TRACKING NEW ELEMENTS IN BULGARIAN DIALECTS

Ronelle Alexander

ralex@berkeley.edu

University of California, Berkeley

The interactive website *Bulgarian dialectology as living tradition* (Alexander and Zhobov 2011–16) presents material of Bulgarian dialects from 69 different Bulgarian villages, recorded by a joint US-Bulgarian field team over roughly the last quarter century. The primary goal of the project was to record, in both written and audio form, stretches of natural conversation (some quite lengthy), each chosen from the full field corpus because it displays both the maximal amount of relevant dialectal traits and for its great ethnographic interest. The website's other goal was to make material more accessible internationally by providing Latin and Cyrillic transcription and English translations. Additionally, the site authors have annotated the texts and provided search options for ease of access and interaction. The interactive aspect involves five search options contained on the website (four of which are currently available to users with the fifth still in preparation). Three are relatively straightforward. These three, which contain only minor innovations, are the searches for Wordform, Lexeme, and Thematic Content. The other two, the searches for Linguistic Traits and Phrases, are quite innovative, a result which derives directly from the fact that they were devised purely on the basis of linguistic elements found within the material itself. This report describes all five searches, with particular concentration on the two more innovative ones.

Keywords: Bulgarian, dialectology, ethnography, corpus linguistics, text annotation

1. Introduction. This paper presents an overview of the *Bulgarian dialectology as living tradition* interactive website (Alexander and Zhobov 2011–16). This project presents written and audio transcriptions of stretches of dialogue from 69 Bulgarian villages, making it an invaluable tool for both linguists and ethnographers. Moreover, the site's authors have designed a level of interactivity that allows for fined-tuned corpus analysis, involving five search options, which I describe below.

2. The five search options. The first search, the **Wordform search**, allows one to select a combination of grammatical or pragmatic tags and to see all tokens within the site that have been assigned these tags; one can also search on this page for either the English gloss of any one token or the standard Bulgarian lemma associated with any one token. One can also, of course, combine any of these searches. Most of the grammatical tags are those normally found in such searches. These include (with respect to nominal and pronominal forms) tags for case, number, gender, and definiteness; and (with respect to verbal forms) tags for person-number, aspect, and descriptor tags for both finite and non-finite verb forms. Additional tags allow one to specify various

James J. Pennington, Victor A. Friedman, and Lenore A. Grenoble, eds. *And Thus You Are Everywhere Honored: Studies Dedicated to Brian D. Joseph*. Bloomington, IN: Slavica Publishers, 2019, 21–28.

syntactic and pragmatic factors. The list also includes traits specific to Bulgarian, such as the category of the "existential" forms *ima* 'there is' and *njama* 'there is no' and the animate masculine counting forms such as *dvama* 'two [humans]', as well as traits specific to Bulgarian dialects, such as the proximal, medial, and distal distinction made in definite forms in certain regions (exemplified by the set *seloso* 'the village [this one here]' vs. *seloto* 'the village' vs. *selono* 'the village [that one there]').[1]

The **Lexeme search** allows one to see all the phonetic implementations throughout the site of any one standard lemma form. To enable this, all tokens (whose main form is in direct phonetic transcription) were tagged with the lemma from the standard language to which they correspond. For dialect forms that do not have a direct correspondent in the standard language, a DIALECTAL LEXEME was created. As opposed to tokens, whose base form is in the Latin transcription, lexemes are transcribed in the standard spelling of literary Bulgarian (using the Cyrillic alphabet, of course). The reference source for all judgments (both the form of the standard lemma, and the decision as to whether or not a particular dialect word can be associated with a lemma in standard Bulgarian) is Andrejčin et al. 2012.[2] One can search for a lexeme in its entirety, but one can also search for lexemes that begin with, or end with certain segments. The latter two options allow one to search for instances of individual prefixes and suffixes, respectively.

The Lexeme search also allows one to isolate lexemes which have been identified as belonging to categories of special interest. One such category is dialectal lexeme, and another is the use of a standard lexeme in a markedly non-standard meaning. Most of the remaining such categories concern lexemes that have been identified as loanwords (from sources including Balkan Latin, German, Greek, Romanian, and Turkish). Finally the Lexeme search allows one to search for LEXICAL VARIATION—instances where the dialectal terms for particular objects or verbal actions show an interesting range of variation. The primary innovation in the Lexeme search is the user's ability to search either by the Bulgarian lemma or its English equivalent (or both, in the cases where it is desirable to resolve ambiguity).

