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

Normalisation

Morphosyntax

Dependency syntax

Ongoing developments

ReLDI ecosystem
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 – 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

<table>
<thead>
<tr>
<th></th>
<th>Slovene</th>
<th>Croatian</th>
<th>Serbian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikipedia</td>
<td>20m</td>
<td>28m</td>
<td>34m</td>
</tr>
<tr>
<td>Web</td>
<td>131m</td>
<td>269m</td>
<td>103m</td>
</tr>
<tr>
<td>Twitter</td>
<td>7m</td>
<td>2m</td>
<td>14m</td>
</tr>
</tbody>
</table>

Tool

- [https://github.com/clarinsi/redi](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 target language probability \( p(\text{še}|\text{se}) \) 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 \( \sim 86\% \))
Normalisation via character-level SMT

CMC data

<table>
<thead>
<tr>
<th></th>
<th>Slovene</th>
<th>Croatian</th>
<th>Serbian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter</td>
<td>102k</td>
<td>89k</td>
<td>92k</td>
</tr>
<tr>
<td>Blog</td>
<td>21k</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Forum</td>
<td>38k</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Comments</td>
<td>23k</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

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
  \[ \text{\_a m a m \_k r \_p r o v \_} \rightarrow \text{\_a \_i m a m \_k a r \_p r a v \_} \]
- Significant improvements by using multiple language models
- Error reduction against weak baseline of 50-90% on both CMC and historical data
Morphosyntax
Tagging

**Data**

<table>
<thead>
<tr>
<th></th>
<th>Slovene</th>
<th>Croatian</th>
<th>Serbian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotated corpus</td>
<td>500k</td>
<td>497k</td>
<td>(→)497k</td>
</tr>
<tr>
<td>Inflectional lexicon</td>
<td>100k</td>
<td>187k</td>
<td>193k</td>
</tr>
</tbody>
</table>

**Tool**

- [https://github.com/clarinsi/reldi-tagger](https://github.com/clarinsi/reldi-tagger)
- CRF with unconstrained tagging – lexicon entries as features
- Tagging accuracy:
  - ReLDI-tagger: 94.27% Slovene, 92.53% Croatian, 92.33% Serbian
  - HunPos: 91.67% Slovene, 89.30% Croatian, 87.20% Serbian
  - RFTagger: 91.84%
Lemmatisation

Data

<table>
<thead>
<tr>
<th></th>
<th>Slovene</th>
<th>Croatian</th>
<th>Serbian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Annotated corpus</td>
<td>500k</td>
<td>497k</td>
<td>(→)497k</td>
</tr>
<tr>
<td>Inflectional lexicon</td>
<td>100k</td>
<td>187k</td>
<td>193k</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th></th>
<th>Slovene</th>
<th>Croatian</th>
<th>Serbian</th>
</tr>
</thead>
<tbody>
<tr>
<td>Twitter</td>
<td>55k</td>
<td>89k</td>
<td>92k</td>
</tr>
<tr>
<td>Blog</td>
<td>5k</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Forum</td>
<td>9k</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Comments</td>
<td>8k</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

- 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

<table>
<thead>
<tr>
<th>Language</th>
<th>Universal Dependencies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slovene</td>
<td>140k</td>
</tr>
<tr>
<td>Croatian</td>
<td>139k</td>
</tr>
<tr>
<td>Serbian</td>
<td>87k</td>
</tr>
</tbody>
</table>

Tool

- MateTools parser
- Labeled attachment score $\sim80\%$
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

<table>
<thead>
<tr>
<th>Tool</th>
<th>Web service</th>
</tr>
</thead>
<tbody>
<tr>
<td>tokenisation</td>
<td>production</td>
</tr>
<tr>
<td>sentence splitting</td>
<td>production</td>
</tr>
<tr>
<td>diacritic restoration</td>
<td>production</td>
</tr>
<tr>
<td>normalisation</td>
<td>production</td>
</tr>
<tr>
<td>tagging</td>
<td>production</td>
</tr>
<tr>
<td>non-standard tagging</td>
<td>production</td>
</tr>
<tr>
<td>lemmatisation</td>
<td>production</td>
</tr>
<tr>
<td>dependency parsing</td>
<td>development</td>
</tr>
<tr>
<td>semantic role labeling</td>
<td>development</td>
</tr>
<tr>
<td>named entity recognition</td>
<td>development</td>
</tr>
</tbody>
</table>

- **tool**: production, development
- **web service**: production, development
ReLDI ecosystem
ReLDI ecosystem overview

- Lexicon
- Normaliser
- Tagger
- Lemmatiser
- Parser

Core

HTTP API

- Python library
- Web interface

Command line client
ReLDI ecosystem overview + WebLicht

- Lexicon
- Normaliser
- Tagger
- Lemmatiser
- Parser

Core

- Command line client
- HTTP API
- WebLicht

- Python library
- Web interface
HTTP API

- http://faustjr.ffzg.hr:8080/api/v1/unauthorized/hr/tag_lemmatise_depparse?format=json&text=Ovo%20je%20primjer%20otvorenog%20servisa.%20Postoji%20onaj%20zatvorenii.&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