living Tradition*, henceforth referred to by the acronym BDLT (Alexander and Zhobov 2011-2016). Although the site is still a work in progress, it has been open to the public since 1 May 2016. Thus, those portions of the site for which data entry has been completed are now readily available, and can be accessed at http://bulgariandialectology.org/.

As noted above, the first goal of the BDLT project was to return the focus of dialectology to the original source material, living village speech. This was done by placing the focus on actual speech samples as recorded in the field rather than on classification or analysis of separate features. In concrete terms, our approach to this goal consisted in the decision to construct the entire website around these individual speech samples. Each of the 181 speech samples chosen as illustrative

2. These include the several volumes of the *Bŭlgarski dialekten atlas*, volumes in the two series bearing the names *Trudove po bŭlgarska dialektologija* and *Bŭlgarska dialektologija*, and numerous other individual articles and monographs.

of the dialectal variation across Bulgaria occupies its own page on the site, where it is available both as an audio file and in transcription. The second goal of the project, developed very soon after the first, was to make these speech samples accessible to the larger world beyond the confines of Bulgaria, or Bulgarian-oriented scholarship. We approached this goal through the decision to transcribe these samples twice – once using the Cyrillic transcription system customary in Bulgarian dialectology and once again using a modified IPA transcription,[3] and then to translate the samples into English and provide annotations for the majority of words within the speech samples. As the project grew in scope, the annotation system was developed to a very broad extent, with the goal of allowing researchers to search the database on a number of different parameters.

The BDLT database includes material from all the major dialects spoken within the current political boundaries of Bulgaria, including large amounts of material from Bulgarian Thrace. Unfortunately, the database does not include any speech samples from Slavic speakers living in those areas of Thrace which are located in Greece and Turkey. My original (and quite idealistic) goal had indeed been to gather speech samples from all the Balkan regions where Slavic is spoken (thus not only from Bulgaria but also from Turkey, Greece, Macedonia, Albania, Romania, and southern Serbia), and to present them all within the same framework. Due to simple human limitations, this goal was not possible to realize, and the database is limited to speech samples representing dialects spoken within the current political boundaries of Bulgaria.

Yet this original goal – of expanding the collection to include speech samples from all of Balkan Slavic (that is, Slavic-based dialects spoken in all Balkan countries) – is not completely out of the range of possibility. Even though this will not be accomplished by me (nor, most likely, will it be accomplished in my lifetime), such a goal is in principle realizable. This is because the BDLT site was constructed using the open source content management system Drupal. The Drupal network consists of a set of modules, freely available on the internet, which can then be customized by users to build a database according to their own specifications. The architect of the BDLT database utilized these several modules to construct a database which displays speech samples representing a broad geographical expanse and provides means to search the data on a number of a different parameters.[4] This basic structure can be adapted to serve a similar purpose for speech samples in other regions, and the team which has produced the BDLT database is willing to share this

3. The transcription uses IPA letters for vowels. The modification consists in the use of the symbols č, š, ž for post-alveolar sounds, and the placement of the accent mark over the accented vowel.

4. The site was created by Quinn Dombrowski in 2011; supplementary design features were added by Cammeron Girvin between 2013 and 2016; I am grateful to them both as well as to the many Berkeley undergraduate students who assisted in data entry through Berkeley's Undergraduate Research Apprentice Program (URAP).

basic structure with other scholars who have access to similar data and who have the ability and desire to prepare and input dialect material from their own regions in a similar manner.

With this hope in mind, I provide here a brief description of the site's capabilities and value to researchers, and of the ways in which data were prepared and entered into the database in order to create such a research tool. Readers are encouraged to visit the site online and browse it for themselves in order to get a more direct sense of its contents and capabilities.