The **Thematic Content search** allows one to see all the passages (each listed as a sequence of numbered lines) within the several texts which concern any one topic. Users can select topics from one of two lists. The first list is ordered hierarchically, according to primary categories such as agriculture, animal husbandry, clothing and textiles, family, foodways, marriage, and the like; each of these includes a number of subsidiary topics. The second list is ordered alphabetically. In the first instance, therefore, the user can browse to get an overview of which topics have been indexed,

[1] This set of tags has been described in more detail in Alexander 2015: 7–10.

[2] Lexeme forms are transcribed as in the main entry in Andrejčin et al. 2012; in the case of doublets, the form chosen was the one which this source considered to be primary. Any form included in this source, even if it was marked as 'archaic' (*arxaično*), 'outmoded' (*ostarjalo*), 'colloquial' (*razgovorno*), or 'rare' (*rjadko*), was included as a lexeme. Only forms that do not appear in this source at all are associated with a dialectal lexeme.

while in the second instance a user can check to see if a particular single topic has been included. In either case, one initiates the search upon locating a topic of interest. Most of the topics in the list are of the sort that might be included in any ethnographic thesaurus list; at the same time the list includes numerous topics which are specific to the particular cultural world being depicted (and the particular conversations which happened to occur and are presented on the website).

In sum, each of the above searches represents a normal sort of inquiry that might be carried out on any such database (grammatical forms, lexical forms, thematic content). At the same time, it is important to note, in the case of all three searches, that the list of elements to be sought in the search (the specific tags added to tokens or lines) was devised in each case on the basis of the material itself: none of the lists was imported from any other source.

The other two searches are more novel, and were designed essentially from scratch. That is, once the site authors realized that the above three searches, although providing access to the data at a basic level, still did not allow searches for some of the most interesting material, they decided to go through the material in detail and to note everything of interest to their own research programs, as well as everything they could conceive might be of interest to other potential users of the site. As expected, this resulted in an extremely long list. Because the items of interest ranged across the spectrum of linguistic analysis, it was decided to call the next search simply Linguistic Traits. These traits were also assigned to individual tokens. The reason they were not incorporated into the Wordform list is because those more basic tags were intended also as interlinears, to be displayed below each token in the GLOSSED TEXT view of each text.[3] Including every one of these additional traits as well would have made that view of the text impossibly unwieldy.

However, the new **Linguistic Traits search** could embrace all of this material. The long list of tags was arranged into categories that are normal for linguistic analysis (phonology, morphology, syntax, etc.) and then the traits in each of these categories were divided into subcategories. As expected for a dialectal region in which the majority of isoglosses concern phonology, the set of traits under PHONOLOGY was by far the longest, and each of its subdivisions (vowels, consonants, and stress) was also fairly long. Some of the traits described synchronic phenomena, such as the elision of various segments, devoicing of vowels, or specific transformations of vowels, consonants, or vowel sequences. Although few of these traits mark significant isoglosses, many are of interest to phoneticians and phonologists, and the ability to locate instances in individual dialects, or to track them over the full Bulgarian dialectal landscape, is useful to such scholars.

[3] The site also allows two other views of each text, neither of which is encumbered with interlinear tags. These two views—one in Latin based transcription with English translation, and the other in the Cyrillic transcription used by Bulgarian dialectologists—allow distraction-free reading of the text for content.

Among other traits of interest are vowel reduction and consonant palatalization, especially when they occur outside the regions where such phenomena are expected. The most interesting traits, however, are those that refer to diachrony and that are listed on the site under the rubrics HISTORICAL SLAVIC VOWELS and HISTORICAL SLAVIC CONSONANTS. These tags allow a user to find modern reflexes of reconstructed jers (front and back), nasals (front and back), jat, jery, syllabic sonorants, the reconstructed sequences **tj* (plus **ktj*, **kt* + front vowel), **stj* (plus **skj*, **sk* +front vowel), **dj* (plus **gtj*), morpheme initial **čr* + front jer, and a few others. What is particularly useful about the latter group of tags is that one may either search for all reflexes of any one reconstructed segment, or one may restrict the search by specifying a number of CONDITIONS. Each set of conditions is keyed to the segment in question, since not all conditions are relevant for all segments. For instance, the conditions for the two jers include only MORPHEME (one can specify whether the segment in question occurs in a root, affix, or ending, or within a preposition or the definite article) and STRESS (one can choose to specify whether the reflex occurs in a stressed or an unstressed syllable). The conditions for the back nasal are similar, but those for the front nasal also allow one to specify whether or not the reflex occurs after a postalveolar or /j/. Finally, one can choose a REALIZATION: this enables one to locate all instances in which a particular reconstructed vowel or consonantal sequence has yielded a particular modern phonetic realization.

