The Creation of SrpKorp RS

Olja Jojić
University of East Sarajevo
olja.jojic@ffuis.edu.ba

Darko Drakulić
University of East Sarajevo
darko.drakulic@ffuis.edu.ba

The paper describes the beginner's steps in the creation of *SrpKorp RS*, the first electronic corpus of contemporary Serbian language as used in Republika Srpska (Serb Republic). After Yugoslavia split up into separate states, Serbo-Croatian language that was spoken in Bosnia and Herzegovina up to then, was replaced by three languages now known outside of the country as Bosnian, Croatian and Serbian (BCS). These are now official languages in the state's two component units: Federacija Bosne i Hercegovine (the Federation of Bosnia and Herzegovina, with predominantly Muslim population) and Republika Srpska (Serb Republic, with predominantly Serb population). Due to reasons based on language planning/policy in Republika Srpska, the aim was to create a corpus of the Serbian language as used in this part of the State. The general-purpose corpus project started in 2015, as a joint project of the Departments of Information Science and Mathematics, Serbian Language and Literature and English Language and Literature at the Faculty of Philosophy (University of East Sarajevo). Currently, the size of the corpus is around 1 million words.

The corpus includes (mostly) complete and also composite written texts from the following genres: creative texts (novels and short stories), legal texts (legal judgments, official gazettes), informational texts- reportage (mostly electronic sources) and informational texts- learned (scholarly articles in humanities, social sciences, natural sciences). At this initial stage no transcripts of spoken language have been included. Even though this was not initially planned, at the moment there is a clear bias towards creative texts (particularly novels). Future expansion will be aimed at removing this bias. Regarding creative texts, permissions were obtained to use copywrited material in the corpus, and the materials were collected with relative ease. The rest of the texts were collected from the Internet sources. In text preprocessing stage problems were encountered with various formats of uploaded texts, particularly legal texts, and many were rejected as unsalvageable. The pictures and tables were removed at this stage of corpus compilation, as were all references to them in the respective text. The length of the text samples vary from 100 words to 90000 words. The shortest texts belong to the category of information texts-reportage, and the lengthiest are creative texts. Since the purpose of the project was to create a corpus of contemporary Serbian language, the compiled texts were written from 2000 onwards. In order to represent the various dialects of Serbian language, the compilation was planned so as to include samples from all major regions of Republika Srpska. Since both Cyrillic and Latin alphabets are used in Serbian language, the texts that were compiled were kept in their original alphabet. The plan was to enable the corpus search in both alphabets. The corpus currently contains annotation only for bibliographical data. Further development of the corpus is envisioned, in the direction of morphological annotation and expansion of the capacity of the corpus.