The Polish Sejm Corpus

Maciej Ogrodniczuk

Institute of Computer Science, Polish Academy of Sciences
Jana Kazimierza 5, 01-248 Warsaw, Poland
maciej.ogrodniczuk@ipipan.waw.pl

Abstract

This document presents the first edition of the Polish Sejm Corpus – a new specialized resource containing transcribed, automatically annotated utterances of the Members of Polish Sejm (lower chamber of the Polish Parliament). The corpus data encoding is inherited from the National Corpus of Polish and enhanced with session metadata and structure. The multi-layered stand-off annotation contains sentence- and token-level segmentation, disambiguated morphosyntactic information, syntactic words and groups resulting from shallow parsing and named entities.

The paper also outlines several novel ideas for corpus preparation, e.g. the notion of a live corpus, constantly populated with new data or the concept of linking corpus data with external databases to enrich content.

Although initial statistical comparison of the resource with the balanced corpus of general Polish reveals substantial differences in language richness, the resource makes a valuable source of linguistic information as a large (300 M segments) collection of quasi-spoken data ready to be aligned with the audio/video recording of sessions, currently being made publicly available by Sejm.

Keywords: written corpora, quasi-spoken data, parliament transcripts, Polish

1. Introduction

Transcripts of Sejm[1] plenary sittings have been taken and published in paper form since the very beginnings of this authority in modern times, i.e. 1918. Since 1993 their current versions are being made available online in the form of publicly available transcripts[2] and the separate simple search interface[3]. Due to their purely informative purpose they only contain text and basic metadata and as such, they cannot be treated as a useful source of linguistic information; even though the full-text search is available, only AND/OR/NOT and asterisk operators are available and no inflectional properties are maintained.

The data from Sejm sittings are a valuable source of linguistic information in two aspects: firstly, as a large (300 M segments) collection of quasi-spoken data and secondly, in the context of making the audio/video recording of sessions available, started with the beginning of the 7th term of office (autumn 2011). This article aims at presenting the linguistically annotated text resource together with a sample of audio/video material corresponding to a selection of transcripts. The resulting Polish Sejm Corpus (PSC) is included into the standard resource set for Polish, made available in the META-SHARE infrastructure[4].

2. The Corpus Format and Structure

A selection of Sejm utterances has already been used by existing general-purpose corpora of Polish, i.e. IPI PAN Corpus (Przepiórkowski, 2004), see `http://korpus.pl` and NKJP (National Corpus of Polish, Pl: Narodowy Korpus Języka Polskiego, (Przepiórkowski et al., 2010), see `http://nkjp.pl`). The NKJP project, being the most advanced large-scale corpus project in Poland, creates an ideal background for the current smaller-scale specialized resource, offering well-tested tools and formats for newly created corpora.

2.1. The NKJP Format

The base format actually employed in NKJP and selected for the Sejm Corpus is TEI P5[5] — a *de facto* standard for encoding and documenting textual data, including the ISO FSR feature structures representation used for the encoding of linguistic information[6]. Its concrete application to Polish linguistic data is thoroughly described in (Przepiórkowski and Bański, 2009b)[7].

The NKJP format assumes stand-off linguistic annotation distributed over various layers: source text, segmentation (paragraph-, sentence- and token-level), morphosyntax, syntactic words and groups, named entities and word sense disambiguation. The Sejm Corpus adopts this approach together with the folder structure and file naming

[1] See `http://www.sejm.gov.pl/english.html` for more information on the institution structure and activity.

[2] For their record concerning the current, 7th term of office, see `http://www.sejm.gov.pl/Sejm7.nsf/stenogramy.xsp` (currently in Polish only).

[3] See `http://orka2.sejm.gov.pl/Debata7.nsf` for the current term (in Polish).

[4] META-SHARE is an interoperable infrastructure for the language technology domain, initiated and maintained by META-NET – a network of excellence dedicated to promote the technological foundations of a multilingual European information society. See `http://www.meta-net.eu/` and `http://www.meta-net.eu/meta-share` for details.

[5] See (Przepiórkowski and Bański, 2009a) and (Przepiórkowski, 2009) for detailed discussion on competitive multi-level XML linguistic annotation formats, including TIGER-XML, MAF/SynAF/LAF/GrAF, PAULA and XCES, and reasons for final selection of TEI P5.

