American Contributions to the 16th International Congress of Slavists

Belgrade, August 2018

Volume 1: Linguistics

Edited by

Christina Y. Bethin

Bloomington, Indiana, 2018

SLAVICA

Cover design: Tracey Theriault

Technical Editor: Phillip Weirich

This volume was published in July 2018.

ISBN 978-0-89357-486-4

Library of Congress Cataloging-in-Publication Data

Names: International Congress of Slavists (16th : 2018 : Belgrade, Serbia) | Bethin, Christina Y. (Christina Yurkiw) editor. | Kornblatt, Judith Deutsch editor.
Title: American contributions to the 16th International Congress of Slavists, Belgrade, August 2018 / [volume 1] edited by Christina Y. Bethin, [volume 2] edited by Judith Deutsch Kornblatt.
Other titles: American contributions to the sixteenth International Congress of Slavists, Belgrade, August 2018
Description: Bloomington, Indiana : Slavica, 2018. Contents: v. 1. Language -- v. 2. Literature.
Identifiers: LCCN 2018008491| ISBN 9780893574864 (v. 1 : pbk.) | ISBN 9780893574880 (v. 2 : pbk)
Subjects: LCSH: Slavic philology--Congresses.
Classification: LCC PG11 .I5 2018 | DDC 491.8--dc23
LC record available at https://lccn.loc.gov/2018008491

Slavica Publishers
Indiana University
1430 N. Willis Drive
Bloomington, IN 47404-2146
USA

[Tel.] 1-812-856-4186
[Toll-free] 1-877-SLAVICA
[Fax] 1-812-856-4187
[Email] slavica@indiana.edu
[www] http://www.slavica.com/

Table of Contents

Digital Processing of Bulgarian Dialects:

New Approaches, Synchronic and Diachronic

Ronelle Alexander

The goal of this paper is to survey the ways in which the availability of digital sound files, together with the easy access of the internet, have been utilized in studies of Bulgarian dialects as of the time of writing (late 2016).

It is first necessary to outline some background and also to point out that the concept "Bulgarian dialects" has two different real-world referents. In the work of Bulgarian dialectologists of the post-war period until around the mid 1980s, and in the work of most scholars outside Bulgaria, this phrase refers to dialects within the boundaries of the Republic of Bulgaria. But for Bulgarian dialectologists at the Institute for the Bulgarian Language (within the Bulgarian Academy of Sciences) from the 1990s onward, and for certain foreign scholars who share their views, this phrase refers to dialects across a much broader area, embracing not only Bulgaria but also Macedonia, southeastern Serbia, and Slavic dialects spoken within Albania and Greece. The majority of scholars outside Bulgaria refer to this complex as Balkan Slavic; they recognize that it forms a cohesive unit but are unwilling to identify that unit by the name of a single ethnic group or political entity.

My goal here in pointing this out is not to take a stand on the issue of names but rather to draw attention to the fact that there exist two sets of research aids, both of which bear the name *Bŭlgarski dialekten atlas*, but which differ considerably both in scope and in method. The first of these (and for many, still the classic set) covers the four quadrants within the political boundaries of Bulgaria, and was published between 1964 and 1981 (IBE 1964, 1966, 1975, 1981). Two additional volumes were subsequently published in the same format, each treating dialect groups outside the boundaries of Bulgaria. One is devoted to dialects in northern Greece (Ivanov 1972) and the other to dialects in far eastern Serbia (Bozhkov 1986).

Christina Y. Bethin, ed. *American Contributions to the 16th International Congress of Slavists, Belgrade, 2018*. Bloomington, IN: Slavica, 1–15.

The four “classic” volumes were undertaken as a single massive project, with the help of Soviet scholars at the outset. After a thorough historical, linguistic, and demographic investigation of 4,680 villages (all of which were determined to have old and homogeneous populations), 1,877 villages were selected and a corps of field investigators was sent out to administer the same questionnaire in each village. Although the published atlases contain cross-referenced lists of all the villages visited and commentary on each of the maps, the indubitable core of the atlas is the maps. Each of the four volumes contains a large number of maps, with each map devoted to a single phonological, morphological, or lexical trait; every village on each map is identified with a symbol denoting the answers to the relevant questionnaire item recorded in that village. The two subsequent atlases are compiled on the same model, and although the number of villages visited in each case is significantly smaller, the coverage in terms of maps is nevertheless relatively comparable. That is, volume I (southeastern Bulgaria) contains 277 maps, volume II (northeastern Bulgaria) contains 290 maps, volume III (southwestern Bulgaria) contains 314 maps, volume IV (northwestern Bulgaria) contains 393 maps, and Bozhkov’s atlas of far eastern Serbia contains 334 maps. The total in Ivanov’s atlas of northern Greece (232 maps) is somewhat smaller, but that atlas is presented as the first volume in a series (which, unfortunately, was not continued).

