Clitics, Particles and Phrases in Bulgarian and Balkan Slavic Dialects

Ronelle Alexander
University of California, Berkeley

One of the many differences between Balkan Slavic and the rest of the Slavic language world is the greatly enhanced role played by clitics and particles in Balkan Slavic. Although each of these small units can be considered a “word” (in that it can be given a definition, and by the convention of writing it separately), as a class they are different from lexical words not only in that they are unaccented, but also because they normally express their full meaning only in conjunction with one or more other words. A classic example is the future particle, which never occurs unaccompanied by the verb form to which it imparts the idea of futurity. The reflexive particle similarly functions in accompaniment to some other form, as do verbal auxiliaries in the expression of the perfect tense. Clitic pronoun objects, while more self-contained in the sense of meaning, nevertheless are closely bound to some other word in the utterance, frequently forming with it a larger “phonological word.” In sum, all these Balkan Slavic clitics and particles can best be described not alone but as part of the phrase in which they occur, which one might define as a “grammatically significant sequence involving particles or clitics” (henceforth “phrasal unit”).

These facts present a considerable challenge to dialectologists: how does one study dialectal variation with respect to these phrasal units? It is no accident that the maps in dialect atlases almost always depict the behavior of individual words, whether it concerns the reflex of a particular Old Slavic vowel in a chosen sample word, the shape of a particular inflectional morpheme, frequently also in a chosen sample word, or individual vocabulary items, such as the local word for “potato.” This stands to reason, since large-scale questionnaire work is most easily accomplished, and elicits the most comparable data, when single-word responses are sought. But the inability of this method to depict dialectal variation of phrasal units poses an implicit challenge to dialectology.

In this brief contribution I will describe how the relational database *Bulgarian Dialectology as Living Tradition* (henceforth BDLT), met this challenge. The basic goal of this database (available on the internet at [http://bulgariandialectology.org]) is to give users direct access to unedited recorded segments of natural speech (called "texts" on the site) from a range of Bulgarian dialects, and to annotate these texts in ways that allow researchers to conduct data searches of various sorts. The texts (which are available in both Latin and Cyrillic transcription, in English translation, and as audio files) are transcribed according to the normal conventions of Bulgarian, which means that the definite article is written together with the lexical word to which it is attached, while all other clitics or particles are rendered as separate words. Texts are divided into lines for ease of data retrieval; each line contains from nine to twelve words (or "tokens") and line breaks correspond, wherever possible, to rhythmic or syntactic breaks.

The content management system used to construct the site allowed tagging of the texts at either the token or the line level. While the latter option allowed searches of the texts for thematic content (through the annotation of chunks of lines that are concerned with a particular conversational topic), the former meant that searches for linguistically relevant material could take place only at the word level. Under this limitation, which amounted in essence to the same as that faced by the compilers of dialect atlases, there was no way to annotate the many phrasal units which are so important to Bulgarian and Balkan Slavic.

I quote below two sets of examples from the BDLT database as illustrative of the frustration created by this limitation. (Numbered examples are quoted as found on the text page of the BDLT site; the text name and line number, given below the example allow one to locate the example on this site so as to both view it, and hear it, in its original recorded context. Examples quoted in the text are also from the site but are not provided with glosses or textual reference.)

The first example is focused on the participial form of the verb *naprav'a* 'make.'

(1)	napravɛ̀li	sa	…	tìkvenìk'
	make pl L.part P	3pl pres aux clt		pumpkin.pie sg m

'They made pumpkin pie' (Bansko 56-57).

(2)	ednì	napraìli	trl’àci
	one pl adj	make pl L.part P	pen pl m

‘Some have made sheep pens’ (Repljana 1:98).

To one who knows Bulgarian grammar, it is clear from each example in context that the participial form in (1) is part of the perfect tense, while the one in (2) is (presumably) the renarrated aorist. But there is no way to alert a user unfamiliar with Bulgarian to this fact, nor is there any way for a user to search the entire database for instances of perfect *vs*. renarrated, since the word-level limitation means one can only search for instances of the participial form. Clearly, one would like a way to mark the sequence *napravɛ̀li sa* in (1) as perfect and the sequence *[Ø+] nəpraìli* in (2) as renarrated.

