

Bulgarian Dialectology as Living Tradition: An Innovative Dialect Corpus

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

Vladimir Zhobov (Владимир Жобов)
Sofia University “St. Kliment Ohridski”
zhobov@slav.uni-sofia.bg

This report presents *Bulgarian Dialectology as Living Tradition*, a new searchable database of oral speech representing a range of Bulgarian dialects, which is slated to become available to the public in late April 2016, with about 60% of the material entered and coded. Overall, the database comprises 179 excerpts from a large corpus of material recorded in 68 different Bulgarian villages over the period 1986–2013. The majority of the material was recorded between 1993 and 1996 by a joint Bulgarian-American field team, and the remaining material comprises recordings made by individual members of this team and their colleagues. Two major criteria governed the selection of the excerpts: each should display as many characteristics of the local dialect as possible, and each should represent a self-contained unit of discourse conveying information about traditional life.

The site’s specific goals are 1) to make both the discipline and the material of Bulgarian dialectology available to a broader, international audience; and 2) to focus on the spoken text as a living document of natural, real-time speech. Although the coverage is both less detailed than of a traditional dialect atlas, and more selective than that of a national corpus, the totality of the text in it nevertheless gives a sufficiently thorough sampling of the breadth of Bulgarian dialectal diversity. The real value of the site is in its ability to provide actual examples of the variability attested in natural speech (rather than categorical statements about dialectal isoglosses). Furthermore, thanks to the full annotation of each of these examples, the site provides ample material for concrete research results on actual attested dialectal variation.

One of the most important features of the site is that each analyzed text is accompanied by a digitized sound file, taken directly from field recordings, and that each line of each text is marked with a time code allowing users to locate the point in the sound file of any one attested word or phrase. Another valuable feature is that each text is transcribed not only in an internationally accessible IPA-based transcription, but also in Cyrillic, using the symbols which are conventionally accepted in Bulgarian dialectology. Finally, each word in each text is annotated to enable searches at three levels: grammatical categories, lexical headwords, and linguistic traits. The latter annotations give information about historical phonology in a manner which allows comparison with data in the All-Slavic Linguistic Atlas. The different excerpts will also eventually be indexed for thematic content.

The first part of the report describes the construction of the site and discusses briefly the rationale behind decisions of data presentation; and the second gives examples of the kinds of research which are now possible using dialectal data presented in this manner.