All the material in these atlases is drawn from actual interviews with informants in the field and can therefore be assumed to be based on living, natural speech recorded in the village context. The fact that each symbol on each map is associated with an identifiable single village and that each map is accompanied by commentaries which often give more information about data from those villages, makes this complex of atlases a very valuable research tool, both in synchronic and diachronic terms. In the first instance the maps give detailed information about the dialectal lexicon, and in the second they give detailed information about the dialectal reflexes of Old Bulgarian/Late Common Slavic vowels and consonant sequences, both in root vowels and in grammatical endings. In both instances the presentation allows one to follow any one statement back to its source, both with respect to the village where it was recorded and the questionnaire item which elicited it. One drawback of the *Atlas* is that the maps are selective: not all topics are covered in all four volumes, which means that one frequently cannot juxtapose maps from different volumes of the *Atlas* maps to make definitive analytic statements for the entirety of Bulgaria on particular topics. Another (more minor) potential drawback lies in the questionnaire method and the fact that not all investigators were equally experienced, raising the possibility that some answers might not be fully grounded in natural village speech. The sheer volume of the collected data, however, more or less balances this out in the larger picture.

The second set of atlases, all of which bear the subtitle *Obobshtavasht tom*, is quite different. First, as noted above, the coverage extends over the broader Balkan Slavic area. Second, the maps do not identify particular locations but rather consist of isoglosses. A preliminary volume, with only 25 roughly drawn maps (IBE 1988), was issued on the occasion of the International Congress of Slavists held in Sofia that year, but the subsequent volumes are much more detailed and contain maps of high graphic quality. The first (IBE 2001) is identified as three volumes in one (I-II-III, *Fonetika, aktsentologiia, leksika*) and the second (IBE 2016) is identified as volume IV, *Morfologiia*. The 2001 volume contains 172 maps in the phonology section, 88 maps in the accentuation section, and 108 maps in the lexical sections; and the 2016 volume (on morphology) contains 145 maps. The maps in all these volumes are printed in bright colors, which make it quite easy to see the broad scope of any one phenomenon at a single glance. At the same time, the maps are much less useful for follow-up research in that the actual source material—the data underlying the decisions as to isogloss placement—is usually recoverable only with difficulty. For regions within Bulgaria proper, the cited source material comprises the same archival material used to produce the original four-volume *Atlas*, which means that with dedication (and some guesswork) one can track the source of the relevant information. For regions outside Bulgaria proper, however, the cited source is the single phrase “materials from written sources,” which means there is no way one can track the specific source.

The largest region outside Bulgaria, of course, is that of Macedonia. Work on the *Macedonian Dialect Atlas* has been underway for many years, but as yet there is no announced date for publication. However, a great number of analytical studies have appeared, most accompanied by maps; the author of the majority of them is the unofficial “dean” of Macedonian dialectology, Bozhidar Vidoeski.[1] Upon Vidoeski’s death in 1998, his research team collected these articles and published them under the title *Dijalektite na makedonskiot jazik*, in three volumes (Vidoeski 1998, 1999a, 1999b). They also published *Fonoloshki bazi na govorite na makedonskiot jazik*, a volume of detailed phonological descriptions of 33 different Macedonian dialects, whose internal consistency, they observe, is assured by the fact that all the transcriptions were made by the same individual—Vidoeski himself (Vidoeski 2000a). Finally,

[1] Some of the lexical maps have appeared in individual monographs devoted to lexical issues, according to the report presented to the 2006 meeting of the All-Slavic Linguistic Atlas (Gajdova 2007).

the team gathered from Vidoeski's archive a large number of texts of transcribed connected speech from different Macedonian dialects and published them as a chrestomathy under the title *Tekstovi od dijalektite na makedonskiot jazik* (Vidoeski 2000b).

