SETimes.SR – A Reference Training Corpus of Serbian

Vuk Batanović[1], Nikola Ljubešić[2], Tanja Samardžić[3]

[1]School of Electrical Engineering, University of Belgrade
[2]Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana
[3]Language and Space Lab, University of Zürich

JTDH 2018, Ljubljana
September 20 – 21, 2018

Introduction

- Manually annotated datasets – crucial for today's NLP, which is based on supervised machine learning
- Extreme scarcity of annotated corpora in Serbian
- Parallel SETimes corpus – news stories in English and languages of southeast Europe
- Chance for cross-linguistic annotation transfer from Croatian to Serbian
- SETimes.SR
  - Selected news stories from the SETimes corpus written in Serbian using the Ekavian pronunciation and the Serbian Latin script
  - Parallel on the document level with the earlier SETimes.HR corpus in Croatian
  - The first publicly available multi-annotated corpus in Serbian that covers all the basic elements of a full NLP pipeline

Corpus overview

- SETimes.SR is annotated on the levels of:
  1. document, sentence and token segmentation,
  2. morphosyntax,
  3. lemmas,
  4. dependency syntax,
  5. named entities.

Documents	163
Sentences	3,891
Tokens	86,726
Types	17,586
Lemmas	8,619
MSDs	557

Morphosyntax and lemmas

Morphosyntax

- Annotated according to the MULTEXT-East V5 specifications
  - Choice motivated by our desire to keep the tagset as close as possible to the one used in SETimes.HR
  - 13 POS categories
- First step – process the Serbian corpus with the best-performing model for Croatian (Ljubešić et al., 2016)
- Second step – manually correct the output (two expert annotators)
- Automatic mapping to UPOS, except for abbreviations

Lemmas

- Harmonization with the inflectional lexicon srLex

Dependency syntax

- Annotated according to UDv2 specification
- First step – process the Serbian corpus with the most up-to-date Croatian model (Agić and Ljubešić, 2015)
- Second step – manual correction (made in 14% of syntactic edges)
- Inter-annotator agreement
  - Measured on a sample of 300 sentences
  - Four annotators – three Croatian native speakers and one Serbian
  - Agreement measure – proportion of identically annotated tokens (same MSD, dependency link and label)
  - Average agreement – close to 93%

Named entity recognition

- Annotation guidelines taken from the Slovene ssj500k corpus (the same ones used in hr500k NE annotation)
- WebAnno annotation by two annotators, conflict resolution by a third annotator
- High saturation of texts with named entities - 42 per document on average, or almost two per sentence
- PER (27%), DERIV-PER (1%), LOC (39%), ORG (28%), MISC (4%)
- Regardless of the strict guidelines and double annotation, some imperfections on the global level still remain

Comparison with SETimes.HR

	SR	HR
Documents	163	163
Sentences	3 891	3 757
Tokens	86 726	83 630

- SETimes.HR preceded SETimes.SR and was instrumental in its creation
- Largest statistically significant differences wrt conjunctions, particularly subordinating conjunctions
  - "*Da*" subordinating conjunction is used much more frequently in Serbian since it appears in complex predicates involving modal and phase verbs, as well as within a complex form of the future tense (SETimes.SR: 2302, SETimes.HR: 507, $\tilde{\chi}^2 = 1099.97$, $p =$ **3.3E-241**, $\Phi = 0.08035$)
- No statistically significant differences wrt named entities

Corpus availability

Download

http://hdl.handle.net/11356/1200

Search

KonText:

https://www.clarin.si/kontext/first_form?corpname=setimes_sr

NoSketch Engine:

https://www.clarin.si/noske/run.cgi/corp_info?corpname=setimes_sr

Models

Annotation tools for Serbian at:

https://github.com/clarinsi/

Planned future activities

Dataset extension

- Enlarge the corpus with additional 500 sentences drawn from other news sources
- Generate new annotation layers
  - Work on a coreference layer is under way

Dataset consolidation

- Insert missing content and achieve maximal sentence parallelism between SETimes.HR and SETimes.SR
- Harmonize specific annotation layers within the SETimes.SR dataset, as well as across languages

Conclusion

- SETimes.SR – the first publicly available gold standard corpus of Serbian annotated on the level of document, sentence, and token segmentation, morphosyntax, lemmas, dependency syntax, and named entities
- Published under a permissive CC license (CC BY-SA 4.0) – feel free to use it!
- Example of a successful linguistic transfer approach to building resources for closely related languages

SETimes.SR – A Reference Training Corpus of Serbian

Vuk Batanović[1], Nikola Ljubešić[2], Tanja Samardžić[3]

[1]School of Electrical Engineering, University of Belgrade
[2]Department of Knowledge Technologies, Jožef Stefan Institute, Ljubljana
[3]Language and Space Lab, University of Zürich

JTDH 2018, Ljubljana
September 20 – 21, 2018