Regional Linguistic Data Initiative (ReLDI)

Tanja Samardžić
Corpus Lab
URPP Language and Space
University of Zurich
tanja.samardzic@uzh.ch

Nikola Ljubešić
Dept. of Information
and Communications Sciences
Faculty of Humanities
and Social Sciences
University of Zagreb
nljubesi@ffzg.hr

Maja Miličević
Dept. of General Linguistics
Faculty of Philology
University of Belgrade
m.milicevic@fil.bg.ac.rs

1 Introduction

Regional Linguistic Data Initiative is a two-year institutional partnership between research units in Switzerland, Serbia and Croatia, funded by the Swiss National Science Foundation grant No. 160501.[1] The partners in the project are the authors of this article. The goal of the partnership is two-fold. First, we will collect and distribute various kinds of linguistic data and tools to support empirical research on Croatian and Serbian. Second, we will organise didactic activities and establish a regional community of researches who will use these data and tools in their research and teaching. In this paper, we describe the key components of the partnership.

2 The Infrastructure

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.

3 Online Content

We collect and distribute two main kinds of data and tools: a natural language processing set and an experimental set. Both sets of resources are associated with courses and tutorials.

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

3.1 NLP Resources and Tools

The natural language processing set consists of Croatian and Serbian corpora, morphological dictionaries and processing tools. While such resources and tools already exist for both languages (Vitas et al., 2003; Agić et al., 2008), they are mostly inaccessible to researches outside the groups that develop them. Our aim is to compile and distribute resources that will be available to all interested researches.

Corpora include smaller manually annotated samples[2] and large automatically annotated corpora (Ljubešić and Klubička, 2014); annotation will be improved and enriched in the course of the project. Free existing morphological dictionaries[3] are extended inside the project in a semi-automated fashion (Ljubešić et al., in press). Tools currently include a state-of-the-art part-of-speech tagger and lemmatiser reaching a new best performance for both languages ($\sim$91% for full morphosyntactic annotation and $\sim$97% for lemmatisation). Development of tools for other kinds of analysis (dependency syntax, semantic role labelling) is planned for the remainder of the project. In addition to the standard tools, we provide a set of scripts for extracting corpus data commonly needed for quantitative linguistic analysis (e.g., for extracting and comparing frequency lists), and scripts for format conversion and file handling. Special emphasis is placed on detailed documentation of all resources.

The presented resources and tools for Croatian are currently more developed than those for Serbian. We take advantage of the large structural and lexical overlap between the two languages to develop Serbian resources starting from the existing Croatian ones.

[2] https://github.com/ffnlp/sethr

[3] https://svn.code.sf.net/p/apertium/svn/languages/apertium-hbs/

Proceedings of the 5th Workshop on Balto-Slavic Natural Language Processing, pages 40–42,
Hissar, Bulgaria, 10–11 September 2015.

3.2 Linguistic Experimental Data

Another important empirical trend in language science concerns experimental research. Linguists increasingly rely on sampling and statistical processing of human judgements about and their reactions to linguistic phenomena (acceptability judgements, reading times, etc.; see e.g., Kraš and Miličević (in press)). Currently, such empirical data tend to be used only within the studies for which they are collected. Through our partnership, experimental data for Croatian and Serbian will be collected and their wider distribution and reuse encouraged. Both instruments (stimuli lists) and results will be included. Papers published based on the given stimuli and results will serve as documentation. Similar initiatives are still rare in linguistics (but see Marsden and Mackey (2014) for instruments in second language acquisition research), so our work in this domain is largely pioneering even beyond the regional context.

3.3 Online Courses and Tutorials

One of the major obstacles to a wider use of both corpus and experimental linguistic data is a lack of skills required for obtaining and analysing them. Despite the growing demand, experimental design, data manipulation and statistical analysis are not yet covered in standard linguistic curricula. An important part of our initiative is thus devoted to online courses and tutorials.

The educational component of the project is intended to equip the interested researchers with the skills needed to fully exploit the resources shared through our initiative. The courses are based on the current teaching activities of the three partners. They cover issues in three main domains:

- 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

All courses will include exercises in which participants will have an opportunity to use the data and tools collected within the project. The courses will emphasise the points in common between the analysis of corpus and experimental data.

4 Activities in the Region

The work on creating a regional research community will be centred around four three-day workshops, two in Belgrade and two in Zagreb, which are planned for the second project year. The targeted participants are graduate students and researchers at universities and institutes, joined by professionals from companies that work with linguistic data.

All four workshops will be composed of invited talks, tutorials given by the project partners (based on the online courses), and a range of activities geared towards encouraging exchange and networking between the participants. Each workshop will have two invited speakers – internationally recognised experts in linguistics or NLP who have worked on Croatian and/or Serbian. Tutorials will have the form of live classes based on the online materials, with hands-on sessions and practical exercises. Exchange and networking will take place during panel discussion and social events. To facilitate participant mobility within the region, we will offer a number of small grants covering travel and accommodation costs.

Different activities will be undertaken to advertise the workshops, promoting at the same time the project goals: the project website (currently under construction), local media, social media, and presentations at various academic events.

5 Expected Outcomes

The partnership is expected to result in a community of researchers working with linguistic data in a shared empirical framework, exchanging ideas, and adhering to common research quality standards. Some of the contacts established through the initiative are expected to result in research collaborations that will extend beyond the duration of the partnership; these collaborations should bring about new research ideas and new projects.

The data and training provided through the initiative are expected to increase the competitiveness of researchers from the region in the international context. The initiative will also help researches contribute to the study of language beyond their specific subject languages.

Finally, Regional Linguistic Data Initiative is expected to help establishing contacts between researchers and professionals in the domain of language technology, identifying common interests and potential for collaboration.

References

Željko Agić, Marko Tadić, and Zdravko Dovedan. 2008. Improving part-of-speech tagging accuracy for Croatian by morphological analysis. *Informatica*, 32(4):445–451.

Tihana Kraš and Maja Miličević. in press. *Eksperimentalne metode u istraživanjima usvajanja drugoga jezika*. Filozofski fakultet, Rijeka.

Nikola Ljubešić and Filip Klubička. 2014. {bs,hr,sr}WaC – web corpora of Bosnian, Croatian and Serbian. In *Proceedings of the 9th Web as Corpus Workshop (WaC-9)*, pages 29–35, Gothenburg, Sweden. Association for Computational Linguistics.

Nikola Ljubešić, Miquel Esplà-Gomis, Filip Klubička, and Nives Mikelić Preradović. in press. Predicting inflectional paradigms and lemmata of unknown words for semi-automatic expansion of morphological lexicons. In *Proceedings of Recent Advances in Natural Language Processing*, Hissar, Bulgaria.

Emma Marsden and Alison Mackey. 2014. IRIS: A new resource for second language research. *Linguistic Approaches to Bilingualism*, (4):125–130.

Duško Vitas, Cvetana Krstev, Ivan Obradović, Ljubomir Popović, and Gordana Pavlović-Lažetić. 2003. An overview of resources and basic tools for processing of Serbian written texts. In *Proceedings of the Workshop on Balkan Language Resources, 1st Balkan Conference in Informatics*, pages 97–104, Thessaloniki, Greece.