The latter volume is of great value, since most dialectological resource works—atlases and lexicons—focus on individual words and do not allow one to study linguistic phenomena above the word level, such as word order, narrative structure, or functional sentence perspective. It is true that some individual dialect descriptions, published both in Macedonia and Bulgaria, also include transcribed texts. There are also a number published individually under the rubric "Dialect texts" in the major Bulgarian linguistics journal *Bŭlgarski ezik*. Further, there are the relatively rare instances where phrases or sentences are cited within dialect descriptions or dialect lexicons in order better to illustrate particular dialectal phenomena. Overall, however, the attention of dialectological research tools remains at the word level.

A tremendous amount of work has been done to provide all these tools, both in the field with informants and then preparing the material for publication. Since nearly all this work was done before entry into the digital age, nearly all the data are available in print form only. Because users of these data cannot listen to the original recordings, they must therefore trust that the transcription is correct. In the case of those data collected before the introduction of sound recording, users are at the complete mercy of the investigator's physical perception of the sounds he heard (in many instances only once), in the knowledge that this investigator had been required, in a single listening, both to follow the speech narrative and to monitor it for linguistic traits. Indeed, the possibility of rewinding a tape and listening multiple times to a recording not only resulted in more accurate transcriptions but also ended up revealing greater dialectal diversity than had previously been known, a fact which suggests that a considerable amount of (largely unconscious) normalization had been taking place in those earlier days.

The ready availability of sound recordings of actual dialectal speech was a great step forward: not only did it allow investigators to make more accurate transcriptions after the fact, but it also allowed them to make more complex analyses, for instance, with the aid of spectrograms. Now, with the ease of digitalization and the ready availability of sound files over the internet, dialectology is ready to take the next step forward and to provide the public with access to original sound recordings of real dialectal speech made in the actual village context (along with, in most cases, analysis of the relevant dialectal material). As of this writing, there are five links on the internet which provide access to actual field recordings of dialect material in the Bulgarian/Balkan

Slavic region. Each gives access to digitized audio files of particular local dialects. Other than that, the five differ considerably in goals and presentation.

Two are the work of official institutes within Bulgaria and Macedonia, respectively: the Institute for Bulgarian language of the Bulgarian Academy of Sciences (IBE n.d.) and the Research Center for Areal Linguistics within the Macedonian Academy of Sciences and Arts (ICAL n.d.). In each case, a map is provided and tabs are placed at locations on the map for which information is available. A relatively small number of the tabs provide both audio recordings and prose material; most provide only prose material. Each of the maps depicts an area broader than that of the state in question. As noted earlier, the Bulgarian map depicts the entire Balkan Slavic region, while the Macedonian map extends into southwestern Bulgaria (the Pirin region) and into northern Greece at least as far as Thessaloniki (the "Aegean" region). The audio selections are not transcribed and no analysis is offered of them. The additional prose material provided for each tab on the Bulgarian site is a short list identifying the reflexes of the most important Old Bulgarian vowels and giving a brief sentence exemplifying the dialect. By contrast, the tabs on the Macedonian site link one to the page in the dialectal chrestomathy (Vidoeski 2000b) which contains transcribed texts from that village. The intent in each instance appears to be to key already published material to a map that allows direct visualization of geographical distribution and to raise the image of dialectology among the local populace. The Bulgarian site, informally called the "talking map," appears to be very popular among the non-academic public in Bulgaria, judging by journalistic reviews of it posted on the Institute's website.

The other three websites all provide transcripts of the audio files posted on them and varying degrees of analysis. Two are devoted to specific dialects outside Bulgaria proper and the third to the overall span of Bulgarian dialects within the boundaries of Bulgaria. The first (Mladenova and Mladenova 2001–2013) is focused on the "Transdanubian region" and presents recordings of Bulgarian dialects spoken in Romania. As is evident from the full title of the site, "Transdanubian Electronic Corpus, Supplement to *Bulgarian Dialects in Romania* by Maxim Mladenov," the material on the site is best studied in conjunction with the classic monograph mentioned in the title (Mladenov 1993). The site is handsomely arranged, with thorough metadata about each site visited, valuable background research articles, and a gallery of photos. The focus of the site, of course, is on the audio files and the transcripts accompanying them. The audio files are digitized versions of field recordings made over the period 1962–1975 by Mladenov and others and transcribed largely by Mladenov throughout the 1970s and 1980s. Each page on the site contains the textual material recorded in a particular village; both the texts and the corresponding audio segments are broken up into thematic sections. The site is very well indexed by thematic content; searching for any one of a number of thematic types brings up segments of conversation concerning that theme.

