Tanja Samardžić (University of Zurich) **& Maja Miličević** (University of Belgrade) **& Nikola Ljubešić** (University of Zagreb)

ReLDI resources for Croatian and Serbian

An important global trend in the study of human language in the past decades has been a growing reliance on empirical data. This is due to the rapid development and wider accessibility of large collections of machine-readable text and data-driven tools needed for its processing. Due to the difficulties brought by the transition period following the breakup of former Yugoslavia, Serbia and Croatia lag behind most other European countries in the implementation of these trends. Regional Linguistic Data Initiative (ReLDI) is a two-year institutional partnership between research units in Switzerland, Serbia and Croatia, funded by the Swiss National Science Foundation grant No. 1605011[1] with the main objective to improve resources and research methodology in the domain of language studies based on empirical data, focusing on Croatian and Serbian and fully exploiting their close relatedness.

The initiative offers resources integrated within a web platform comprising portions dedicated to corpus-based and experimental research, complete with detailed documentation, as well as an e-learning environment for web courses and tutorials dedicated to resource use. The resources are embedded in an educational component intended to equip the interested researchers with the skills needed to fully exploit the resources shared through our initiative. In addition to the online courses, the educational activities will also include traditional live seminars, serving as both training and networking events.

The collected data and tools are mostly managed through the infrastructure at the University of Zurich. In collaboration with the S3IT support service, we have set up a virtual server running most of the software used in the project:

- WordPress for the main project website / access point for data and tools
- WebAnno for collaborative manual annotation of language corpora
- NoSketch Engine for searching corpora
- R and Python for data processing
- EdX for online courses

We also use a public GitHub repository to share the source code of specialised NLP tools and the related documentation.

The courses are based on the current teaching activities of the three partners. They cover issues in three maindomains:

- Methodological: general principles of experimental design, corpus-based studies, statistical analysis, basics of machine learning
- Theoretical: the role of data in language science, corpus annotation as a form of linguistic analysis
- Technical: data processing with R and Python, data visualisation, use of annotation tools and other NLP resources

In this talk, I will present the resources and the courses offered by the partnership with suggestions on how they can be used for studying clausal complementation.

[1] http://p3.snf.ch/project-160501