The home page of the BDLT site includes a map which identifies each of the 70 locations in Bulgaria where speech samples were recorded. With very few exceptions, all material was recorded by the same team of field workers. This is not a necessity, of course, but it is an obvious advantage in data preparation that those with direct memory of the field experience are also those doing the transcription and analysis. In this instance, the recordings were made by an international team working in Bulgaria. The original idea of joint Bulgarian-American fieldwork was developed in the late 1980s by one Bulgarian and one American scholar; the Bulgarian scholar was Maksim Mladenov and I was the corresponding American scholar. Mladenov unfortunately was able to take part only in one of the early field trips (in August 1992) before his untimely death in the fall of that year. After that point the primary Bulgarian field workers were Georgi Kolev and Vladimir Zhobov, with Todor Bojadžiev working in an advisory capacity; I continued to be the primary American team member.

Most of the data on the site were gathered in two lengthy field trips. During the first, which took place in the fall of 1993, the focus was exclusively on recording, and speech samples were recorded in 21 different villages. During the second, which took place in the summer of 1996, the focus was both on recording and on the training of students (both American and Bulgarian).[5] Only nine villages were visited during this second trip but the amount of material recorded in two of these (the two Erkeč villages of Kozičino and Golica) was vastly greater than that recorded in all the other sites.[6] Shorter trips had been made both before this time (to one village in 1991 and to five villages in 1992) and after it (to two villages each in 2002 and 2013). Once it was decided to create the database by choosing representative selections from each of these 40 villages, the data set was supplemented by material which Zhobov had recorded earlier in 1986 (14 villages) and 1988 (nine villages), some in conjunction with Kolev. Finally, in order to round out the coverage, material was added from seven more locations; these

5. Funding for both field trips was provided by the International Research and Exchanges Board (IREX), to whom gratitude is extended here.

6. This expedition resulted in a volume of papers which not only described the expedition but also included original research by each of the team members, faculty and students alike, derived from the field material gathered on the Erkeč dialect. See Alexander and Zhobov 2004.

were from recordings by students or colleagues of Kolev and Zhobov (one village each in 1993, 1998, 2003, 2011 and 2012, and two villages in 2000).

The full amount of recorded material fills some 200 hours of cassette tapes. Rather than posting the entirety of this material on the site, we decided to select excerpts and then to process each excerpt in detail. Each of the selected excerpts, chosen after careful listening to the entire set of recordings, needed to satisfy two criteria. First, it needed to contain as many of the specific linguistic elements characteristic of the dialect in question as possible. Second, it needed to constitute a well-formed instance of discourse whose content either conveyed material of ethnographic interest or constituted a narrative of some sort (legend, folktale, or account of personal experience). This aspect of the work was accomplished by Bulgarian team members, who eventually chose 186 excerpts as representative of these 70 different Bulgarian villages. As the numbers indicate, most villages are represented by more than one excerpt; indeed, only 15 of the 70 villages are represented by a single excerpt. These excerpts, the sum of which totals somewhat less than thirteen hours of natural speech, form the core of the database. Henceforth the term «text» will refer to one of these excerpts.

The website itself was created by American team members, and is maintained on the server of the Berkeley Language Center. Its home page identifies the 70 geographical locations, with a link to each individual location. Each Location page gives information about the village in question, and contains links to the text (or texts) recorded in that location. The Contents page is the best way to locate individual texts, however, for two reasons. The first and most obvious is that it lists all 186 texts on the same page; and the second is that it gives a synopsis of basic informational items about each text. Each of these informational items is listed in a separate column. The first column gives the name of the text, with each text named after the village it was recorded in, followed by a number if that village is represented by more than one text. The second column names the dialect group to which each text belongs. There then follow three columns describing the length of the text: one gives the number of "tokens" (words spoken by informants); another gives the temporal length of the audio segment, and the final one gives the number of lines in the text. The central column gives a brief synopsis of the thematic content of the text, and the final columns give the status of data entry with respect to those items which are still in the process of being entered. Another useful factor about the listing on the Contents page is that it can be sorted on any of these parameters except the one noting the time of the audio.