The site's basic name is appropriate: it is a corpus. That is, the focus is on natural speech recorded in context and on presenting the largest amount of natural speech possible. As in most corpora, the texts can be searched for individual words. Search results present the word in question embedded in a line of context. Hovering the mouse over the word brings up a box with the name of the village and the segment of text where the word appears. No further linguistic analysis is given, however. Furthermore, texts are available in Bulgarian Cyrillic only; no translations or transliterations are given.

The second set of materials is part of the Pangloss Collection (LACITO n.d.), a massive French-based site whose subtitle labels it "an archive for endangered languages." It includes material from all over the world, catalogued by the continent and country where the relevant endangered language is spoken. Under the heading of Greece, one finds two sets of recordings of Slavic speech, one labeled "bulgaro-macédonien" and the other "nashta." The recordings of "bulgaro-macédonien" were made in Edessa in 1976 by Georges Drettas; they comprise eight selections for a total of approximately 40 minutes. The recordings of the speech called "nashta" were made between 2002 and 2004 by Evangelina Adamou in a village ten kilometers north of Thessaloniki which she identifies by its Greek name Liti, but whose Slavic name has been identified by Lindstedt (2011:339) as Ajvatovo. The dialect is clearly Balkan Slavic; however, Adamou prefers to call it as the locals do, "nashta," both in the entries on this website and in her monograph about the dialect (Adamou 2006). The website contains nine selections for a total of approximately 50 minutes.

As is all material on this impressive website, every text is accompanied by the metadata about its recording plus a small map locating it geographically. All texts are broken up into lines with a link at the beginning of each line to the appropriate spot in the audio file, and all are transcribed and translated. Some are translated only into French; others include more information. With respect to the Balkan Slavic texts, those recorded by Drettas (the ones in "bulgaro-macédonien") are translated into both French and English. Those recorded by Adamou (the ones in "nashta") are presented in more detail: they are translated into French, English, and Greek and are provided with interlinear glosses.

I move finally to the third site, entitled "Bulgarian Dialectology as Living Tradition" (Alexander and Zhobov 2011–2016), which is the outgrowth of a research endeavor I began over a quarter century ago. When it first became possible for Americans to travel freely in the Bulgarian countryside, I organized a series of joint American-Bulgarian field trips, comprising two longer expeditions in 1993 and 1996, and four shorter ones in 1990, 1992, 2002, and 2013.[2] These expeditions were organized with several research goals in mind, all of which required the collection of large amounts of free-flow, natural conversational dialectal speech over a broad area. Two of these goals arose in direct response to the fact that published dialect atlas material was limited to the word level and in general did not provide sufficient data to study phrasal phenomena at the dialect level. The first was my own concrete research goal, which centered upon the relationship between clitic sequences and accentuation in various regions of Bulgaria. The second was more broadly phrased and was a goal of the entire research team. This was to provide material for the study of questions that are rarely, if ever, addressed in dialectal research, questions such as word order, functional sentence perspective, narrative structure, intonation, and conversational analysis. The third goal was more intangible and underlay the entire enterprise: it can best be phrased as a desire to return the focus of dialectology to its original source, actual living speech. This philosophical goal is expressed by the phrase "living tradition" in the project title. The decision to limit the scope to dialects within the boundaries of Bulgaria was taken for purely practical reasons, since to have also organized fieldwork at a similar level of coverage in other regions of Balkan Slavic simply was not humanly possible.

The total corpus of material collected on these field trips within Bulgaria amounts to some two hundred hours, a fact which presented us with a dilemma. Should we present the material as a "corpus" by simply uploading the raw tapes? Should we focus all our attention on transcription of as many tapes as possible and then make this available as a resource? Or should we take a different approach and present a selection of the material, which we could then annotate and analyze in more detail?

[2] Core members of the team were Georgi Kolev, Vladimir Zhobov, and Ronelle Alexander. This trio undertook the central 1993 and 1996 expeditions (both with support from the International Research and Exchanges Board [IREX]); the latter expedition also included three Bulgarian students (Tanya Delcheva, Kamen Petrov, and Petŭr Shishkov) and three North American graduate students (Matthew Baerman, Jonathan Barnes, and Elisabeth Elliott). The shorter expeditions were undertaken by Kolev, Alexander, and Radko Shopov (1990); Alexander and Maksim Mladenov (1992); Kolev, Alexander and two students, Petŭr Shishkov and Traci Lindsey (2002); and Kolev, Zhobov, and one student, Cammeron Girvin (2013). The site also includes material recorded by Zhobov and/or Kolev between 1986 and 1989, and excerpts from field recordings made by a Russian colleague of Zhobov's, Elena Uzeneva (2000), and by two students of Kolev's, Marieta Nikolova (2011), and Krasimir Mirchev (2012).

