ParlaCLARIN Workshop: Creating and Using Parliamentary Corpora

Miyazaki, 7 May 2018
EuroParl-UdS: Preserving and Extending Metadata in Parliamentary Debates

Alina Karakanta, Mihaela Vela, Elke Teich

Department of Language Science and Technology, Saarland University
Outline

● Introduction
● Motivation
● EuroParl-Uds
● Corpus processing
  ○ Metadata – proceedings
  ○ Metadata – MEPs
  ○ Sorting and alignment
● Corpus structure
● Corpus statistics
● Possible applications
Introduction

Parliamentary corpora as a rich high-quality resource for linguistic applications

**Metadata**

Structure, organise, filter

**Related projects:**

- Europarl (Koehn, 2005)
- Corrected and Structured EuroParl corpus (Gräen et al., 2014)
- European Comparable and Parallel Corpora (Calzada Perez et al., 2006)
- Digital Corpus of the European Parliament (Hajlaouiet al., 2014)
- Talk of Europe – Travelling CLARIN Campus/LinkedEP (van Aggelen et al., 2017)
Motivation

- Machine translation
- Gender identification (Koppel et al., 2002)
- Topic detection (Yang et al., 2011; Blei, 2012)
- Translation studies (translationese)
  - Status: Original vs. Translation (Rabinovich et al., 2015)
  - Producer: Native vs. Non-native (Nisioi et al., 2016)
EuroParl-UdS

- **Parallel corpora** where the source language (SL) sentences come from native SL speakers and are aligned to sentences in the required target language (TL) (SL\_native - TL) and
- **comparable monolingual corpora** of the target languages, where the sentences come from native TL speakers (TL\_native).
- A complete **pipeline** to compile such a corpus from European Parliament debates.

Supported languages (so far): EN, DE, ES
Corpus processing

1. Download proceedings in HTML
2. Download MEPs’ metadata in HTML
3. Extract MEPs’ information
4. Model proceedings as XML
5. Filter out text not in the expected language
6. Add MEPs’ metadata to proceedings
7. Add sentence boundaries
8. Annotate token, lemma, PoS
9. Separate originals from translations and filter by native speakers
10. Extract text into raw format
11. Sentence align
Corpus processing

1. Download proceedings in HTML

2. Download MEPs' metadata in HTML

3. Extract MEPs' information

4. Model proceedings as XML
Metadata - MEPs

- **meps**
  - id
  - name
  - nationality
  - birth date
  - birth place
  - death date
  - death place

- **national_parties**
  - Id
  - start date
  - end date
  - name of the party

- **political_parties**
  - Id
  - member state
  - start date
  - end date
  - name of the group
  - role in the group
Corpus processing

1. Download proceedings in HTML

2. Download MEPs’ metadata in HTML

3. Extract MEPs’ information

4. Model proceedings as XML

5. Filter out text not in the expected language

6. Add MEPs’ metadata to proceedings

7. Add sentence boundaries

8. Annotate token, lemma, PoS
Corpus processing

1. Download proceedings in HTML
2. Download MEPs’ metadata in HTML
3. Extract MEPs’ information
4. Model proceedings as XML
5. Filter out text not in the expected language
6. Add MEPs’ metadata to proceedings
7. Add sentence boundaries
8. Annotate token, lemma, PoS
9. Separate originals from translations and filter by native speakers
10. Extract text into raw format
11. Sentence align

Diagram:
- XML
- CSV
- xml_metadata
- EN_o
- EN_t
- EN_n
- DE
Sorting and alignment

- Language identifiers to filter text not in the expected language (Python: `langid`, `langdetect`)
- PoS-Tagging, lemmatisation (`TreeTagger`)
- Sentence split (`Punkt`, `NLTK`)
- Filter by:
  - Originals vs. Translations
  - Native vs. Non-native

For the parallel corpus:
- Automatically align text per intervention (`hunalign`)
## Corpus structure

