Manually PoS-tagged corpora in the CLARIN infrastructure

Tomaž Erjavec, Jakob Lenardič and Darja Fišer

CLARIN Annual Conference 2019
1 October 2019
Leipzig, Germany
Introduction

• Motivation:
  – Manually tagged corpora provide gold-standard data as important resources for training and testing PoS/MSD-taggers

• Goal:
  – Compare encodings and tagsets of such corpora in CLARIN to determine their interoperability
The surveyed corpora

• Starting point:
  – CLARIN Resource Families overview of 74 manually annotated corpora available in the CLARIN infrastructure

• The selection:
  – 14 corpora manually annotated with PoS/MSD
  – treebanks excluded, as they’re typically smaller than corpora only annotated for PoS/MSD
## Overview

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Language</th>
<th>k-tokens</th>
<th>Tagset</th>
<th>Licence</th>
</tr>
</thead>
<tbody>
<tr>
<td>ssj500k 2.2</td>
<td>Slovenian</td>
<td>586</td>
<td>MULTEXT, UD</td>
<td>CC BY-NC-SA</td>
</tr>
<tr>
<td>Janes-Tag 2.0</td>
<td>non-standard</td>
<td>75</td>
<td>MULTEXT</td>
<td>CC BY-SA</td>
</tr>
<tr>
<td>hr500k 1.0</td>
<td>Croatian</td>
<td>500</td>
<td>MULTEXT, UD</td>
<td>CC BY-SA</td>
</tr>
<tr>
<td>ReLDI-NormTagNER-hr 2.0</td>
<td>non-standard</td>
<td>89</td>
<td>MULTEXT</td>
<td>CC BY</td>
</tr>
<tr>
<td>SETimes.SR 1.0</td>
<td>Serbian</td>
<td>87</td>
<td>MULTEXT, UD</td>
<td>CC BY-SA</td>
</tr>
<tr>
<td>ReLDI-NormTagNER-sr 2.0</td>
<td>non-standard</td>
<td>92</td>
<td>MULTEXT</td>
<td>CC BY</td>
</tr>
<tr>
<td>MDET</td>
<td>Estonian</td>
<td>513</td>
<td>MULTEXT</td>
<td>CLARIN ACA</td>
</tr>
<tr>
<td>Szeged Corpus 2.0</td>
<td>Hungarian</td>
<td>1,500</td>
<td>MULTEXT-like</td>
<td>NORED-NC-ND</td>
</tr>
<tr>
<td>MATAS</td>
<td>Lithuanian</td>
<td>1,600</td>
<td>Lith. PoS tagset</td>
<td>CLARIN ACA</td>
</tr>
<tr>
<td>NKJP1M</td>
<td>Polish</td>
<td>1,000</td>
<td>IPI PAN tagset</td>
<td>GNU GPL</td>
</tr>
<tr>
<td>CINTIL</td>
<td>Portuguese</td>
<td>1,000</td>
<td>CINTIL PoS tagset</td>
<td>ELRA</td>
</tr>
<tr>
<td>BNC Sampler</td>
<td>English</td>
<td>2,000</td>
<td>CLAWS 7</td>
<td>BNC (NORED)</td>
</tr>
<tr>
<td>MULTEXT-East 1984</td>
<td>Multiling</td>
<td>374</td>
<td>UD</td>
<td>CC BY-NC-SA</td>
</tr>
<tr>
<td>XLiME Twitter Corpus</td>
<td>Multiling</td>
<td>364</td>
<td>UD</td>
<td>MIT</td>
</tr>
</tbody>
</table>
Main findings

• Uneven distribution of languages
  – Western European languages less represented than Eastern European languages
  – Less WE language corpora are included in VLO
  – Authors of WE language corpora consider them too valuable to make them freely available?

• The most common tagset
  – MULTEXT (9/14 corpora)

• Others tagsets
  – IPI PAN (the Polish NKJP1M corpus)
  – CINTIL (the Portuguese CINTIL corpus)
  – CLAWS 7 (The English BNC samples corpus)
  – Universal Dependencies (multilingual)

• Licences usually limit the use of corpora to non-commercial
MULTEXT-East (Erjavec, 2012)

• A positional tagset
  – first character determines the part of speech
  – the rest are lexical and inflectional morphosyntactic features

• For instance (example from *Janes-Tag 2.0*; Erjavec et al. 2017)

```xml
<s>
  <w lemma="ta" ana="#Pd-nsn">To</w><c> </c>
  <w lemma="danes" ana="#Rgp">danes</w><c> </c>
  <w lemma="biti" ana="#Va-r3p-n">so</w><c> </c>
  <w lemma="zgolj" ana="#Q">zgolj</w><c> </c>
  <w lemma="slab" ana="#Agpfpn">slabe</w><c> </c>
  <w lemma="igralka" ana="#Ncfpn">igralke</w>
  <pc ana="#Z">.</pc>
</s>
```