The fact that we chose the third path[3] means that the resulting site cannot be termed a “corpus” in the usual sense. Instead, the language material is presented as samples in the form of carefully chosen excerpts which represent the speech of each of the 68 villages we visited. The criteria guiding us in selecting the excerpts were that each should display as many of the characteristic features of the dialect as possible, and each should constitute a well-formed piece of discourse. These excerpts were then processed in detail. Each was transcribed, both in a modified IPA (i.e., Latin-based) transcription and in the Cyrillic transcription commonly used by Bulgarian dialectologists, and then translated into conversational English (with the goal of rendering the tone of the original conversation). Every word in the text was then glossed in detail, provided with the standard Bulgarian lemma to which it corresponded, and supplied with a number of “linguistic traits” that identified characteristics likely to be of interest to linguists. In terms of running time, the aggregate of the 181 different excerpts is somewhere near thirteen hours, and the full text (excluding lines spoken by investigators) totals some 300,000 tokens. For a corpus, this is very small—practically negligible. At the same time, the fact that all the language material in the site is from natural, spontaneous speech means that the corpus principle is being honored.

The unique feature of this site, and the innovation it brings to the digital processing of dialectal material, is that it combines two very different approaches to dialectal speech. One of these is indeed the corpus approach, in which large amounts of natural speech are gathered together and the entire sum of material is provided with at least some search capabilities. The other is the atlas approach, in which a great many individual items are annotated for historically relevant factors in such a way that the geographical distribution of these items with respect to these factors can be displayed on maps. One could call the first of these “synchronic” in that it focuses on actual speech in its natural context. The availability of transcribed and annotated text, keyed to the original audio files, allows for synchronic study of many elements of the actual speech stream. As to the other, it is properly diachronic since once the material is tagged for the relevant “linguistic traits” one is able to extract from

[3] We began this stage of the project in 2009, intending to produce an electronic publication containing annotated transcripts, commentary on the local dialects, and the relevant streaming audio files. In 2011 Quinn Dombrowski joined the team and created a relational database in order to make the material more accessible to researchers. Design of the site was taken over in 2013 by Cammeron Girvin, who worked actively in this capacity through 2016. I am very grateful to them both and to the many UC Berkeley undergraduates who did data entry as participants in UC Berkeley’s Undergraduate Research Apprentice Program (URAP).

these dialect texts all the elements that are relevant for the historical study of any one feature and display them on a map. Both goals are further served by the fact that all material on the site is translated into English, which exponentially increases the availability of this material to researchers who may not know Bulgarian well (or at all).

The basic unit of material on the site is the individual excerpt. Each of the 181 excerpts is referred to as a Text, and each occupies its own page. One can link to the texts either via the Home page or via the Contents page. The home page includes a map with tabs representing each village visited followed by a list naming each of these villages. Clicking on either the tab or the name will take one to the Location page for that village, which gives metadata about the location in question plus a more detailed map and links to the text (or each of the texts) recorded in that location. The Contents page presents a table containing information about each of the 181 texts, and clicking the name of the text in the far left column takes one directly to the Text page. Other columns specify the dialect group to which the text belongs, the number of tokens in the text (informant speech only), the length of the text both in minutes and lines, a brief synopsis of the text contents, and the status of data entry.[4]

When one links to the page of any one text, one sees metadata in a sidebar to the left and the audio link at the top right corner of the page; this link follows the text as one scrolls down the page. The longest text is slightly over 16 minutes and the shortest is 29 seconds; most are in the range of three to five minutes. Each text is broken up into lines of nine to twelve tokens; wherever possible, the line divisions correspond to rhythmic pauses or natural syntactic breaks. Lines are numbered for retrieval of information, and the speaker of each line is identified by a code letter. Furthermore, each line spoken by an informant is supplied with a time-code, which allows one to move easily to the relevant spot in the audio file. Everything on the tape is transcribed, including non-verbal cues such as laughter and speech by outsiders that could conceivably have affected the flow of conversation.