The Contents page summarizes the status of texts now in the database. Getting each text to this point, however, was a complex process. The audio cassette was first digitized, and the segment in question was prepared as an mp3 file for eventual posting to the site. The text was then transcribed using the modified IPA system (that is, in Latin letters), and translated into English. In general, work on both these processes was straightforward, but it sometimes required more con-

sultation (with phonetician colleagues as to the transcription and with various dialectal lexica as to the translation). Once both the transcription and the translation were verified to be as correct as we could make them, the text was divided up into numbered lines. Although the division into lines makes the text look somewhat artificial, it was necessary to do this so that individual items in each text could be precisely located for the purposes of eventual data retrieval. Each line could contain a maximum of twelve tokens, though most are shorter than this. Line breaks in the text were made wherever possible in correspondence with natural breaks in syntax or speech rhythm. Line breaks also needed to be made in such a way as to correspond to natural semantic and syntactic breaks in the English rendering of the text.

These were the principles followed with reference to longer passages by the same speaker. Some line entries are quite short, however. This is because each line is marked as to the speaker, and it was thus necessary to make a new line each time there was a shift of speaker. If the only «speech» was a single word or syllable, or even laughter or a coughing sound, this was nevertheless transcribed. We made the decision to transcribe texts in this fashion not only because we felt it necessary to reproduce fully the content of the texts, but also in recognition of the fact that these interviews constitute natural conversations between individuals and that the transcripts could be valuable for conversational analysis as well as more for traditional linguistic analysis. Indeed, since one of our primary goals was to elicit longer stretches of natural speech, we purposefully constructed our interviews to resemble natural conversation. Some of our texts, in fact, consist of conversations between informants (and in several instances an actual argument). Although the overlapping speech that almost always occurs in such instances is much harder to transcribe, the resulting naturalness of the speech situation made it worthwhile, in our opinion, to include such texts. As concerns earlier texts in which only the speech of the informant appears on tape (due to the fact that a directional mike was used and investigator regularly paused the machine in order to economize on tape space and battery power), we reconstructed the portions of the conversation spoken by the investigator. Although these segments are transcribed on the page, and given a number in the overall sequence of lines, the graphic notation on the page indicates clearly that they do not appear on the audio.

Once the text had reached this stage of preparation, with transcription and translation complete and with numbered line breaks finalized, it was run through a macro which replaced each of the Latin-based symbols with the corresponding Cyrillic symbol used by Bulgarian dialectologists to represent that phonetic sound. Now each line of each text was represented in three separate forms: one in the Latin-based IPA transcription, one in the Cyrillic transcription customarily in use by Bulgarian dialectologists, and one in English translation. Each line with the same content, whether in translation or in one of the two transcriptions, is identified by the same number, by the same alphabetic code identifying the

speaker, and by the same time-code identifying the point in the audio file where the spoken line begins. Each full text, therefore, appears in three different versions.

In response to the fact that users will have different goals in accessing the texts, we have structured the page devoted to each text so as to display different combinations of the three different versions. Each of these is called a «View». The view consulted by Bulgarian users (who do not need the text to be translated, and – if they are dialectologists – do not need annotations) is called the «Cyrillic Line Display». This view displays the running text in Cyrillic transcription only. The view consulted by external users who are interested primarily in content is called the «Line Display». Here, each line text is presented in two forms. The IPA-based transcription is given first, followed by the English translation directly underneath it. The absence of other information in this view allows for distraction-free reading of the content of the text. The third view, called «Glossed Text», has the most information, and this is the view that will be consulted by users interested in the texts for analytic purposes. This view in fact contains four lines of information for each line of text. First is the English translation, and second is the primary text, in the IPA-based transcription. The third line glosses each token (word) with grammatical tags and (for lexical items) the literal English gloss, and the fourth line gives the standard Bulgarian lemma (in Cyrillic) for each token.

Here is an example of the three views, quoting the first line of the text Drabišna 1, from the Thracian sub-group of Bulgarian dialects.[7]

Example 1: Text displays

Cyrillic Line Display

1(a) [0:01] бу̀бички hрàнеhме трѝ hрàнеh трѝ кутѝи бу̀бички

Line Display

1(a) [0:01] bùbički hrànehme trì hràneh trì kutìi bùbički
We used to feed [= keep] [silk]worms. I kept three boxes of silkworms.