• The MSD tags for the N *igralke*:
  \textbf{N}(noun) \textbf{c}(ommon) \textbf{f}(eminine) \textbf{p}(lural) \textbf{n}(ominative)
The IPI PAN tagset

- Used in the **NKJP1M** corpus – a manually annotated subset of the *National Corpus of Polish* (Przepiórkowski, 2010)
- Uses the CTAN package for PoS categories
- Morphosyntactic features from the Morfeusz SGJP analyzer

<!-- ofierze [72,7] -->
<f name="interps">
  <fs type="lex" xml:id="morph_1.16.1-lex">
    <f name="base"><string>ofiara</string></f>
    <f name="ctag"><symbol value="subst"/></f>
    <f name="msd">
      <vAlt>
        <symbol value="sg:dat:f" xml:id="morph_1.16.1.1-msd"/>
        <symbol value="sg:loc:f" xml:id="morph_1.16.1.2-msd"/>
      </vAlt>
    </f>
  </fs>
</f>

- The N **ofierze** (‘victim’)
  - Category defined as **SUBST** (CTAN package)
  - MSD features: **SG:DAT/LOC:F**; note the case syncreticism
The Lithuanian MATAS corpus

• Uses a dedicated tagset tailored to Lithuanian (Daudaravičius et al., 2007). For instance:

```xml
<word="griežliu" lemma="griežlė" type="dktv mot.gim dgsk K">
<space>
<word="gyvenamose" lemma="gyventi(-a,-o)"
  type="dlv teig nesngr neveik.r esam.l nei_vardž mot.gim dgsk Vt">
<space>
<word="vietose" lemma="vieta" type="dktv mot.gim dgsk Vt">
<sep=".">
```

• MSD attributes are in Lithuanian
• Difficult to interpret by non-Lithuanian speakers
Intermediate summary

- All the examples discussed use XML
- However, the examples differ in MSD features and attributes, even in the case of Slovenian and Polish corpora, which both use TEI
- Nevertheless, format conversion from any of the XML schemas presented to a common one should not be too difficult
Universal dependencies

• The UD project offers the largest (100 treebanks, 70 languages) multilingual manually annotated corpus (Nivre et al. 2018)

• On the basis of this corpus, the **UD-Pipe** tool (Straka and Straková 2017) was trained

• UD-Pipe
  – annotates texts in UD languages for morphosyntactic features, lemmas, and syntactic dependencies
  – often used for PoS/MSD tagging as well

• This raises the question: *are the dedicated PoS/MSD corpora introduced in the previous section still relevant for PoS/MSD tagger training at all?*
Manually tagged PoS/MSD corpora vs. UD corpora

<table>
<thead>
<tr>
<th>Language</th>
<th>PoS/MSD corpus</th>
<th>k-tokens</th>
<th>UD treebank</th>
<th>k-tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>Slovenian</td>
<td>ssj500k 2.2</td>
<td>586</td>
<td>UD SSJ</td>
<td>140</td>
</tr>
<tr>
<td>Croatian</td>
<td>hr500k 1.0</td>
<td>500</td>
<td>UD SET</td>
<td>197</td>
</tr>
<tr>
<td>Estonian</td>
<td>MDET</td>
<td>513</td>
<td>UD EDT</td>
<td>434</td>
</tr>
<tr>
<td>Hungarian</td>
<td>Szeged Corpus 2.0</td>
<td>1,500</td>
<td>UD Szeged</td>
<td>42</td>
</tr>
<tr>
<td>Lithuanian</td>
<td>MATAS</td>
<td>1,600</td>
<td>UD ALKSNIS + HSE</td>
<td>46</td>
</tr>
<tr>
<td>Polish</td>
<td>NKJP1M</td>
<td>1,000</td>
<td>UD LFG + SZ</td>
<td>214</td>
</tr>
</tbody>
</table>

- The UD treebanks are generally much smaller in size
Comparison

• PoS/MSD taggers trained on the dedicated PoS/MSD-tagged corpora will achieve greater accuracy than those trained on UD corpora, as shown by Dobrovoljc et al. (2019)

• Important: the dedicated PoS/MSD corpora use different tagsets, while the UD counterparts use a harmonised set of features

• Consequently, there is a need to investigate the optimal conversion of native PoS tagsets to UD

• Previous work by Zhang et al. (2012) and Sagot (2018) employed automatic mapping of language-specific PoS tags to the Universal PoS tagset successfully. A similar approach could be used for our cases.
Conclusions

• Manually tagged PoS/MSD corpora in the CLARIN infrastructure should be used to train (better) PoS/MSD taggers than are currently available
  - Mapping of corpus-specific tagsets to UD morphological features would improve the accuracy of training existing multilingual taggers with a harmonised tagset

• Manually PoS/MSD tagged gold-standard corpora are part of basic language resource kits → Strategically important for CLARIN to actively encourage the integration of the existing gold-standard corpora in the infrastructure

• Quality and interoperability of such corpora should be promoted and supported through shared tasks, data camps and hackathons