Each text is available in three different views; all views contain the line number, the speaker identification code, and the time code. The most detailed of the three views is called “Glossed text.” It contains an English translation of

[4] Because these columns will be removed once data entry is complete, it is likely that they will no longer exist when this report is in print. For the period between the release of the site (30 April 2016) and the completion of data entry, their obvious function will have been to inform the user of availability of specific data. The research team vacillated about releasing the site before all data entry was completed, finally deciding both that the value to researchers was sufficiently great and that by April 2016 enough texts had been entered to give a good coverage of Bulgarian dialects.

the text and then the transcribed text itself separated into “tokens” (individual words or pieces of utterance). Each of these tokens is annotated for grammatical and pragmatic information and, in the case of non-function words, provided with an English gloss. Additional tags identify the standard Bulgarian lemma to which the form corresponds (or provide a “dialectal lexeme” if no such standard lemma exists), and mark any linguistic traits of potential interest for further analysis.

The second view, intended to allow distraction-free reading of the text for content, is called “Line display”: it contains only the transcribed text itself (with normal spacing between words) and the English translation. The third view, intended for Bulgarian dialectologists (and others who prefer to read the text in Cyrillic transcription) is called “Cyrillic line display”: it contains only the text itself, transcribed according to the system used in Bulgarian scholarly publications.[5] Below I quote examples of a single line (from the text “Babjak 3”) in the three different views. Boldface in the examples indicates that the material in question appears in a different color on the page.

Glossed text

2 (b) [0:02]   We used to celebrate St. George’s Day here in the village.

gerg′òvden	**tuk**	**f**	**selòtu**	**gu**	**praznùvahme**
St.George’s.Day sg m	here adv	in	village sg n def	acc m 3sg clt	celebrate 1pl impf I
Гергьовден	**тук**	**в**	**село**	**той**	**празнувам**

Line Display

2 (b) [0:02]   gerg′òvden tuk f selòtu gu praznùvahme
We used to celebrate St. George’s day here in the village.

Cyrillic Line Display

2 (b) [0:02]   герг′о̀вден тук ф село̀ту гу празну̀ваhме

[5] The Latin-based transcription system is a simplified version of the International Phonetic Alphabet, using the more familiar Slavic symbols *č, š, ž* for post-alveolar sounds and marking stress on the vowel itself. Because the Bulgarian and the international symbols often differ considerably, the two transcription systems are listed on the site (in the “Principles of Data Presentation” section under the Site Information tab).

The abbreviations in the glosses are usually comprehensible (a full list of them is available in the “Wordform Tags” section under “How to Use This Site” under the Site Information tab).

Because the “glossed text” view is complex enough already, the tags called “linguistic traits” are not given on that view. However, because each Token has its own page, one can link directly to that page and see which linguistic traits have been assigned. In the case of this line, only the first token bears such a tag in the form “fjer root s0 /e/”. That tag itself links to the page for the particular linguistic trait, which spells out the information that the tag refers to a Historical Slavic Vowel (clearly a front jer), that a Morpheme Type has been specified (the front jer is in a root morpheme), that the absence of Stress is indicated (had the vowel been stressed, the tag would have contained the notation “s1”), and that the phonetic Realization /e/ has been specified. All of this information refers to the form *den* within the token *gerg'ovden* ‘St. George’s Day’.

The purpose of all these different tags, of course, is to enable searches of the database. There are three different kinds of searches. The first is called Wordform Search, which can search for any combination of tags added to tokens (except for linguistic traits). The second is called Lexeme Search, which can search for information keyed to individual lexemes. The third is called Linguistic Trait Search, which (as its name indicates) can search for combinations of linguistic traits.

The Wordform search page allows the user to select any combination of lexeme, English gloss, and wordform tags. For instance, the token *selòtu* from the line quoted above would be one of a number of results if the user had selected the lexeme “село,” and/or the English gloss “village,” and/or any of the grammatical tags “sg” (singular), “n” (neuter), or “def” (definite). The search results are then sorted according to texts, listed alphabetically. The token is listed first followed by all its tags (other than linguistic trait tags), the text and line in which the token appears, the full line context, and the full line translation. Below is an example of the form of the search result:

BABJAK 3
Token: **selòtu** village n sg def село
Babjak 3 Line 2
gerg′òvden tuk f selòtu gu praznùvahme
We used to celebrate St. George’s day here in the village.

The searched-for item is the token, and the output allows the user to see at a glance all the relevant information about the token: not only the line in which it occurs but also all the tags associated with it and the actual context of the line (with the English line included for those whose Bulgarian is rudimentary or non-existent). The boldfaced items—the token name **selòtu** and the text name **Babjak 3**—indicate that one can link from them directly to either the page of the token or the page of the text.