Glossed Text

1(a) [0:01] We used to feed [= keep] [silk]worms. I kept three boxes of silkworms.

bùbički	**hrànehme**	**trì**	**hràneh**	**trì**	**kutìi**	**bùbički**
silkwormpl f	feed 1pl impf	three	feed 1sg impf	three	boxpl f	silkwormpl f
бубичка	**храня**	**три**	**храня**	**три**	**кутия**	**бубичка**

7. In all citations from the website, boldface indicates that the forms in question are hyperlinks on the site, taking one to a separate page entry.

In addition to these tags on these two bottom lines, there are two additional sets of tags that are not shown directly on the Glossed Text page. One of these identifies «linguistic traits» of interest. The traits themselves are spelled out on a long list accessible on the Site Information page, but the tags that are appended to individual tokens are highly abbreviated. One can see which tags have been added to any one token by linking on the token itself, which takes one to the individual page for that token. This page displays not only all the relevant tags for that token but also identifies any other lines in the database (in any text) where this token also appears. Among the «relevant tags» for that token are any of the linguistic trait tags assigned to it, and it is here that one can expand the abbreviated form of the tag in order to see its full meaning. Here, for instance, is the form in which information is listed on the page for the token *hràneh*.

Example 2: Token page display

hràneh
Meaning: feed
Aspect: I
Verb form: impf
Lexeme: **хрaня**
Person: 1sg
Linguistic trait: **jat end s0 pal0 c0 /e/**
Lines where hràneh appears
Drabišna 1:1 – bùbički hràneh̀me trì hràneh trì kutìi bùbički
Drabišna 1:51 – lež'àh nevr̀nkə bùbičkį̀ tvà trì kutìi bùbički hràneh

To the see full description of a linguistic trait, one clicks on the abbreviation, which takes one to a full path description of the linguistic trait and all its components.

Example 3: Expanded single linguistic trait[8]

/e/ REALIZATIONS → jat m s p c R | Historical Slavic Vowels →| Vowels → | Phonology →| unstressed | s Stress | not before palatalizing environment | p Palatalizing environment | ending |m Morpheme Type | not before /c/ | c Affricate /c | CONDITIONS →

This linguistic trait, which is composed of a number of components, means that the word contains a segment identified as the Historical Slavic Vowel called «jat» (categorized under Vowels, which category itself falls under Phonology), that this segment is found in a Morpheme Type identified as «ending», that its Condition with respect to Stress is «unstressed», that its Condition with respect to

8. Symbols which are underscored in the sample appear on the website encircled; for typological reasons it was not possible to reproduce these encircled symbols here.

Palatalizing Environment is that it does not occur «before palatalizing environment», that its Condition with respect to the Affricate /c/ is that it does not occur before /c/, and that its Realization in this segment is «/e/». The function of such complex traits will be discussed below in the section about searching the database.

A further set of tags identifies elements of «thematic content». Rather than being linked with individual tokens, these tokens are appended to lines of text. These tags identify segments of text which speak of a particular topic. If one wishes to see for any one line which thematic content tags have been appended to it one can link to the page for the line, which itself is accessed via the page of any of the tokens in that line. For instance, the example given above for the page of the token *hràneh* lists the lines in which this token appears in the form of a hyperlink. Clicking on the hyperlink for the first of these lines will bring up the following page, which itself includes hyperlinks to each of the tokens, to the full text itself, and to the thematic content tag.

Example 4: Line page display

Drabišna 1:1

Full line: bùbički hrànehme trì hràneh trì kutìi bùbički

Cyrillic full line: бубички hранеhме три hраneh три кутии бубички

Translation: We used to feed [= keep] [silk]worms. I kept three boxes of silkworms.

