DOI 10.31168/2618-8597.2021.16.06

Digital analysis of Balkan phraseology

Ronelle Alexander
University of California (Berkeley, USA)
ralex@berkeley.edu

One of the most characteristic features of the Balkan Sprachbund is the way in which idiomatic phrases cross from one language to the next with such apparent ease, and the way in which they give such a Balkan "flavor" to each individual language. The collection and analysis of such instances is an ongoing, and important, contribution to our understanding of Balkan culture and the *homo balcanicus*.

The focus of such contrastive phraseology studies is primarily semantic. In contrast, my focus here is the phrase as a grammatical construct, defined roughly as a "grammatically significant group of tokens, the meaning of which it is not possible to tag at the level of the individual token". My examples are taken from Slavic (primarily Bulgarian), but because many represent well-known morphosyntactic "Balkanisms", the discussion frame is broader.

Not all such phrases are specifically Balkan. Reflexive verbs, for instance, which consist of a main verb form and a reflexive particle (such as *разбира се* 'it is understood') are found in non-Balkan South Slavic as well. Similarly, many compound verb forms are found in both Balkan and non-Balkan languages: these include perfect-like tense forms such as *дошли сме* 'we have come' or *бяхме дошли* 'we had come' and modal-like future forms such as *ще дойда* 'I will come' or *щях да дойда* 'I would have come / was about to come'.

Other compound verbs are more specifically Balkan. One is the Romance perfect, which is much better known in Macedonian (e.g. *имам дојдено* 'I have come') but is also found in Bulgarian dialects. The most well-known, however, is the renarrated mood, also called the evidential, which in the third person forms the phrase *Ø дошли* 'they [apparently] came'. The obvious common feature to all these forms is the combination of an auxiliary and a main verb form: in the Bulgarian examples, the auxiliary is a form of *съм* in the perfect-like tenses, a form of *ща* in the modal-like future tenses, and zero in the third person forms of the renarrated.

Grammatical phrases containing nominal forms tend to be specific to the Balkans. One notable example is the doubled object, in which an orthotonic object form (either a nominal full pronominal) is accompanied by a reduplicated clitic pronoun, as in *сиренето ще го нарежеш* 'you slice the cheese' or *нас ни пазят* 'they protect us'. Another is the dative of possession, in which a dative clitic pronoun imparts the meaning of possession to the preceding noun, as in *майка ми* 'my mother'. Indeed, one of the most striking Balkanisms, the marking of definiteness by the addition of the post-posed article, clearly should be labeled a "grammatically significant group of tokens" but is excluded for purely graphic reasons, since the standard orthographies of both Bulgarian and Macedonian require the article to be written together with the form which it renders definite. But just as the above two examples consist of an orthotonic nominal form accompanied by a clitic, so do the definite forms *селото* 'the village', in which the article is affixed to the noun, and *голямото село* 'the big village', in which the article is affixed to the adjective.

The above listing is a typology of sorts, which is striking in that nearly all items consist of an orthotonic word plus a clitic (with two exceptions: the pluperfect auxiliary is a fully stressed word, and the future auxiliary, although it is unstressed, can nevertheless occur in initial position). But a full catalog of "grammatically significant groups of tokens" would also include combinations of the above. Compound verbs frequently occur with pronoun objects (which themselves can be doubled) and/or the reflexive particle, doubled objects can occur in non-verbal predicates (such as *мене ме е срам* 'I'm ashamed'), and any of the above sequences can occur with the negative particle.

These facts about the grammar of Bulgarian form the backdrop for my discussion of dialectal variation. The primary reference tool of dialectology, the dialect atlas, contains maps which almost always depict the form taken by individual words, whether the intent is to illustrate the reflex of particular Old Slavic vowels, the shape of particular inflectional morphemes, or the range of lexical variation with respect to a particular lemma. Maps focus on words because they require readily comparable data from a very large number of individual sites; little to no information is available about phrases. It was to fill this gap that the collaborative project *Bulgarian Dialectology as Living Tradition* was undertaken. Fieldwork intentionally elicited long stretches of natural conversation in order to allow analysis at the phrase, sentence and discourse level.