The Lexeme search page has two functions, with a third yet to be developed at the time of this writing. Its most basic function is to display all the phonetic renditions across the site of any one standard lexical item. Secondarily, it can search for all of the lexemes which are tagged for traits such as "dialectal lexeme," "dialectal usage," "folk etymology," or loans from various source languages. Finally, it will be able to search for semantic groups—all the dialectal expressions across the site of any one particular concept; it is expected that this capability will be part of the site by the time this report appears in print.

The Linguistic trait search page is set up to do searches for two kinds of traits. These searches are accomplished by means of a hierarchy: one selects first from among a list of domains comprising Phonology, Morphology, Syntax, Lexemes, Pragmatics and then follows through the subsequent hierarchical lists to find the trait of interest (a full list, with all hierarchies expanded, is available in the "Linguistic Trait Tags" section under "How to Use This Site" under the Site Information tab). The first kind of trait is simplex: one simply finds the right trait on the list and then initiates the search. The second kind of trait is complex: this is the case with all the traits relating to "historical Slavic vowels or consonants". As in the example of the front jer in the token *den* above, one can specify a number of different Conditions (such as stress, morpheme type, presence/absence of palatalizing environment, presence/absence of postalveolar or /j/, presence/absence of labial consonant, or presence/absence of /c/), as well a Realization (the present phonetic shape of the historical vowel or consonant). Not all conditions apply to all segments (for instance, the condition "presence/absence of /c/" applies only to the historical Slavic vowel jat′); those that are relevant are noted when one selects the historical segment in question. Normally one adds all the relevant specifications; however, this is not necessary. One could, for instance, search for all instances of front jer in stressed syllables or all instances of front jer in stressed root syllables without specifying a phonetic realization.

The output of the search is similar to that of the Wordform search: each citation gives the token in question, the text and line number, and the full line in which the token appears. As in the Wordform search results, one can go directly to the page of any token or text. But this search has something additional, in that it plots results on a map. This is the aspect of the site that replicates a dialect atlas; it is particularly appropriate that the site is able to map the behavior of the major Old Bulgarian/Late Common Slavic vowels and consonant sequences throughout Bulgarian dialects—that is, to give responses to some of the major questions posed by the All Slavic Linguistic Atlas.

By the time this report is in print, it is expected that data entry of all the above tags will have been completed and that work will have advanced to at least one, and possibly two or three, of the next four envisioned tasks. The first of these is to index the texts for thematic content. This will be done by creating a list of thematic tags and identifying lines that speak to that content; a search will then list all the relevant chunks of lines, identified by text. The second of these is to create the semantic groupings and append that search capacity to the Lexeme search page. The third is to append prose summaries to the site, summarizing the major traits of each village's dialect, and providing commentary to the particularities of each text.

The fourth is the most ambitious. This is to create a new data set composed of Phrases. In the current setup, only individual tokens or individual lines can bear tags. The research team is now sketching out a new framework which will allow the tagging of phrases, a category comprising a number of different word combinations that deserve comparative analysis. Among these are "verb + *se*," compound verbs (auxiliary + L-participle, future in the past, etc.), phrases of clitic(s) plus headword, doubled object pronouns, approximative numerals, compound pronouns, and the like. When this new section is complete, one will be able to search not only for the phrases themselves but also for combinations of relevant traits of each phrase, such as word order, effect of negation, accentual properties, and the like.

Even in its current state, the site has proved its worth to researchers, and several research projects have been initiated using data from the site. Two have already been completed: one concerning dialectal accentuation (conducted by myself) and one concerning dialectal vocalism (conducted by the site's Bulgarian co-director, Vladimir Zhobov), both of which are published in a single monograph (Zhobov and Alexander, to appear). Each of these two research projects is quite innovative, and in each instance the innovative nature was possible because of the way the site is structured. In the instance of dialectal accentuation, I was able not only to identify the presence of three different types of dialectal accentual patterns in phrases containing clitics, but was also able to make statements about frequency of occurrence. Because I had at my disposal large amounts of natural speech, I was able first to establish the frame in which each accentual pattern was possible and then to plot the number of instances when it actually did occur against the number of instances in which it could have occurred. Such a result is significantly more meaningful than a simple statement that the accentual pattern in question either occurs or it does not. In the instance of dialectal vocalism, Zhobov was able to carry out two significant experiments, both possible because he had large amounts of digitized audio data to work with. With respect to "vowel space" in any one dialect, he was able to isolate individual segments of dialectal vowels from the actual tapes and do listening tests with native speakers in order to test "auditory distances." With respect to vowel reduction, he was able to extract comparable sequences from different dialects and measure the relationship between the length of stressed vowels and the degree of vowel reduction.

