Development and Applications of the Croatian 1984 Corpus for the MULTEXT-East Resources

Željko Agić**, Danijela Merkler*, Daša Berović*, Marko Tadić*

* Department of Linguistics
** Department of Information Sciences
Faculty of Humanities and Social Sciences, University of Zagreb

{zagic, dmerkler, dberovic, marko.tadic}@ffzg.hr

SlaviCorp 2011
Dubrovnik
2011-09-14
MULTEXT-East project

- MULTEXT-East resources
  - multilingual dataset for LT research and development
  - covering today
    - Bulgarian, Croatian, Czech, English, Estonian, Hungarian, Lithuanian, Macedonian, Persian, Polish, Resian, Romanian, Russian, Serbian, Slovak, Slovene, and Ukrainian
  - language resources
    - morphosyntactic specifications
      - Croatian included since 1998
    - lexica
    - annotated 1984 corpus
    - MULTEXT-East parallel and comparable text and speech corpora
    - associated documentation
The 1984 corpus

- the central component of the MULTEXT-East corpus
- XML marked-up in accordance with the TEI recommendations
- annotation with hand validated MSDs and lemmas
  - suitable for MSD tagging and lemmatisation experiments
- separate alignment files
  - hand-validated pair-wise sentence alignments between English and the translations
- version 4 adds pair-wise alignments between all the languages
  - automatically induced from the alignments with English
The 1984 corpus

- the Croatian translation of 1984 semi-manually annotated for lemmas and morphosyntactic tags
  - first step
    - text of 1984 and Croatian Morphological lexicon matched yielding all lemma and MSD interpretations for tokens
    - manual selection of proper interpretation by 20 students
  - second step
    - two expert annotators
    - checking the corpus for errors in the previous round
      - sentence segmentation, tokenization, lemmatization, assignment of MSD tags
    - overlap of 1/4 of the overall corpus size in sentences
      - in order to calculate the inter-annotator agreement on tokens, lemmas and morphosyntactic tags, including PoS and other MSD categories
Problems

- manual verification of the semi-manually lemmatized and MSD-tagged corpus
- two kinds of problems
  - problems of processing
    - incorrect lemmatization and MSD tagging
    - such errors were corrected
  - problems of the text
    - errors in lemmatization and MSD tagging because of the errors in the text itself
    - we did not intervene, errors were marked
Problems of processing

- correcting problems of processing by consulting two Croatian grammars and a Croatian dictionary
  - nouns with numerals 2, 3 and 4
  - adverbialised nouns and pronouns
  - conjunctions
  - modal particles
Nouns with numerals 2, 3 and 4

- nouns that occur with the numbers 2, 3 and 4 always appear in a special morphosyntactic category
- masculine, neuter and feminine nouns corrected to genitive case singular

\[
\begin{align*}
\text{tri (tri, Mc-p-l) } & \text{ čovjeka (čovjek, } \text{Ncmsa--y)} \rightarrow \text{tri (tri, Mc-p-l) } \text{čovjeka (čovjek, Ncmsg)} \\
\text{dva (dva, Mc-p-l) } & \text{mišljenja (mišljenje, } \text{Ncnpa)} \rightarrow \text{dva (dva, Mc-p-l) } \text{mišljenja (mišljenje, Ncns}}
\end{align*}
\]
Adverbialised nouns and pronouns

- adverbialization of nouns
  - result of lexicalization of certain word-forms that become a new lexical unit and move to new PoS
  
  *Mjesecima prije toga nije se dijelila čokolada.*
  
  mjesec, Ncmpd
  mjesecima, Rt

- adverbialised pronouns
  - occur as intensifiers of an adverb
  
  *Onda se udalji od mene što brže možeš.*
  
  što, Pi3n-n--n-n-n
  što, Rn
Conjunctions

- different types of words can have function of the conjunction in the sentence
  - adverbs (pronominal adverbs with interrogative semantics)
    \[ U \text{ tom trenu, iznenada mu je sinulo } \text{kako} \text{ su potpuno sami.} \]
    \[ \text{kako, Rn} \]
    \[ \text{kako, Css} \]
  - pronouns (relative - Što, koji → conjunctions introduce causal, temporal and comparative clauses)
    \[ Ubrzo su bili u neprilici Što tu sjede tako šutke. \]
    \[ Što, Pi3n-a--n-n-n \]
    \[ Što,Css \]
Conjunctions

- combination of multiple words
  - conjunction + conjunction or
  - conjunction + non-conjunction word (adverb, present participle, etc.)

Ali **budući** da u zbilji Veliki Brat nije svemoćan, a Partija nije nepogrešiva, postoji stalna potreba za neumornom i uvijek budnom elastičnošću u postupanju s činjenicama.