<table>
<thead>
<tr>
<th>Directory</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>html</td>
<td>The crawled proceedings and MEPs’ information in HTML</td>
</tr>
<tr>
<td>metadata</td>
<td>MEPs’ metadata in CSV</td>
</tr>
<tr>
<td>txt</td>
<td>Raw text of the proceedings</td>
</tr>
<tr>
<td>xml</td>
<td>Proceedings transformed from HTML to XML</td>
</tr>
<tr>
<td>xml_langid</td>
<td>Proceedings in XML where the text not in the expected language is filtered out</td>
</tr>
<tr>
<td>xml_metadata</td>
<td>Proceedings in XML with added MEPs’ metadata</td>
</tr>
<tr>
<td>xml_sentences</td>
<td>Proceedings in XML where text is split into sentences</td>
</tr>
<tr>
<td>xml_translationese</td>
<td>Proceedings in XML filtered by factors relevant for translation</td>
</tr>
<tr>
<td></td>
<td>– original, translation, native speaker</td>
</tr>
<tr>
<td></td>
<td>For each language $a$, it contains</td>
</tr>
<tr>
<td></td>
<td>- the originals in $a$,</td>
</tr>
<tr>
<td></td>
<td>- the originals in $a$ only by native speakers,</td>
</tr>
<tr>
<td></td>
<td>- all translations from any language into $a$ and</td>
</tr>
<tr>
<td></td>
<td>- all translations into $a$ from a specific $SL$ where the speakers are native speakers of the $SL$</td>
</tr>
<tr>
<td>xml_ttg</td>
<td>PoS-tagged and lemmatized proceedings in XML</td>
</tr>
<tr>
<td>raw_parallel</td>
<td>For each language the corresponding parallel corpora</td>
</tr>
<tr>
<td>raw_comparable</td>
<td>For each language the comparable corpus of original texts by native speakers</td>
</tr>
</tbody>
</table>
## Corpus statistics

<table>
<thead>
<tr>
<th></th>
<th>EN→DE</th>
<th>EN→ES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>words</td>
<td>sents</td>
</tr>
<tr>
<td>all</td>
<td>42.08 M/38.93 M</td>
<td>1.91 M</td>
</tr>
<tr>
<td>translationese_orig</td>
<td>6.43 M/6.22 M</td>
<td>296.7 K</td>
</tr>
<tr>
<td>translationese_native</td>
<td>3.18 M/3.10 M</td>
<td>137 K</td>
</tr>
</tbody>
</table>

Table 3: Statistics of the parallel corpora after every processing step

<table>
<thead>
<tr>
<th></th>
<th>EN</th>
<th>DE</th>
<th>ES</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>words</td>
<td>sents</td>
<td>words</td>
</tr>
<tr>
<td>html</td>
<td>95.21 M</td>
<td>5.11 M</td>
<td>91.48 M</td>
</tr>
<tr>
<td>xml</td>
<td>95.60 M</td>
<td>5.11 M</td>
<td>92.43 M</td>
</tr>
<tr>
<td>langidfilter</td>
<td>65.55 M</td>
<td>3.23 M</td>
<td>40.23 M</td>
</tr>
<tr>
<td>translationese_orig</td>
<td>19.69 M</td>
<td>0.84 M</td>
<td>11.74 M</td>
</tr>
<tr>
<td>translationese_native</td>
<td>8.67 M</td>
<td>0.37 M</td>
<td>7.86 M</td>
</tr>
</tbody>
</table>

Table 2: Statistics of the comparable corpora after every processing step
Possible applications

- Careful data selection for improving MT (Kurokawa et al., 2009; Lembersky et al., 2012a)
- Translationese features

- Modelling human translation choice (Teich, E. and Martínez Martínez, J., in press)

\[
\text{Adequacy} = \text{Fidelity to SL} + \text{Conformity with TL}
\]
EuroParl-UdS

The corpus: http://fedora.clarin-d.uni-saarland.de/europarl-uds/

The code: https://github.com/hut-b7/europarl-uds

Thank you!

Questions: alina.karakanta@uni-saarland.de

Special thanks to our colleague Jose Martinez Martinez for by providing us his scripts (https://github.com/chozelinek/europarl) for crawling the data - while building this resource.