Degrees of non-standardness. Feature-based analysis of variation in a Torlak dialect corpus

Teodora Vuković, Anastasia Escher, Barbara Sonnenhauser
Slavisches Seminar, University of Zurich

Abstract

The South Slavic dialect continuum is characterized by an intricate encounter of affiliations: Genealogically, it is intersected by an old bundle of isoglosses differentiating Western and Eastern South Slavic. Areally, parts of it share a number of morphosyntactic innovations with their neighboring non-Slavic languages. The resulting variation becomes most distinct in the Torlak dialects spoken in Southern Serbia (gray area in Figure 1).

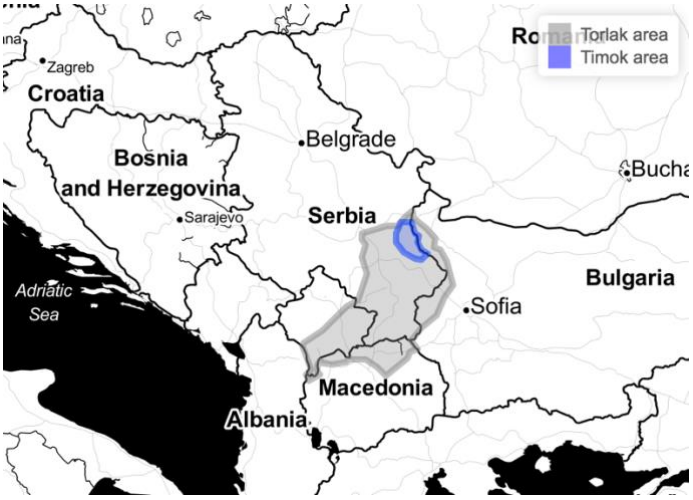

Figure 1: The Timok area within the Torlak dialect zone

A corpus-based method for assessing the range of dialect-standard variation is designed to identify samples with the highest prevalence of dialect features. It provides insight into areal and inter-speaker variation and allows the extraction of maximally non-standard manifestations of the dialect, which may then be sampled and used for the study of language change and variation. The focus is on the non-standard Timok variety of Torlak (blue area in Figure 1) undergoing considerable change under the influence of standard Serbian. The degree of variation is assessed by measuring the frequencies of five distinguishing linguistic features:

- accent position
- dative reflexive clitic *si*
- auxiliary omission in compound perfect

- post-positive article
- analytic case marking in indirect object and possessive

Taking the dialect and standard Serbian as two poles of a variation scale, it aims to empirically distinguish between gradual levels of non-standardness, and to establish accent position as a predictor for the use of other dialect features.

Speakers with the greatest and the least influence of the standard are revealed using hierarchical clustering. The analysis has revealed three clusters of speakers based on the presence of the dialectal features in their speech, see Figure 2.

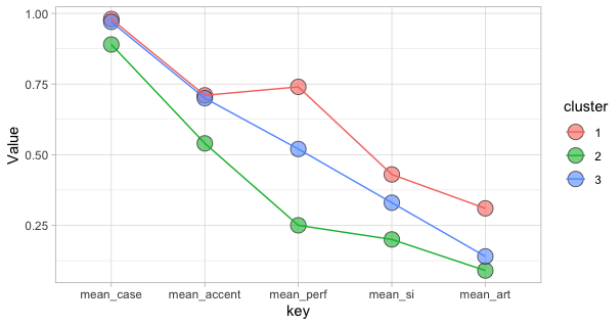

Figure 2: Cluster mean values for each feature.

A positive correlation between the occurrence frequencies reveals which non-standard feature is the best predictor of the other ones. Pearson's correlation method demonstrated that accent and the auxiliary omission can serve as an indication of non-standardness in the Timok sample. At the same time, the inability to identify a single reliable predictor of non-standardness illustrates the complexity of the linguistic variation present in Timok.

The results of the methods presented in this study will be used for classification of the corpus data, as well as to identify data on different poles to be used for contrastive analysis of the use of dialect. It can be expanded to include more features to further test the hypotheses. As such, the analysis presented here paves way to creating a method for adding information on variation in dialect corpora in a principled way, derived empirically from the data.