**budući biti1 Vcpp, da Csp**

**budući Csc, da Csc**
Modal particles

- many words in a sentence can have a function of modal particles
- in the most cases in 1984 these are conjunctions *i* and *ni*, that have function of intensifiers

...*presjeći će se i posljednja karika koja vezuje s prošlošću.*

*i, Css*

*i, Qo*

**Ni tisuću je raketnih bombi ne bi razorilo.**

*ni, Css*

*ni, Qo*
Problems of the text

- problems of the text imply errors in lemmatization and MSD tagging because of the errors in the text itself:
  - spelling errors: čeovjek (čovjek), šte (što), čene (neče), veli (voli), i i...
  - grammar and spelling errors: zaspe < zasuti (zaspi < zaspati)...
  - disagreement: ovaj (Pd-msn--n-a--) puta (Ncmsg), urazini oko (Spsg) metar (Ncmsga--n), poslije (Spsg) podne (Ncnsn)...
- not corrected, but marked for possible later correction
Corpus stats

- 6625 sentences, 106632 tokens
  - 18846 different wordforms, 8671 different lemmas
- annotated by 802 different MSD-tags
- distribution of word types in the corpus

<table>
<thead>
<tr>
<th>Word Type</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>adjective</td>
<td>9244</td>
</tr>
<tr>
<td>conjunction</td>
<td>8553</td>
</tr>
<tr>
<td>interjunction</td>
<td>45</td>
</tr>
<tr>
<td>numeral</td>
<td>1177</td>
</tr>
<tr>
<td>noun</td>
<td>20524</td>
</tr>
<tr>
<td>pronoun</td>
<td>11627</td>
</tr>
<tr>
<td>particle</td>
<td>1680</td>
</tr>
<tr>
<td>adverb</td>
<td>7881</td>
</tr>
<tr>
<td>adpostition</td>
<td>7760</td>
</tr>
<tr>
<td>verb</td>
<td>23659</td>
</tr>
<tr>
<td>abbreviation</td>
<td>56</td>
</tr>
<tr>
<td>other</td>
<td>14426</td>
</tr>
</tbody>
</table>
Corpus stats

- distribution of different MSD-tags on parts of speech
- adjectives, nouns, pronouns and verbs expected to be the most difficult to annotate

<table>
<thead>
<tr>
<th>Part of Speech</th>
<th>Count</th>
</tr>
</thead>
<tbody>
<tr>
<td>adjective</td>
<td>215</td>
</tr>
<tr>
<td>conjunction</td>
<td>3</td>
</tr>
<tr>
<td>interjunction</td>
<td>2</td>
</tr>
<tr>
<td>numeral</td>
<td>58</td>
</tr>
<tr>
<td>noun</td>
<td>78</td>
</tr>
<tr>
<td>pronoun</td>
<td>329</td>
</tr>
<tr>
<td>particle</td>
<td>6</td>
</tr>
<tr>
<td>adverb</td>
<td>18</td>
</tr>
<tr>
<td>adpostition</td>
<td>8</td>
</tr>
<tr>
<td>verb</td>
<td>74</td>
</tr>
<tr>
<td>abbreviation</td>
<td>6</td>
</tr>
<tr>
<td>other</td>
<td>5</td>
</tr>
</tbody>
</table>
Experiment setup

- an illustrational experiment in morphosyntactic tagging of Croatian by using the 1984 corpus
  - CroTag HMM tagger (and lemmatizer)
  - tenfold cross-validated
  - cross-tagging by using the CW100 newspaper corpus of Croatian (differring in domain)
- four different scenarios
  - train on 1984, test on 1984
  - train on CW100, test on CW100
  - train on 1984, test on CW100
  - train on CW100, test on 1984
- results predefined by the size of the model
  - CW100 ca 10% larger than 1984 in terms of tokens and different MSDs used in the annotation
## Results

<table>
<thead>
<tr>
<th></th>
<th>overall</th>
<th>known</th>
<th>unknown</th>
</tr>
</thead>
<tbody>
<tr>
<td>1984</td>
<td>88.46</td>
<td>91.09</td>
<td>70.48</td>
</tr>
<tr>
<td>cw100</td>
<td>84.80</td>
<td>88.70</td>
<td>64.94</td>
</tr>
<tr>
<td>1984 → cw100</td>
<td>71.11</td>
<td>85.44</td>
<td>47.67</td>
</tr>
<tr>
<td>cw100 → 1984</td>
<td>79.26</td>
<td>86.43</td>
<td>59.51</td>
</tr>
</tbody>
</table>
Future work

- completion of Croatian MULTEXT-East lexica and relevant documentation
- inclusion of Croatian *1984* corpus in the next version of MULTEXT-East Resources
- use also the Croatian *1984* for the experiments that all other MULTEXT-East resources are submitted to
- usage of Croatian translation of *1984* in experiments within the ACCURAT project
  - si, ro also included
Thank you for your attention.

The research within the project Accurat leading to these results has received funding from the European Union Seventh Framework Programme (FP7/2007-2013), grant agreement nº 248347.