Still under the major heading PHONOLOGY is the third subheading, labeled stress. Here one can search for unexpected instances of stress in certain morphological categories, instances of lexicalized accent advancement or retraction, or instances of the well-known double accent—a dialectal phenomenon in which longer words can be spoken with two primary accents instead of just the etymologically expected one. (See below for more on this and related accentual phenomena.)

The set of traits under MORPHOLOGY is organized according to morphological category, and includes the rubrics definite article, nouns, pronouns, adjectives, and verbs, with the largest amount of traits to be found (as expected) under the latter. In most instances, only unexpected or interesting traits are tagged. In the case of the masculine definite article, however, it was adjudged that all instances of it are interesting, and all are tagged. Here too it is possible to choose conditions (to specify whether or not the article morpheme bore stress, and whether or not it occurred in a palatalizing environment) and to specify a particular realization.

Other points of interest within nominal forms included the presence of a count form in the feminine (or lack of one in the masculine), the presence of nonstandard oblique personal pronoun forms, gender marking in plural adjectives, and similar. Under verbs, five of the six the subcategories referred to separate verbal forms (present, aorist and imperfect, *l*-participles, passive participles, and verbal nouns), and the sixth to STEM UNIFICATIONS. Under this latter rubric were marked instances where the stem derivation of a verbal forms was other than expected. Tags within the categories referring to verbal forms identified instances of interesting endings (such as *-men*, *-mo*, or *-ne* for the 1st plural aorist or imperfect); and other interesting features, such as reduplication of /l/ in the *l*-participle, the occurrence of an unexpected suffix in the passive participle, and similar.

Relatively few categories were listed under SYNTAX and LEXEMES. In the first instance, this is due to the fact that tags had to be framed in such a manner as to be associated only with single words. Under gender, for instance, are found tags for the generalized use of the *l*-participle or non-standard gender marking; under case are found tags for feminine case markings in masculine nouns, and the historical feminine accusative singular used in nominative function; under aspect is found a tag for the use of a perfective verb in the negative imperative; and under transitivity is found a tag for the transitive use of an intransitive verb.

As to lexemes, one might think that all possibilities would already have been covered under the search capabilities of the lexeme page, especially since that page allowed one to search for the presence of individual prefixes (by searching for lexemes "beginning with …") or suffixes (by searching for words "ending with …"). In fact, however, there are at least two rubrics which transcend this formal level and are of interest to researchers. One is the category diminutive, which encompasses a number of different suffixes, and the other is a rather catch-all category entitled nonstandard usage.

The creation of the Linguistic Traits search was both a satisfying achievement, and a frustrating failure, since it soon became obvious that a number of critically important dialectal traits could not be tagged by the system in its current state. This, of course, is because each such trait requires reference to more than one word whereas the system could only encode traits pertaining to single words. This new set of traits includes elements such as reflexive verbs (verb forms which are incomplete without the added reflexive particle), compound verbal tenses (perfect, Romance perfect, pluperfect, future, future in the past, future perfect in the past), and various sequences in which the presence of clitics is of central importance. This latter group was particularly important, in that it embraced possessive constructions such as *majka mi* 'my mother', phrases including the so-called ethical dative, clitic strings (phrases comprising verbs plus pronoun objects expressed by clitics), and particular accentual phenomena found in certain clitic strings.

Paradoxically enough, it was at this point that what at first seemed like a crippling limitation of the site construction model became an asset instead. The limitation was that the basic data entry program in the Drupal content management system allowed data to be tagged only in two ways: either at the level of an individual token or at the level of a line of text. Consequently, the tags in three of the four searches discussed above (Wordform, Lexeme, and Linguistic Traits) had been assigned to individual tokens, while the tags in the remaining search (Thematic Content) had been assigned to chunks of text, one line at a time. Clearly, however, the set of traits which required reference to a group of words could not be fit into this system: it would not be possible to tag them satisfactorily by putting the tag only on one of the words within the sequence, nor would it be useful simply to tag the line in which they occurred (since they rarely constituted a full line all on their own).