Text: **Drabišna 1**

Timecode: 0:01

Token: **bùbički**

hrànehme

trì

hràneh

trì

kutìi

bùbički

Thematic content: **silkworms**

The reason for the existence of all these tags, of course, is to make the database a research tool, and to allow users to search the database for items within the speech samples that are of particular interest, either from the point of view of linguistic structure or thematic content. There are five different types of searches.

The first of these (the Wordform Search) allows one to choose from among the many different «word form tags» and find all tokens that have been marked by any single tag or any combination of tags. One can also search simply by English gloss or by standard Bulgarian lemma («lexeme»); and one can include either of these in the combination of tags to be searched. A list of all the available wordform tags can

be found within the document entitled «How To Use the Site» under the Site Information tab; the list is given both in alphabetical order, and by category. The tags can also be seen, arranged by category, on the actual search page. A search is initiated by simply marking the relevant tags (and/or entering the Bulgarian lemma and/or English gloss in the relevant boxes) and hitting Apply. The results of such a search give one not only the token in question and its place in the database (text name and line number) but also the entire line (in transcription and in translation) so that one has some sense of the context in which the token occurs; their geographicla distribution is also shown on a map.

For instance, if one wanted to search the database for all instances within the excerpted texts of infinitive forms, one would choose the grammatical tag «inf» and hit Apply. Since Bulgarian dialects have generally lost the infinitive, only a few items will be found (for instance, a total of ten items are retrieved for this particular search). Nevertheless, it is very useful to be able to see them all at a glance. The following (one of the ten items retrieved) illustrates the form in which the retrieved data are listed.

Example 5: Wordform search result

Malevo/Hsk 1
Token: **kazə** say inf P **кажа**
Malevo/Xsk 1 Line 68
ne mògə ti kàzə
I can't say.

In this citation format, the several forms are listed alphabetically by text name (the form cited above occurs after two forms from the text Kolju Marinovo 1 and before a single form from text Stojkite 1). The entry then lists the token in question and all its tags, followed by the text name and line number, and the context of the token (the full line in transcription and translation). One can thus both see the immediate context of any one token and, by means of the line number, track the token back to its original recorded context.

If one wishes to search for a combination of tags, one simply ticks all the relevant boxes before hitting Apply. If, for instance, one wanted to find all instances of the verb *храня* in the imperfect tense, one would enter the form *храня* in the Lexeme box, and mark the word-form tag «impf». Each of the individual tokens which satisfies this combination of tags would appear in the form seen above.

The second of these searches (the Lexeme Search) allows one to search by Bulgarian standard lemma to find all the phonetic implementations of any one standard form. One can also search by English gloss on this page as well. For instance, if one did not want to see every instance of the lexeme *време* but only those with the meaning «weather», one could search by the English gloss. This

page also includes a list of «lexical traits»; these are tags that have been added to the entries for lexemes. This search is of particular interest to those interested in ethnolinguistic aspects of the texts, since they allow one to search for loans from Turkish, Greek, and other sources. For instance, if one searches for instances of «Turkish loan», one receives a very long list of alphabetized lexemes. Each lexeme is followed by a list of every phonetic implementation of that lexeme found within the database. Any one lexeme also has its own page; this page identifies all lexical traits associated with that lexeme and then gives a compact list of all the phonetic implementations, this time without separating them out by text. This allows one to see this phonetic information in two different forms: if one wants simply to see at a glance the range of variation for any one lexeme, one goes straight to the individual Lexeme page, but if one wants to check the forms by village, one does this from the Lexeme search page.

