Teodora Vuković (University of Belgrade)

Towards the Torlak vernacular corpus

Each research starts with a sample. When it comes to South Slavic linguistics, its focus often lies on non-standard varieties and dialects which lack in digitized data to be analyzed. One such dialect is Torlak, a dialect spoken on the borderline area between Serbia and Bulgaria, officially considered an endangered language by the UNESCO.

A recent project, *Protecting the immaterial heritage of the Torlak vernacular*, took to collecting linguistic samples from the Serbian Timok region. This was perhaps the last chance to do so since it is currently almost exclusively spoken by older people in rural areas. A large amount of video and audio data was recorded in more than 50 villages across the Timok region, through ethnolinguistic interviews with over 100 informants, thus collecting more than 150 hours of audio and video material.

The final aim of the project is to create an online archive of this language that would include a corpus as the main resource for analysis. The corpus would focus on morphosyntax in order to enable researchers to explore the dialect and compare it to other South Slavic and Balkan languages. The current attention of the authors is at three grammatical features, namely the post-positive article, comparative and superlative forms, and the use of imperfect and aorist tense, all of which make Torlak stand out from other Serbian dialects. Furthermore, the corpus will be used to study geographic distribution of certain dialect features, as well as the influence of gender, age and education, their effect on use of language in individuals, and linguistic change and evolution.

I will present the steps towards creating this corpus, from transcription over normalization to annotation, drawing attention to potential problems and their solutions. I will also give examples of its use, predominantly in the field of linguistics, but also in other disciplines, such as anthropology, ethnology and history following the rich content of the interviews.