

ReLDI+JANES data and tools

Nikola Ljubešić

Dept. of Knowledge Technologies, Jožef Stefan Institute, Ljubljana

Dept. of Information and Communication Sciences,
Faculty of Humanities and Social Sciences, University of Zagreb

CLARIN workshop, Ljubljana, 10th Nov 2016

Overview

- 1 Projects
- 2 Normalisation
- 3 Morphosyntax
- 4 Dependency syntax
- 5 Ongoing developments
- 6 ReLDI ecosystem

Projects

ReLDI

- ReLDI – Regional Linguistic Data Initiative
- 2015–2017
- <https://reldi.spur.uzh.ch>
- Funded by the Swiss National Science Foundation inside the SCOPUS (Scientific Cooperation between Eastern Europe and Switzerland) programme
- Partners
 - University of Zürich (Tanja Samardžić)
 - University of Belgrade (Maja Miličević)
 - University of Zagreb (Nikola Ljubešić)
- Relevant task: share expertise in developing datasets and tools for both languages

JANES

- JANES – Linguistic Analysis of Non-standard Slovene
- 2014–2017
- <http://nl.ijs.si/janes/>
- Funded by the Slovene National Science Foundation
- Partners
 - Faculty of Arts (Darja Fišer)
 - Jožef Stefan Institute (Tomaž Erjavec)
- Relevant task: develop datasets and tool for processing non-standard Slovene

Interaction between the two projects

- Languages
 - ReLDI: Croatian and Serbian
 - JANES: Slovene
- Technologies
 - ReLDI: cutting-edge standard language technologies
 - JANES: technologies for non-standard language
- Synergy
 - ReLDI: annotating Croatian and Serbian non-standard data following the JANES guidelines
 - JANES: using ReLDI tools for Slovene, adapting them to non-standard language
- Limited funding, interaction between multiple teams and individuals

Normalisation

Diacritic restoration

Data

	Slovene	Croatian	Serbian
Wikipedia	20m	28m	34m
Web	131m	269m	103m
Twitter	7m	2m	14m

Tool

- <https://github.com/clarinsi/redi>
- *Jaz se ne vem* → *Jaz še ne vem*
- Token-by-token transformation via the noisy channel model
- Transformation probability ($p(\text{se}|\text{se})$ and $p(\text{še}|\text{se})$) and target language probability ($p(\text{jaz se ne vem})$, $p(\text{jaz še ne vem})$)
- Accuracy 99.2% on non-standard and 99.6% on standard data (weak baseline ~86%)

Normalisation via character-level SMT

CMC data

	Slovene	Croatian	Serbian
Twitter	102k	89k	92k
Blog	21k	-	-
Forum	38k	-	-
Comments	23k	-	-

Historical data

- Available only for Slovene
- goo300k corpus of historical Slovene from the 18th and 19th century

Normalisation via character-level SMT

Tool

- <https://github.com/clarinsi/csmtiser>
- Wrapper around SMT system Moses to train and apply a character-level translation model
_ a m a m _ k r _ p r o v _ → _ a _ i m a m _ k a r _ p r a v _
- Significant improvements by using multiple language models
- Error reduction against weak baseline of 50-90% on both CMC and historical data

Morphosyntax

Tagging

Data

	Slovene	Croatian	Serbian
Annotated corpus	500k	497k	(→)497k
Inflectional lexicon	100k	187k	193k

Tool

- <https://github.com/clarinsi/reldi-tagger>
- CRF with unconstrained tagging – lexicon entries as features
- Tagging accuracy:

	Slovene	Croatian	Serbian
ReLDI-tagger	94.27%	92.53%	92.33%
HunPos	91.67%	89.30%	87.20%
RFTagger	91.84%		

Lemmatisation

Data

	Slovene	Croatian	Serbian
Annotated corpus	500k	497k	(→)497k
Inflectional lexicon	100k	187k	193k

Tool

- Part of the ReLDI-tagger
- MSD (morphosyntactic description) annotation prerequisite
- (token, MSD) pairs seen in training data or lexicon via lemma frequency
- For unseen pairs classifier per MSD which predicts quadruples (prefix_cut, prefix, suffix_cut, suffix)
- *najdražemu* → *drag* (3, '', 4, 'g')

Tagging non-standard text (work in progress)

Data

	Slovene	Croatian	Serbian
Twitter	55k	89k	92k
Blog	5k	-	-
Forum	9k	-	-
Comments	8k	-	-