The third of these searches (the Linguistic Trait Search) allows one to locate forms marked by any of a number of linguistic traits of interest. Here too, one can find a list of all available linguistic trait tags within the document entitled «How To Use the Site». Unlike the Wordform Search, however, where all the options are listed on the page and one need only check the desired one(s), the choices here are made through a hierarchically arranged series of drop-down menus. One chooses first whether to search in the domain of Phonology, Morphology, Syntax, Lexemes, or Pragmatics; this starting menu also gives one the option to specify within a search Conditions or Realizations. Subsequent menus narrow the search until one is finally ready to hit the Apply button. When the results of this search are displayed, one first sees a map, with pins identifying the geographical location of all the individual results. There follows a list of the results, which is a simplified version of that in the Wordform search. For instance, if one searched for all tokens satisfying the complex linguistic trait cited in Example 3 above, one would see among the cited examples the following:

Example 6: Linguistic trait search dataform result

Drabišna 1
hràneh
Drabišna 11
Bùbički hrànehme trì hràneh trì kutìi bùbički

Some of the traits are simply stated. The trait named «elision of unstressed vowel» (found by working through the hierarchy «Phonology → Vowels → Elision →») is self-explanatory, as is the trait «diminutive» (found directly under «Lexemes →»). More complex traits are those such as the one seen in Example 3 above. Such complex tags are mostly used to search for reflexes of various historical vowels, consonants, or consonant sequences. The «complexity» here refers to the fact that one can either search simply for all reflexes of such a historical

vowel, or one can restrict the search in various ways. These include the addition of various Conditions which would be relevant to the specification of that vowel. For instance, there are only two conditions that would make a difference in the description of the current replacements of the historical vowels called «jers»: these are the specification of morpheme type (preposition, affix, root, ending, or definite article) and the specification of stress (stressed or unstressed). To get the most information about the historical vowel called «jat», however, one can choose to specify not only these two conditions but also the fact of whether or not it is in the presence of a palatalizing environment or after the affricate /c/. In addition, one can specify in a search the fact of Realization (by asking for all instances in which that historical vowel, specified by whichever conditions one wishes to choose, appears in the form of a specific modern realization, such as /ɛ/).

The fourth search is the Thematic Content. Here, one specifies a content term from a long list, which again is available in two forms, both as a listing by thematic categories and subcategories, and as an alphabetic listing of all tags. In the first instance, one can look under a major heading (such as Agriculture, Family, Foodways, or the like) to see what individual topics are catalogued; and in the second one can see whether a particular term one has in mind appears in the alphabetical ordering of the entire list. The results are displayed in the form of chunks of text which speak to the particular topic. The list is available both in thematic subdivisions and as a single alphabetic list.

The fifth search, the Phrase Search, is devoted to grammatically significant groups of tokens whose meaning must be tagged at the phrase level (and not at the level of any individual token, as in the first three searches). Like the Wordform search, one chooses from among a number of different tags which describe possible phrase components (such as reflexive, negation, types of clitic sequences, types of phrasal accentuation, or particular types of word order). The results appear in a format similar to those for the Wordform Search, and a listing of all the phrase types and their components can be found within the document «How To Use the Site».

The listings within this lengthy document comprising word-form tags, linguistic trait tags, thematic content tags and phrase component tags, are the result of a great deal of analysis and thought on the part of the site creators. Indeed, it is important to note that the lists were not taken over from any ready-made reference source but were rather constructed on the basis of the texts themselves. That is, it was the material in the speech samples (the texts which form the core of the database) which determined which tags should be including on the lists, and in what form. The set of word-form tags, for instance, includes most of the common grammatical tags that are relevant for a Slavic language, but it also includes tags such as «proximal» and «distal» under the category of deixis, since these are relevant in several regions of Balkan Slavic. It also includes, under the category of pragmatics, tags for the conversational function called «backchannelling» (when

the listener makes a remark whose function is to assure the interlocutor that s/he is paying attention) and the type of conversational-pragmatic marker called «ostensive» (when the speaker points in some way to something relevant for the conversation). The set of linguistic trait tags is considerably more complex, but it was also compiled using the same principle. In this instance, however, it was necessary to create separately the hierarchical organization which is seen on the search page and in the reference list under Site Information. Similarly, it is not only the set of phrase tags that was compiled anew, but also the entire structuring of their components into the several phrasal categories that organize the search.