In the preceding, I have discussed the great advantages which the digital age has brought to dialectology and described several examples, paying primary attention to the site which my research team has developed. Although I believe it is an important research tool in itself, well worth all the work, there is one additional, highly practical feature about this site, and this is the fact that it is not dependent on major grant funding. Instead of relying on custom programming, the site architect was able to construct it using the open source content management system Drupal. Under this method, one constructs a site from "modules" freely available on the internet, and then customizes it according to one's needs. Because the work of the site architect (in the original construction) and the project design coordinator (in further customization) accomplished the technical aspects of the site with relatively little financial outlay, this meant that work on the site is basically the same as that of any research undertaking, many hours of careful attention to textual material and processing of data. Another desirable outcome of this method is that the structure of the site is exportable: anyone wishing to undertake a similar project may freely adopt the model we have created and adapt it to their own data. It is my hope that those working in related languages and dialects will adopt this model and create similar research tools.

University of California, Berkeley
ralex@berkeley.edu

REFERENCES

Adamou, Evangelina. 2006. *Le nashta: Description d'un parler slave de Grèce en voie de disparition*. Munich.

Alexander, Ronelle, and Vladimir Zhobov. 2011–2016. *Bulgarian Dialectology as Living Tradition*. http://bulgariandialectology.org

Bozhkov, Rangel. 1986. *Bŭlgarski dialekten atlas: severozapadni bŭlgarski govori v caribrodsko i bosilegradsko*. Sofia.

Gajdova, Ubavka. 2007. “Makedonski dijalekten atlas (MDA)”. *Studia z filologii polskiej i slowiańskiej* 42:131–142.

IBE (Institute for Bulgarian Language). 1964. *Bŭlgarski dialeketen atlas I: Jugoiztochna Bŭlgariia*. Sofia.

—. 1966. *Bŭlgarski dialeketen atlas* 2*: Severoiztochna Bŭlgariia.* Sofia.

—. 1975. *Bŭlgarski dialeketen atlas* 3*: Jugozapadna Bŭlgariia.* Sofia.

—. 1981. *Bŭlgarski dialeketen atlas* 4*: Severozapadna Bŭlgariia.* Sofia.

—. 1988. *Bŭlgarski dialeketen atlas: Obobshtavasht tom*. Sofia.

—. 2001. *Bŭlgarski dialeketen atlas: Obobshtavasht tom* 1–3*: Fonologiia, leksika, akcentuaciia*. Sofia.

—. 2016. *Bŭlgarski dialeketen atlas: Obobshtavasht tom* 4*: Morfologiia*. Sofia.

—. n.d. *Karta na dialektnata delitba na bŭlgarskiia ezik*. http://ibl.bas.bg//bulgarian_dialects/

Ivanov, Iordan. 1972. *Bŭlgarski dialekten atlas: Bŭlgarski govori ot egeiska Makedonia* 1*: Dramsko, siarsko, valovishko.* Sofia.

LACITO (Langues et civilisations à tradition orale). n.d. *Collection Pangloss*. http://lacito.vjf.cnrs.fr/pangloss/index.html

Lindstedt, Jouko. 2011. “Review of Adamou 2006”. *Journal of Slavic Linguistics* 19 (2):339–45.

ICAL (Research center for areal linguistics). n.d. *Digitalna zbirka na tekstovi od makedonskite diialekti*. http://ical.manu.edu.mk/index.php/dialect-collections

Mladenov, Maksim. 1993. *Bŭlgarskite govori v Rumŭniia*. Sofia.

Mladenova, Olga, and Darina Mladenova. 2001–2013. *Transdanubian Electronic Corpus, Supplement to Bulgarian Dialects in Romania by Maxim Mladenov*. http://corpusbdr.info

Vidoeski, Bozhidar. 1998. *Dijalektite na makedonskiot jazik, tom* 1. Skopje.

—. 1999a. *Dijalektite na makedonskiot jazik, tom* 2. Skopje.

—. 1999b. *Dijalektite na makedonskiot jazik, tom* 3. Skopje.

—. 2000a. *Fonološki bazi na govorite na makedonskiot jazik*. Skopje.

—. 2000b. *Tekstovi od dijalektite na makedonskiot jazik*. Skopje.

Zhobov, Vladimir, and Ronelle Alexander. To appear. *Bulgarian Dialects: Living Speech in the Digital Age.*