Fortunately, the Drupal system of interlocking modules allowed the addition of a new and separate module, which could then be configured to set up such a search. This led to the creation of the **Phrase search**, which will allow users to search the site for instances of such phrases. The fact that this search module needed to be created from scratch, and that the system did not allow the more easy adoption of existing search options (by which one specifies both a searched item and the item(s) that would precede and/or follow it), made it possible to customize this search to a much greater degree than would have been allowed in a system already set up to search for sequences of words. In particular, this meant that the list of phrases to be searched can (and does) include all (and only) those of interest to a scholar of Bulgarian and Balkan linguistics. To take only a few examples, it allows instances of the RENARRATED mood to be tagged whether or not an auxiliary is present or for instances of complementizer deletion to be noted, among other such innovative moves.

Furthermore, it became possible to modify searches in a number of different ways. For instance, if the phrase type selected includes a verb form, one can modify the search to include negation; and if the phrase in question includes clitics, one can further modify the search to specify the position of the negative particle within the string. One can also modify the search for any phrase including more than one clitic to specify the order of the clitics. Similarly, if the phrase type selected is DOUBLED NOUN/PRONOUN, one can modify the search to specify the order and type of the components.

The potential of this type of search can be illustrated by the way in which dialectal accentual patterns are tagged. It was noted above that certain instances of the well-known double accent of southwestern Bulgarian dialects could be tagged under the stress section within the Linguistic Traits search. But this is possible, of course, only if both of the accents occur on a single word form. However, as all who are familiar with the phenomenon of double accent are aware, it is frequently the case that double accent is found not just on lexical words but also on phonological words—sequences of words plus attendant clitics. Consider, for instance, the following examples, all drawn from material on the database.

(1) lèb-uvè-tu
bread-PL-the
'the [loaves of] bread'

(2) dètencè-tu sì mi uzdravè
child-the REFL to-me get.well-3SG.AOR
'my child recovered'

(3) vèemè gu
we-winnow it
'we winnow it'

(4) dadème sì gu
we-give REFL it
'we give it [back]'

The first three examples above contain a single word attested with two accents. These three words (*lèbuvèto*, *dètencètu*, *vèemè*) carry the tag DOUBLE ACCENT and will show up in a search for that phenomenon under Linguistic Traits. But this is a deceptive result, since only in 1 and 2 is the occurrence of double accent completely defined by that word, whereas in 3 it is the larger, four-syllable phonological word that conditions the double accent. Furthermore, when the single-word search retrieves the initial four-syllable word in 2, it misses the fact that the two following clitics in fact create a six-syllable phonological word, also characterized by double accent.[4]

The Linguistic Traits search does allow one to see these larger contexts, because the search results list not only each tagged token, but also the full line of speech in which the token was attested. But there is no way for the Linguistic Traits search to tag the double accent seen in 4, since only one of the accents on this five-syllable phonological word occurs on the main token (the verb form). Consequently, the Phrase search is the only way to mark the double accent in 4. Indeed, the Phrase search is by far the preferred way to mark all these data, since the obvious goal is to capture them all together.

There is also another accentual pattern found widely in this group of dialects, a pattern which also includes accented clitics. But whereas certain instances of double accent can in principle be tagged under Linguistic Traits, instances of this second accentual pattern can only be recovered by the Phrase search. This is because the accent of interest always occurs on a clitic form. In this pattern, discovered by the field team whose data form the basis of the website, and provisionally named by them ADDITIONAL ACCENT (see Zhobov et al. 2004), a poststressing conjunction (or particle) triggers accent on a following clitic (usually within a verb phrase). Here are examples.

(5) ako gò razbṛ̀kaš
if it mix.2SG.PRES
'if you mix it'

(6) ž′ə vì gu kàža
FUT to.you it show.1SG.PRES
'I'll show it to you'

(7) kat sè vəzgènaha
when REFL twist.3PL.AOR
'when [things] all went bad'

Because a stressed clitic could in principle be part of either sort of accentual phrase, there is no way to tag these examples unambiguously at the token level. Since it is obviously important to be able to distinguish clitics after poststressing conjunctions from those in other sorts of phrases, some solution needed to be found. The Phrase search resolves this ambiguity nicely, and will easily allow users to compile data lists of the two types separately.

[4] The term double accent also encompasses instances of more than two accents within a phonological word, if the word is long enough.

