The Polish Parliamentary Corpus

Encoding format

Maciej Ogrodniczuk

Linguistic Engineering Group
Institute of Computer Science
Polish Academy of Sciences

ParlaFormat Worskhop

Amersfoort, May 23–24, 2019

The Polish Parliamentary Corpus

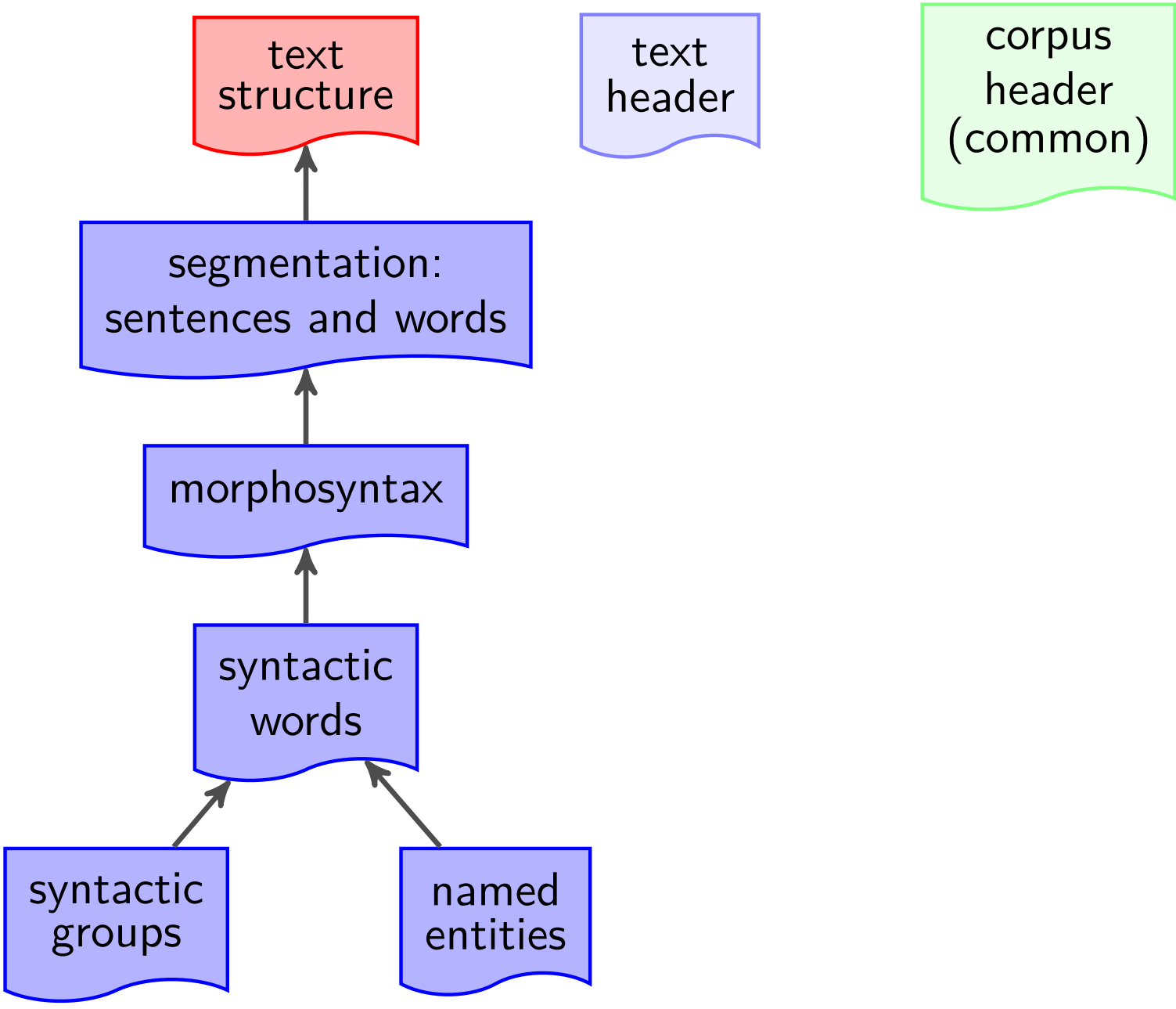
In a nutshell:

- a 600M-token collection of linguistically annotated documents from the proceedings of Polish Parliament (Sejm and Senate)
- based on the Polish Sejm Corpus prepared in October 2011 and recently extended in CLARIN-PL
- with new data updated live
- available at: `http://clip.ipipan.waw.pl/PPC`

Corpus data

Corpus format and structure:

