The Polish Sejm Corpus

Maciej Ogrodniczuk
Institute of Computer Science, Polish Academy of Sciences

INTRODUCTION

General information

The Polish Sejm Corpus is a new specialized resource containing transcribed, automatically annotated utterances of the Members of Polish Sejm (lower chamber of the Polish Parliament).

Source

- Transcripts of modern-time Sejm plenary sittings have been taken and published since 1918.
- Since 1993 their current versions are being made available online in the form of PDF transcripts and via a simple search interface.
- Due to their purely informative purpose they cannot be treated as a useful source of linguistic information.

Motivation for the task

- Sejm transcripts are a valuable source of large (300M segments), publicly available collection of quasi-spoken data.
- Previous attempts of their linguistic processing (IPI PAN corpus, NKJP) treated this resource selectively, both in terms of size and available structural information.
- The task prepares the ground for integration of audio/video recordings started by Sejm with the beginning of the 7th term of office (Q4 2011).

CORPUS ANNOTATION

NKJP-based format

The Polish Sejm Corpus format is taken over from the National Corpus of Polish format (NKJP, see `http://nkjp.pl`):

- TEI P5-based encoding format for documenting textual data,
- ISO FSR feature structures representation for the encoding of linguistic information,
- their concrete application to Polish linguistic data as put forward by NKJP.

Corpus structure

General information about the corpus is represented in a unique corpus header, gathering all common general metadata (place of the speech act, information of the type and formality of the utterance etc.)

Files related to each Sejm session are stored in separate folders containing a header file with content-related metadata (such as sitting number/day, list of speakers etc.) and several stand-off annotated files:

- `text_structure.xml` — text layer of the session, including basic structure of the session record, whenever available,
- `ann_segmentation.xml` — segmentation into sentences and tokens,
- `ann_morphosyntax.xml` — morphosyntactic description,
- `ann_words.xml` — syntactic words,
- `ann_groups.xml` — syntactic groups,
- `ann_named.xml` — named entities.

Example:

```
<seg xml:id="named_3.3-s_n2">
  <fs type="named">
    <f name="type">
      <symbol value="Person"/>
    </f>
    <f name="orth">
      <string>Lech Wałęsa</string>
    </f>
  </fs>
  <ptr target="ann_morphosyntax.xml
               morph_3.3.11-seg"/>
  <ptr target="ann_morphosyntax.xml
               morph_3.3.12-seg"/>
</seg>
```

PROCESSING TOOLS

1. Morfeusz SGJP — sentence- and token-level segmenter, morphological analyser and lemmatizer,
2. Pantera — morphosyntactic disambiguating tagger,
3. Spejd — shallow parser using a cascade grammar of Polish,
4. NERF — statistical named entity recognizer using CRF modelling.

BASIC STATISTICS

PSC vs. balanced NKJP

Basic statistics of the Polish Sejm Corpus as compared to the balanced subcorpus of the NKJP:

	Balanced NKJP	PSC
Segments	219 946 994	113 536 955
Unique analyses	2 341 623	718 267
Unique segments	1 795 722	427 598
Unique lemmata	1 188 737	189 321
Unique MSD tags	812	898

Working data set

Average frequency of the 10 most frequent POS tags in the Polish Sejm Corpus and the balanced subcorpus of the NKJP:

POS tag	Balanced NKJP	PSC
`subst`	26.58%	29.43%
`interp`	18.51%	15.06%
`prep`	9.43%	10.03%
`adj`	9.30%	11.68%
`qub`	4.67%	4.44%
`fin`	5.00%	5.86%
`praet`	4.41%	2.54%
`conj`	4.21%	3.76%
`adv`	3.55%	3.03%
`inf`	3.51%	1.76%

AUDIO/VIDEO SAMPLE

- Currently: one sitting day.
- Audio: 64kbps, video: 280 kbps, 320 x 240.
- `<encodingDesc>`-based association with textual data.

NEXT STEPS

- More content:
  - sessions of the 7th term,
  - parliamentary questions or Sejm committee meeting transcripts,
  - similar data textual (e.g. from the Polish Senate),
  - current audio/video recordings.
- More annotation:
  - word-sense disambiguation,
  - coreference resolution,
  - deeper annotation of existing layers (e.g. finer-grained categories of NEs),
  - SMIL descriptions for synchronizing content with audio/video.
- Better technical operation:
  - "live corpus",
  - information extraction based on the corpus data.

ACKNOWLEDGEMENTS

The work reported here was carried out within the Central and South-East European Resources (CESAR) project, part of META-NET, co-funded by the European Commission under the CIP Information and Communications Technologies (ICT) Policy Support Programme (Grant Agreement No 271022). See `http://www.meta-net.eu/projects/cesar` for details.

michal.lenart@gmail.com, maciej.ogrodniczuk@ipipan.waw.pl