A trickier situation arises in the case of clitic forms bearing accent. Consider the following three examples, each of which represents a very different type of phrasal accentuation. Yet once again, if information can be marked only on an individual token, all one can say is that each includes an instance of the accented clitic *si*.

(3)	s’àkuj	ni	sì	dàvə	rəkɤ̀tə
	each sg m	neg	dat refl clt	give 3sg pres I	hand sg f def

‘Not everyone gives their hand [to him]’ (Kolju Marinovo 6:39)

(4)	dədème	sì	gu
	give 1pl pres P	dat refl clt	acc m 3sg clt

‘We give it back [to them]’ (Babjak 3:21).

(5)	tùka	što	sì	sàdime
	here adv	what rel	dat refl clt	plant 1pl pres I

here [= in this spot] that we’re planting’ (Eremija 2:37).

In (3), the accented clitic follows the negative particle. Such accentuation is expected: not only does it appear in the vast majority of dialectal examples, but it is also prescribed in the standard language (in those few instances where it does not occur in dialectal texts, accent always falls on the negative particle, a fact tagged elsewhere on the site). The other two examples, however, are quite different. Example (4), although it contains three separate tokens, is a single phonological word exhibiting the well-known “double accent”; while example (5) exhibits an accentual pattern documented only by the BDLT team and named “additional accent”: this is a pattern in which certain conjunctions or particles appear to cause stress on a following clitic. For one who studies phrasal accentuation, it would be highly desirable to be able to isolate such phrases in question and mark them according to the type of phrasal accentuation exhibited. But with the limitation of token-level annotation, a researcher could only search for all instances of accented clitics, and then disentangle instances of the three types on a case-to-case basis.

It was dilemmas such as these that prompted the design of the Phrase Search component on BDLT, whose stated goal on the site is “to allow the user to locate grammatically significant groups of tokens, the meaning of which is impossible to tag at the level of the individual token.” (For a full description of the capabilities of this component, see [http://bulgariandialectology.org/how-use-site#phrasesearch].) Once such a component was set up, it was then possible to include in it a large number of phrases. Indeed, although the original intent was simply to identify and collate different types of phrasal units including particles or clitics, it soon became clear that other goals could be accomplished as well. One was the ability to modify the search for any one type of phrasal unit in a number of ways, another was to allow the inclusion of “grammatically significant sequences” that did not include clitics or particles, and a third was to include discourse commentary that might be relevant to the analysis of the searched data.

The first goal was accomplished by establishing a number of different categories. The most basic is called “Tense-Mood”: under this rubric one can search for any of the compound tenses (future, perfect, pluperfect, future-in-the-past and the like). By adding tags from additional, separate categories, one can then modify such searches by specifying whether pronoun objects, the reflexive particle or negation is present. Simplex verb forms (such as aorist, present, imperfect, imperative and the like) are also included under Tense-Mood, but are coded as part of a phrase only if accompanied by clitics (pronoun objects or the reflexive particle). Renarrated forms of the verb posed a special case, in that such forms can express both the idea of renarration and the meaning of a specific tense. Because the search function stipulates that only one tag within any one category can be assigned to the chosen phrase, it was necessary to set up a separate category called “Evidential,” comprising the tags “renarrated” and “dubitative.”

The category devoted to “Clitic Objects” allows one to search for different combinations of clitic objects (including the reflexive particles *se* and *si*) as well as instances where such clitic objects are attached to non-verbal forms. Usually these latter express possession (as in *màjka mi* ‘my mother’) but they can also express a looser relation (as in *vkə̀šti si* ‘at home’). But for phrases including clitic objects in reduplicative function (as in *sìreneto go narèžeš* ‘you slice the cheese’), a separate category, called “Doubled Phrase,” was created. This is because it was judged useful to identify not only the components of such phrases (whether the reduplicated object was a nominal form or a full-form pronoun, or whether it was an instance of subject doubling, as in *onò devòjčeto* ‘the girl’) but also the ordering of elements in the phrase. Thus, tags in this category also specify whether the reduplicative pronoun object (or subject pronoun) precedes or follows the full form to which it refers.