- each session day in a separate folder
- header and linguistic annotation in stand-off XML TEI P5 files
- generated automatically with in-house tools:
  - morphological analyser Morfeusz SGJP (text structure, utterance-level segmentation, tokenization, lemmatization)
  - disambiguating tagger Concraft (disambiguated morphosyntactic description)
  - syntactic parser Spejd (syntactic words and syntactic groups)
  - named entity recognizer Liner2 (people, organizations, geographical names)
- compatible with NKJP (National Corpus of Polish), see `http://nlp.ipipan.waw.pl/TEI4NKJP/`


text structure
text header
corpus header (common)
segmentation: sentences and words
morphosyntax
syntactic words
syntactic groups
named entities

Text header

Basic metadata:

```
<sourceDesc>
  <bibl>
    <title>Sprawozdanie stenograficzne z obrad
      Senatu RP z 7 grudnia 2016 r. (kadencja IX,
      Posiedzenie Plenarne 31, dzień 6).</title>
    <publisher>Kancelaria Senatu RP</publisher>
    <note type="system">III RP</note>
    <note type="house">Senat</note>
    <note type="termNo">9</note>
    <note type="type">Posiedzenie Plenarne</note>
    <note type="sessionNo">31</note>
    <note type="dayNo">6</note>
    <date>2016-12-07</date>
  </bibl>
</sourceDesc>
```

Text header

List of all speakers:

```
<profileDesc>
  <particDesc>
    <person xml:id="SenatorBogdanBorusewicz"
            role="speaker">
      <persName>Senator Bogdan Borusewicz</persName>
    </person>
    ...
    <person xml:id="GlosZSali" role="speaker">
      <persName>Głos z sali</persName>
    </person>
    </person>
  </particDesc>
</profileDesc>
```

Linguistic annotation format

Text structure:

```
<teiCorpus>
  <xi:include href="corpus_header.xml"/>
  <TEI>
    <xi:include href="header.xml"/>
    <text>
      <body>
        <!-- ... -->
        <div xml:id="txt_7-div">
          <u xml:id="txt_7.1-u" who="#The_Speaker">
            <!-- ... -->
          </u>
          <u xml:id="txt_7.2-u" who="#MP_Jan_Nowak">
            <!-- ... -->
```

Linguistic annotation format

Sentence- and token-level segmentation:

```
<p xml:id="..."
   corresp="text_structure.xml#...">
  <s xml:id="...">
    <!-- Dziękuję  -->
    <seg xml:id="..."
         corresp="text_structure.xml
         #string-range(txt_7.1-u,0,8)"/>
    <!-- bardzo -->
    <seg xml:id="..."
         corresp="text_structure.xml
         #string-range(txt_7.1-u,9,6)"/>
    <!-- . -->
    <seg xml:id="..." nkjp:nps="true"
         corresp="text_structure.xml
         #string-range(txt_7.1-u,15,1)"/>
```

Linguistic annotation format

Morphosyntactic annotation:

```
<p xml:id="...">
  <s xml:id="..."
     corresp="text_structure.xml#...">
   <seg xml:id="..."
        corresp="ann_segmentation.xml#...">
    <fs type="morph">
      <f name="orth">
       <string>posiedzenia</string>
      </f>
      <f name="interps">
        ...
      </f>
      <f name="disamb">
        ...
      </f>
```

Linguistic annotation format

Morphosyntactic annotation:

```
<f name="interps">
  <fs type="lex" xml:id="...">
    <f name="base">
     <string>posiedzenie</string>
    </f>
    <f name="ctag">
     <symbol value="subst"/>
    </f>
    <f name="msd">
     <vAlt>
      <symbol value="pl:nom:n" xml:id="..."/>
      <symbol value="sg:gen:n" xml:id="..."/>
      <symbol value="pl:acc:n" xml:id="..."/>
      <symbol value="pl:voc:n" xml:id="..."/>
     </vAlt>
```

Linguistic annotation format

Morphosyntactic annotation:

```
<f name="disamb">
  <fs type="tool_report">
    <f fVal="#morph_1.1.3.1.2-msd" name="choice"/>
    <f name="interpretation">
      <string>posiedzenie:subst:sg:gen:n</string>
    </f>
  </fs>
</f>
```

Linguistic annotation format

Named entities:

```
<seg xml:id="...">
  <fs type="named">
    <f name="type">
     <symbol value="persName"/>
    </f>
    <f name="orth">
     <string>Stanisław Karczewski</string>
    </f>
  </fs>
  <ptr target="..."/>
  <ptr target="..."/>
</seg>
```

Questions

Shouldn't we have...

- some more sophisticated text structure?
- 'more computationally optimal' format?
  - a little lengthy for 600M tokens...
  - isn't it too sophisticated for our application?
  - regexps 'more popular' than parsing
  - human-readable comments in XML files?

Thank you!

And several funding institutions:

- a European (CIP ICT-PSP) project CESAR: Central and South-East European Resources (grant agreement 271022), part of META-NET
- part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education (DIR/WK/2016/02 and DIR/WK/2018/01)
- CLARIN ERIC for inviting us!