Parsing Texts of Non-Native Czech

Jirka Hana & Barbora Hladká

Charles University Prague
NLPTEA 2017
CzeSL – Czech as a Second Language

• Part of AKCES – Acquisition Corpora of Czech
• Essays written by non-native speakers of Czech
• A1 – C2 CEFR proficiency levels
CzeSL – two releases

• CzeSL-SGT
  • 8,600 essays, 1.1M tokens
  • http://hdl.handle.net/11234/1-162, CC BY-SA-3.0

• CzeSL-man ← we work with this here
  • 645 essays, 120K tokens
  • Manually corrected and annotated for errors
  • https://bitbucket.org/jhana/czesl, CC BY-SA-3.0
## CzeSL – number of texts by CEFR Level

<table>
<thead>
<tr>
<th>Level</th>
<th>Documents</th>
</tr>
</thead>
<tbody>
<tr>
<td>Basic user</td>
<td></td>
</tr>
<tr>
<td>A1</td>
<td>57</td>
</tr>
<tr>
<td>A1+</td>
<td>3</td>
</tr>
<tr>
<td>A2</td>
<td>111</td>
</tr>
<tr>
<td>A2+</td>
<td>145</td>
</tr>
<tr>
<td>Independent user</td>
<td></td>
</tr>
<tr>
<td>B1</td>
<td>176</td>
</tr>
<tr>
<td>B2</td>
<td>124</td>
</tr>
<tr>
<td>Proficient user</td>
<td></td>
</tr>
<tr>
<td>C1</td>
<td>12</td>
</tr>
<tr>
<td>Unknown</td>
<td>17</td>
</tr>
<tr>
<td>Total</td>
<td>645</td>
</tr>
</tbody>
</table>
Non-native and native language are different

Non-native language has:

• Errors in spelling, grammar, vocabulary, collocations
• Different distribution of vocabulary and syntactic constructions

ADAM, 18 Let

Bydlim v Cechagh už 6 měsíc.
Task

Work in progress

Ultimate: Create high-level NLP tools for non-native Czech

Intermediate: Create a parser for non-native Czech

For now: Identify top level syntactic functions (Predicate, Subject, Object, Nominal predicate)
What is the problem

• Syntactic annotation is expensive

• For non-native language, it is not always clear what the gold standard should be.

• Can non-native parser be trained on native text?
What did we do

• Annotated a corpus of non-native Czech (CzeSL) with selected syntactic functions

• Trained an MST parser on
  • a corpus of standard Czech (PDT)
  • a subset of standard Czech corpus with simple sentences (STYX)

• Evaluated the results
Syntactic annotation of CzeSL-man

- Only main syntactic functions are distinguished
  - Predicate
  - Subject
  - Object
  - Nominal predicate
Syntactic annotation of CzeSL-man

• Annotation of original text not of target hypothesis

• Standard:

Vstoupit do místnosti.
enter into room.
`Enter a room.'

Annotated as directional adjunct (obl in UD)

• Possible non-native:

Vstoupit místnost.
enter room.
Intended: `Enter a room.'

Annotated as object (obj in UD)
But some interpretation is needed ...

_Jsem Mongolska._ `I am Mongolian / from Mongolia’

Could be viewed as:

• _Jsem mongolský._ – adjective
  • nominal predicate (UD: root)
  • not in standard language

• _Jsem Mongol._ – inhabitant, noun
  • nominal predicate (UD: root)

• _Jsem z Mongolska._ – country, preposition + noun
  • adjunct (UD: obl)
Prague Dependency Treebank (PDT)

- Czech Newspaper texts
- Annotated for morphology, dependency, ...
- Example:
  Ráno půjdu se svým kamarádem na houby .
  'I will go mushrooming with my friend in the morning.'

- Available: [http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3](http://hdl.handle.net/11858/00-097C-0000-0023-1AAF-3)
STYX

• Simpler language subset of PDT
• Complex phenomena discarded
• Rule-based transformation of PDT annotation into Czech school annotation
• Available: http://hdl.handle.net/11234/1-2391
Results – Parser trained on PDT – All

$r^2 = 0.52$

A1  A2  A2+  B1  B2  C1
Results – Parser trained on PDT – Predicate

\[ r^2 = 0.94 \]
Results – Parser trained on PDT – Subject

\[ r^2 = 0.52 \]
Results – Parsers trained on PDT/Styx – All

$r^2 = 0.52$

$r^2 = 0.77$
Conclusion

• parser trained on native language corpus works for non-native corpus (at least on basic function labels)

• using simpler native language does not seem to help
Future work

• (nearly) full UD annotation
• two parallel annotations