Entry of all this data into the database is time-consuming, but straightforward once all the structural decisions are taken. Since the central focus of the site is the text, this is where the basic data entry takes place. Texts having been previously divided up into lines, it is then necessary to enter each line as a unit. At this stage of data entry one enters the name of the text, the line number, the identification code of the speaker, the three forms of the line itself (IPA transcription, Cyrillic transcription, English translation), and any thematic content codes. One must then move to the next stage, which is to enter each token within that line as a separate unit, together with its various tags (wordform tags and English glosses). This must be done in order to key each of the tokens to the specific line of which they are component parts, and to have them appear in order underneath the full form of the line. Other tags associated with tokens, namely linguistic traits and lexemes (standard lemma forms) are entered from a separate page. Finally, it was of course necessary to create all the separate tags: word-form tags, lexeme tags, thematic content tags, phrase component tags and the linguistic trait tags. Simplex linguistic trait tags are stand-alone, but each unique combination of complex tags must be individually created before it can be assigned to a token.

The resulting database is intended both as a research tool for scholars and as a general cultural tool, and it has proven to be of value in both these spheres. The first goal is more quantifiable, in that research results tend to appear in print (or electronically) and to quote sources. The second goal is more intangible, but no less real. Some speakers of regional dialects in Bulgaria wish to leave their rural past behind them and take on city ways as soon as possible. A very large number of them, however, retain an attachment with their traditional roots and a number of these individuals have expressed their deep satisfaction at the idea that someone outside their village (and even outside their country) recognizes the value of their way of speech, and sees it as «living tradition».

It is certain that the same feeling holds true for Pomak speakers in Greece and Turkey (and for non-Muslim Slavic speakers there as well). Though they may learn Greek and Turkish for necessary external reasons, their native speech is still the Slavic-based «Pomak», and they are deeply attached to it. Scholarly efforts which have been taken thus far to record this speech are of great value. I hope in

this brief report to have proposed yet another model by which future scholars could bring more recognition to this «living tradition» in the Balkans.

REFERENCES

Alexander, R. and V. Zhobov (2011-2016). *Bulgarian Dialectology as Living Tradition*. Retrieved from http://bulgariandialectology.org/.

Alexander, R. and V. Zhobov (eds.) (2004). *Revitalizing Bulgarian Dialectology*. UCIAS Edited Volumes, uciaspubs/editedvolumes/2. Retrieved from https://escholarship.org/uc/item/9hc6x8hp.

Bojadžiev, T. (1991). *Bŭlgarskite govori v zapadna (Belomorska) i iztočna (Odrinska) Trakija*. Sofia: Universitetsko izdatelstvo «Kliment Ohridski».

Edited by Christina Markou, Evangelia Thomadaki, Xenophon Tzavaras

Perspectives of ethno-linguistic research on the «historical» area of Thrace

K. & M. STAMOULIS PUBLICATIONS
THESSALONIKI

Χριστίνα Μάρκου–Ευαγγελία Θωμαδάκη–Ξενοφών Τζαβάρας
(Επιμέλεια)

Η ΙΣΤΟΡΙΚΗ ΘΡΑΚΗ
ΥΠΟ ΤΟ ΠΡΙΣΜΑ ΤΗΣ ΕΘΝΟΓΛΩΣΣΟΛΟΓΙΑΣ

ΕΚΔΟΤΙΚΟΣ ΟΙΚΟΣ Κ. & Μ. ΣΤΑΜΟΥΛΗ
ΘΕΣΣΑΛΟΝΙΚΗ

ΕΚΔΟΤΙΚΟΣ ΟΙΚΟΣ
Κ. & Μ. ΣΤΑΜΟΥΛΗ
Π. Π. Γερμανοῦ - Ἰω. Μιχαὴλ 2
546 22 ΘΕΣΣΑΛΟΝΙΚΗ
Τηλ. 2310-264.748
69.46.461.460
email: anstamoulis@hotmail.com - xarpantidisioannis@gmail.com
ΘΕΣΣΑΛΟΝΙΚΗ 2021

ISBN: 978-960-656-059-0