[6] Use of TEI and feature structures for encoding linguistic information has a long history in Poland; see e.g. (Ogrodniczuk, 2000).

[7] See also `http://nlp.ipipan.waw.pl/TEI4NKJP/` for samples of NKJP files.

convention of NKJP in the final representation of the data, extending it with audio/video capabilities.

2.2. The Corpus Structure

General information about the corpus is represented in a unique corpus header file. Due to considerable uniformity of the source data, the corpus header also gathers the common general metadata such as place of the speech act, information of the type and formality of the utterance etc.

Files related to each Sejm session are stored in a separate folder, named with term and session number. Each folder contains a text header file with content-related metadata (such as sitting number/day, list of speakers etc.) and several annotation files (compressed for practical reasons):

- `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.

Annotations were created with Morfeusz SGJP (Woliński, 2006) morphological analyser, also responsible for sentence-level segmentation and tokenisation, Pantera (Acedański, 2010) tagger which produces disambiguation information and scripts used by the NKJP project, Spejd (Buczyński and Przepiórkowski, 2009) shallow parser with a cascade grammar of Polish and NERF (Savary et al., 2010) statistical tool based on the Conditional Random Fields modelling method.

2.2.1. Text Structure

The text layer consists of `<div>`s representing continuous statement of a single speaker. Individual utterances are enclosed in `<u>` elements. Stenographer's comments, referring to non-spoken events in the sitting hall are separately marked; so are the undelivered speeches, also formally encompassed by the official version of transcripts and therefore included in the Sejm Corpus for completeness. Each utterance is explicitly marked (by `who` attribute) with reference to the speaker, referencing the header file:

```
<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_Kowalski">
            <!-- ... -->
```

The text layer is the only one containing text content of the record; all other layers maintain reference to preceding descriptions using TEI `@corresp` attribute values.

2.2.2. Sentence- and Token Level Segmentation

Segmentation layer identifies sentences and individual tokens by enclosing empty `<seg>` (token) elements in `<s>` (sentence) elements. Tokens do not store their text representation (available in the text layer and referenced with string ranges), but XML comments are used to make the content human-readable:

```
<p corresp="text_structure.xml
            #txt_7.1-u"
   xml:id="segm_txt_7.1-u">
  <s xml:id="segm_txt_7.1-u.1-s">
    <!-- ... -->
    <!-- Proszę -->
    <seg corresp="text_structure.xml
         #string-range(txt_7.1-u,0,6)"
         xml:id="segm_txt_7.1-u.4-seg"/>
    <!-- państwa -->
    <seg corresp="text_structure.xml
         #string-range(txt_7.1-u,7,7)"
         xml:id="segm_txt_7.1-u.5-seg"/>
    <!-- ... -->
```

2.2.3. Morphosyntactic Annotation

The morphosyntactic layer is encoded by a list of segments containing feature structure specifications of morphosyntactic information on the segment. Disambiguated interpretations are separately marked:

```
<p xml:id="txt_7.1-u">
 <s corresp="ann_segmentation.xml
             #segm_txt_7.1-u.1-s"
    xml:id="txt_7.1-u.1-s">
   <!-- ... -->
   <seg corresp="ann_segmentation.xml
        #segm_txt_7.1-u.415-seg"
        xml:id="morph_txt_7.1-u.45-seg">
    <fs type="morph">
     <f name="orth">
      <string>państwa</string>
     </f>
     <!-- państwa [7,7] -->
     <f name="interps">
       <fs type="lex"
           xml:id="morph_txt_7.1-u
                   .45-seg_0-lex">
        <f name="base">
         <string>państwo</string>
        </f>
        <f name="ctag">
         <symbol value="subst"/>
        </f>
     <f name="msd">
      <vAlt>
       <symbol value="pl:nom:n"
               xml:id="morph_txt_7.1
                   -u.45-seg_0-msd"/>
       <symbol value="pl:gen:m1"
               xml:id="morph_txt_7.1
                   -u.45-seg_1-msd"/>
       <symbol value="sg:gen:n"
               xml:id="morph_txt_7.1
                   -u.45-seg_2-msd"/>
       <symbol value="pl:acc:m1"
               xml:id="morph_txt_7.1
                   -u.45-seg_3-msd"/>
       <symbol value="pl:acc:n"
               xml:id="morph_txt_7.1
                   -u.45-seg_4-msd"/>
       <symbol value="pl:voc:n"
               xml:id="morph_txt_7.1
                   -u.45-seg_5-msd"/>
      </vAlt>
     </f>
    </fs>
  </f>
  <f name="disamb">
   <fs feats="#pantera"
       type="tool_report">
     <f fVal="#morph_txt_7.1
              -u.45-seg_3-msd"
        name="choice"/>
     <f name="interpretation">
      <string>państwo:subst
              :pl:acc:m1</string>
     <!-- ... -->
```