The construction of the Phrase search component is now complete and available, although data entry is still ongoing. Indeed, this search capability may well prove to be one of the most valuable aspects of the database, since it will allow users to track in detail, and with considerable sophistication, those traits of Bulgarian dialects which are particularly Balkan, and which have excited great interest among researchers in recent years. It is characteristic of the site overall, which has aimed in every instance to perform analysis from the bottom up, inspired by the material itself, rather than top down, as in the majority of searchable sites.

3. Conclusion. The *Bulgarian dialectology as living tradition* interactive website (Alexander and Zhobov 2011–16) offers exciting possibilities for linguists and ethnographers of Bulgarian language and culture to conduct in-depth research of stretches of transcribed and recorded audio from 69 Bulgarian villages. In addition, the site's authors have made strides in providing a truly interactive corpus to aid researchers in their searches. Moreover, the site's authors have specifically designed the available five search options to benefit linguistic and ethnographic investigations. Finally, although not all data have been fully entered into the website, it has the potential to aid tremendously in the advancement of Bulgarian studies on the international scene.

REFERENCES

Alexander, Ronelle. 2015. Bulgarian dialectology as living tradition. *Balkanistica* 28.1–13.

Alexander, Ronelle, and Vladimir Zhobov. 2011–16. *Bulgarian dialectology as living tradition* (text portions only). Online: http://bulgariandialectology.org/

Andrejčin, Ljubomir; Ljuben Georgiev; Stefan Ilčev; Nikola Kostov; Ivan Lekov; Stojko Stojkov; and Cvetan Todorov. 2012. *Bălgarski tălkoven rečnik*, 4th ed., expanded and revised by Dimitŭr Popov. Sofia: Nauka i izkustvo.

Zhobov, Vladimir, Ronelle Alexander, and Georgi Kolev. 2004. Hierarchies of stress assignment in Bulgarian dialects. *Revitalizing Bulgarian dialectology*, ed. by Ronelle Alexander and Vladimir Zhobov, 226–40. Online: https://escholarship.org/uc/item/9hc6x8hp#page-226.

THE SYNCHRONY OF THE SERBIAN INFINITIVE: A SYNTACTIC PERPECTIVE

Bojan Belić

bojan@uw.edu

University of Washington

Joseph (1983:131) noted that 'Serbo-Croatian presents special problems with regard to the study of the Balkan infinitive-loss', the first of which 'is that there is some difficulty in finding reliable information on colloquial usage due to the strong prescriptivist tradition in the codification of the grammar of the language.' Hoping to avoid this problem, Belić's (2005) account relied on data collected by means of a questionnaire administered to a representative sample of native speakers of Serbian. The present contribution expands on that by including data from transcripts of parliamentary debates in the National Assembly of the Republic of Serbia, as well as from the Natural Language Processing group's Serbian web corpus. While Joseph (1983:147) indicated that one way 'in which Serbo-Croatian contributes to the study and understanding of the Balkan infinitive-loss is through the fact that an infinitive-replacement process is still in progress (which provides an opportunity to see first hand the variety of factors, social as well as purely linguistic, that can interact in the manifestation of this process)', the present contribution focuses exclusively on the syntax of the process.

Keywords: Bosnian-Croatian-Serbian, infinitive loss, Balkan linguistics, syntax, corpus linguistics

> *Le deuxième trait marquant des langues balkaniques est le manque complet ou partiel de l'infinitif et son remplacement par des propositions subordonnées. ... En serbo-croate, on peut employer ces dernières constructions [est remplacé par des propositions subordonnées] dans bien des cas, mais l'emploi de l'infinitif est prépondérant.*
>
> Sandfeld (1930:173–74)

1. Introduction. In his 1983 monograph *The synchrony and diachrony of the Balkan infinitive*, Brian Joseph offers an insightful account of the use of the infinitive in various so-called Balkan languages, one of which, at the time, was referred to as Serbo-Croatian. Joseph (1983:143) describes Serbo-Croatian as a language that 'represents a microcosm of the general Balkan situation—internal to Serbo-Croatian one finds the range of complete infinitive-loss to fluctuation between infinitive and infinitival-replacements, and infinitive-retention, as well as some indication of sporadic renewals of infinitives functionally'. It is precisely 'the distribution of the infinitive in the modern Serbo-Croatian dialects' that Joseph (1983:145) sees as 'the real value

James J. Pennington, Victor A. Friedman, and Lenore A Grenoble, eds. *And Thus You Are Everywhere Honored: Studies Dedicated to Brian D. Joseph*. Bloomington, IN: Slavica Publishers, 2019, 29–40.