Manually PoS-tagged corpora in the CLARIN infrastructure

Tomaž Erjavec, Jakob Lenardič and Darja Fišer

CLARIN Annual Conference 2019
1 October 2019
Leipzig, Germany

CLARIN

Introduction

- Motivation:
  - Manually tagged corpora provide gold-standard data → 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

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

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

```
<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*:
  N(oun) **c**(ommon) **f**(eminine) **p**(lural) **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>
```

- 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:

```
<word="griežliu¸" lemma="griežl˙e" type="dktv mot.gim dgsk K">
<space>
<word="gyvenamose" lemma="gyventi(-a,-o)"
type="dlv teig nesngr neveik.r esam.l neı¸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

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

- 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