- Combination with standard data

Tool

- <https://github.com/clarinsi/reldi-tagger>
- Extending the feature set with Brown clusters of various depth
- Brown clusters – hierarchical word clustering technique, based on words' contexts

```
^0100101001100 jaz jst jest js jz
```

Dependency syntax

Dependency syntax

Data

	Slovene	Croatian	Serbian
Universal Dependencies	140k	139k	87k

Tool

- MateTools parser
- Labeled attachment score ~80%

Ongoing developments

Ongoing developments

Named entity recognition

- Datasets and StanfordNER models for Slovene and Croatian already available
- Recently started a new campaign with unified annotation guidelines for all three languages
- Work on our own CRF-based tool

Semantic role labeling

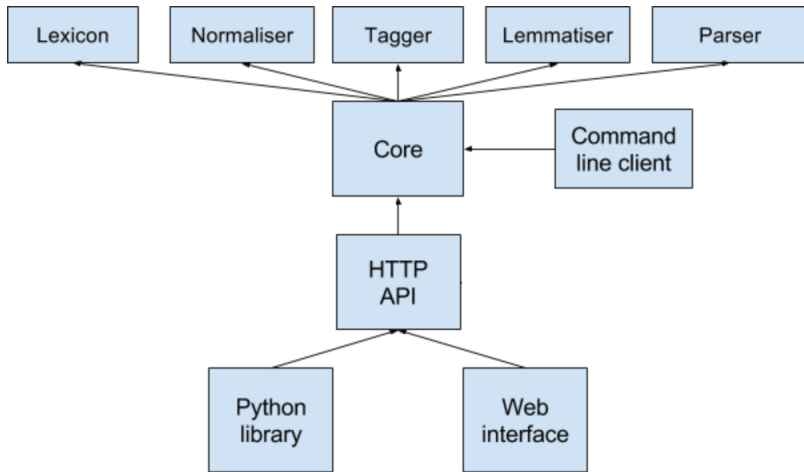
- Slovene-Croatian bilateral project
- Annotate UD data with shallow semantics
- CroVallex lexicon with no counterparts in other languages
- Produce valency lexicons that are in tune with corpus data

Current status of the technologies

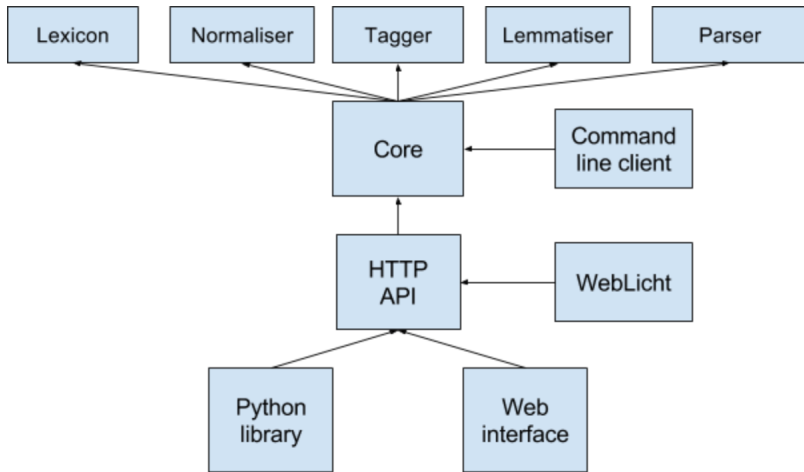
	tool	web service
tokenisation	production	production
sentence splitting	production	production
diacritic restoration	production	production
normalisation	production	
tagging	production	production
non-standard tagging	development	
lemmatisation	production	production
dependency parsing	development	development
semantic role labeling	data annotation	
named entity recognition	development	

ReLDI ecosystem

ReLDI ecosystem overview



ReLDI ecosystem overview + WebLicht



HTTP API

- `http://faustjr.ffzg.hr:8080/api/v1/authorized/hr/tag_lemmatise_depparse?format=json&text=0vo%20je%20primjer%20otvorenog%20servisa.%20Postoji%20i%20onaj%20zatvoreni.&request-id=10`

Web interface

- `http://nl.ijs.si/services`

Python library

- <https://github.com/clarinsi/reldi-lib-doc>

ReLDI+JANES data and tools

Nikola Ljubešić

Dept. of Knowledge Technologies, Jožef Stefan Institute, Ljubljana

Dept. of Information and Communication Sciences,
Faculty of Humanities and Social Sciences, University of Zagreb

CLARIN workshop, Ljubljana, 10th Nov 2016