Nikola Ljubešić
Jožef Stefan Institute
Faculty of Computer and Information Science,
University of Ljubljana
nljubesi@gmail.com

Tanja Samardžić
University of Zürich
tanja.samardzic@uzh.ch

Tomaž Erjavec
Jožef Stefan Institute
Faculty of Computer and Information Science,
University of Ljubljanatomaz.erjavec@ijs.si

Darja Fišer
University of Ljubljana
darja.fiser@ff.uni-lj.si

Maja Miličević Petrović
University of Belgrade
m.milicevic@fil.bg.ac.rs

Simon Krek
Centre for language resources and technologies,
University of Ljubljana simon.krek@guest.arnes.si

Vuk Batanović
University of Belgrade
vukbatanovic@sbb.rs

The "ReLDI effect": Collaborative development of manually annotated datasets for Slovene, Croatian and Serbian

With the rapid development and increasing accessibility of natural language processing (NLP) techniques, the exploitation of NLP inside electronic lexicography is on a rise. Textual datasets manually annotated with linguistic information are a backbone of the currently dominating

paradigm in NLP based on supervised machine learning. However, developing such manually annotated datasets is a very costly activity, which is one of the reasons for limited availability of NLP technologies for languages with fewer speakers, and especially for less dominant language varieties such as the language of the Internet.

In this talk we present a series of collaborations between researchers developing such datasets for Slovene, Croatian and Serbian, three languages with just a few million speakers each. Close relatedness of these languages brings an opportunity for a synchronized approach to the development of resources and technologies, to the benefit of all parties. Due to the complex political environment, however, such an approach has not been established until the start of the ReLDI (Regional Linguistic Data Initiative) project. The main synergistic effect of the collaborations presented here is achieved by drastically lowering the efforts required to produce datasets in additional languages, primarily in the areas of (1) the development of annotation guidelines, (2) setting up the technical requirements for the annotation campaigns and (3) pre-annotation of data with models trained for another, but very close language.

The linguistic levels covered in the resulting datasets are those of tokenisation, sentence segmentation, normalisation, morphosyntax, lemmatisation, dependency parsing, semantic role labeling, named entity recognition and coreference resolution. Two varieties of each of the three languages are covered: the standard variety and the variety of the language of the Internet.