The project website (available at http://bulgariandialectology.org) presents selections from the team's field recordings, annotated to provide five different search procedures. The Wordform search allows a user to choose any combination of the various grammatical or pragmatic tags that have been assigned to individual tokens, and to see each resulting token within the context of the line of text. The Lexeme search allows a user to see all phonetic implementations of any one lemma or to choose among various tags assigned to individual lexemes. The Linguistic Trait search allows a user to work through a detailed hierarchy to isolate very specific traits, again assigned to individual tokens. The Thematic content search allows a user to locate chunks of text devoted to a particular topic, each line of which has been tagged for that topic.

The fifth search is the Phrase search, and it is completely different. The tagged item is the phrase, the particular "grammatically significant group of tokens". The definition of any one phrase depends upon its components, and the search is set up to allow a user to search either for a single trait or for a combination of them. The components are grouped into categories, with titles such as "Tense-mood", "Evidential", "Reflexive", "Clitic objects", "Doubled phrase", "Negation", "Word order" and the like; these titles obviously correspond to the various examples given above.

The Phrase search greatly extends the options open to a user of the site. Consider the instance of reflexive verbs. Choosing the tag "reflexive" under Wordform search will elicit all instances of the reflexive particle. Choosing this tag alone under Phrase search will give a similar result, except that now the particle will be listed together with its headword. But it is only in Phrase search that one can modify the search to specify the desired tense or mood of the reflexive verb, and/or to specify (if desired) the presence of an additional pronoun object and/or the presence of the negative particle. One can similarly find instances of any compound verb form, negated or not, and with or without pronoun objects (simplex verb forms are included only if accompanied by pronoun objects; simplex verb forms standing alone can be found easily under Wordform search).

The Phrase search is particularly valuable in three specific instances. One of these concerns renarrated forms. Here, the Wordform search is of no help: since the form in question is tagged simply as the L-participle, the only way to tell whether it is the perfect tense or the renarrated mood is by combing through all the individual examples. The Phrase search disambiguates the two directly: one can search for "perfect" on the one hand and for "aorist renarrated" on the other.

The Phrase search also allows one to see differences in word order, not just as concerns the sequencing of the elements in doubled pronoun phrases, but also as concerns dialectal variation in word order. Whereas the norm in standard Bulgarian is for clitic objects to follow the negative particle and for object pronouns to follow the auxiliary in all but 3rd singular, there is dialectal variation on both these points. Thus, in addition to identifying each such phrase as to verb tense, number of clitic objects and presence or absence of negation, tags for the Phrase search also specify the position of the negative particle with respect to clitics, and of clitic objects with respect to the auxiliary.

Finally, the Phrase search allows the identification of phrasal accentual phenomena. The most well known of these, called "double accent", can occur within individual words (as in *гра̀довѐте* 'the cities' or *кра̀ставѝца* 'cucumber'), but it occurs just as frequently in phrases composed of orthotonic words and following clitics or particles (as in *гледа̀ли смѐ го* 'we've seen him' or *дѐтенцѐто сѝ ми оздравѐ* 'my dear child recovered'). In another type of phrasal accentuation (called "additional accent" on the site), clitics following certain phrase initial particles are accented (as in *ако гѝ разбр̀каш* 'if you mix it'). It is true that two other search options on the site (Wordform search, Linguistic trait search) can mark the fact of an accented clitic. But this is of no use, since such a list would lump together not only accented clitics from the two different types of phrases described above, but also the more common (and indeed ubiquitous) instances of accented clitics after the negative particle.

The website *Bulgarian Dialectology as Living Tradition* was created in order to present particular material of Bulgarian dialects, and all its search functions have been specifically designed for the most efficient and productive analysis of the actual data on the site. But because the site was constructed using open source content management tools, its structure is readily exportable to the data of other languages; and because Balkan phraseology — not only semiotic but also grammatical — is shared to such a great extent, the materials described here can be a tool not just for Bulgarian dialectology but also for Balkan linguistics overall.