Indeed, in linguistic systems where clitic forms proliferate, the issue of word order is of great interest, especially if there is dialectal variation on this point. To reflect this properly, three additional categories devoted to issues of word order were created. The first, entitled simply “Word Order” marks the ordering of verbal auxiliaries (or copula forms) and pronoun object(s) with respect to one another; instances of clitics in initial position are also noted here. An additional tag (located of necessity in the category “Miscellaneous”) alerts the user to an instance of non-standard word order.

The second, entitled “Negation,” marks the position of the negative particle with respect to the clitic string (if compound verb forms are negated, this simple fact of negation is noted within this category). The third, entitled “Syntactic Cohesion” marks instances where grammatically extraneous material breaks up elements of the phrase as noted. Most such instances are tagged with the basic cover term “nonsequential,” but there is a separate tag denoting the disruption of the sequence “auxiliary + L-participle.” For instance:

(6) tò se odgòre izbìstri
nom sg n acc refl clt from.above adv clear.up 3sg pres P
‘It gets clear on top’ (Belica 3:17).

(7) kət sme nìe ràždali
when conj 1pl pres aux clt nom 1pl give.birth pl L.part I
‘When we gave birth …’ (Dolno Draglište 1:21).

In the case of (6), the tagged phrase is entered simply as *se izbistri*. In addition to the tags "present" and "reflexive," however, there also appears the tag "nonsequential," which alerts the user to the fact that the phrase elements did not occur consecutively within the speech stream. Since the list of data retrieved for any search (including those with the tag "nonsequential") always gives the full line in which the phrase was uttered, the identity of the intervening word(s) can be easily discovered. Examples such as (7) are similar: the tagged phrase is entered as *kət sme rəždàli* (the conjunction is included because it could in principle have triggered an instance of the additional accentuation seen in (5)). It is given the more specific tag "aux-X-verb" as an aid to users whose particular focus of interest may be word order phenomena involving the perfect tense.

Next, the simple ability to catalogue phrases allowed the inclusion of all sorts of phrases, a few of which include clitics (such as the ostensive *èto go* 'there he is' and non-verbal predicate phrases such as *sràm me e* 'I'm ashamed'), but most of which do not. For example, users can locate instances of compound imperative forms (such as *nemòj pità* 'don't ask' or *idì otkinì* 'go and pluck'), approximate numerals (such as *dve tri* 'two or three'), what can be called multiple determination (such as *tìa stàrite* 'the old ones'—instead of the expected *tìe stàri*), complementizer deletion (such as *nèma vìdim* 'we won't see,' or *št'a pòčneme* 'we would begin'—instead of the expected *nèma da vìdim* or *št'a da počneme*), instances of second accusative (such as *vìkame go krosnò* 'we call it a beam'), coordinated idioms of the type *gorèšto ne gorèšto* 'whether it's hot or not,' and several other types of phrases. Some of the phrases catalogued are of historical and comparative interest, such as instances of the future tense with complementizer (such as *če da putue* 'he will travel') or the future tense with conjugated future particle, such as *ču ga poparim* 'I'll scald it').

The choice to include a category devoted to discourse elements, which is called simply "Style," was prompted by the fact that the site includes several instances of a clear shift in register, either to the narration of a folktale or the quotation of song lyrics. Although these two speech registers are not intrinsically connected with phrases, it was thought useful to include them, especially to allow users to compare the frequency of certain verbal forms (such as the renarrated) with these two marked speech styles.

In sum, the Phrase Search component of BDLT is perhaps one of the site's most valuable contributions to Bulgarian and Balkan Slavic dialectology. Not only can scholars of Bulgarian dialectology now find data on many features beyond the word level, but scholars of classically Balkan features such as the analytic future, object doubling and the renarrated mood can now have recourse to a great deal more data. The BDLT data are the more valuable since everything on the site is presented exactly as it was recorded in the field.