Morphosyntactic descriptions conform to the NKJP positional tagset (Przepiórkowski and Woliński, 2003), containing 36 grammatical classes (roughly, parts of speech, e.g., adjective). For each grammatical class there is a list of obligatory and optional grammatical categories (e.g., case and number), with the total of 13 different categories in Polish. Each grammatical category has an associated list of possible values (e.g., singular and plural for the grammatical number). In this definition of the NKJP Tagset, all grammatical classes are complex/open Data Categories (DCs), grammatical categories are complex/closed DCs, and the values of grammatical categories are simple DCs.

2.2.4. Syntactic Words and Groups

Syntactic word is a non-empty sequence of tokens and/or constituent syntactic words (which amounts to a non-empty sequence of tokens) which can be referenced from the shallow parsing layer.

Syntactic word layer is intended to help with syntactic parsing; since the process of morphological analysis (and, in turn, tagging) could result in identification of segments shorter than space-to-space words, construction of parsing rules would be much more complex than with "traditional" understanding of segments. In practice, this method also helps with "genuine" compounds (such as "*marszałek-senior*": en. *Senior Marshal*[8]):

```
<seg xml:id="words_3.3-s_42">
<!--  rule="subst-subst" -->
  <fs type="word">
    <f name="orth">
      <string>marszałka
              -seniora</string>
    </f>
    <f name="interps">
      <fs type="lex">
        <f name="base">
          <string>marszałek
                  -senior</string>
        </f>
        <f name="ctag">
          <symbol value="Noun"/>
        </f>
        <f name="msd">
          <symbol value="sg:nom:m1"/>
        </f>
      </fs>
    </f>
  </fs>
  <ptr target="ann_morphosyntax.xml
               #morph_3.3.14-seg"/>
  <ptr target="ann_morphosyntax.xml
               #morph_3.3.15-seg"/>
  <ptr target="ann_morphosyntax.xml
               #morph_3.3.16-seg"/>
</seg>
```

Syntactic groups represented in the annotation model contain pointers (<ptr>s) to immediate constituents of the group — syntactic words specified in the preceding layer or other syntactic groups of the same layer, irrespective of their continuity. <ptr> elements may specify the type of the constituency relation (head or nonhead).

```
<seg xml:id="groups_2.2-s_5">
<!--  rule="NGs: Noun
            + n-Noun (nom)" -->
  <fs type="group">
    <f name="orth">
      <string>marszałek-senior
              Małachowski</string>
    </f>
    <f name="type">
      <symbol value="NGs"/>
    </f>
  </fs>
  <ptr type="head"
       target="ann_words.xml
               #words_2.2-s_14"/>
```

[8] See e.g. http://en.wikipedia.org/wiki/Senior_Marshal.

```
  <ptr type="nonhead"
       target="ann_words.xml
               #words_2.2-s_15"/>
</seg>
```

2.2.5. Named Entities

Named entities are assigned basic types (date and time expressions, names of organizations, locations and persons); they reference the morphosyntactic layer directly:

```
<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>
```

3. The Corpus Data

The source data have been received from Sejm in two batches in 2011 and consisted of approx. 230 000 files from all six modern-time terms between 1991 and 2011 (557 sessions in total). Although the character of the data is public, obtaining it directly from Sejm enabled easier access to more detailed descriptions of the sessions (as compared to the interfaces presented to the general public) with references to the discussion agenda, clear identification of voting procedures etc. The data of the last two terms were available in the form of XML files; earlier terms only in HTML format.

The first edition of the Sejm Corpus was prepared in October 2011 and contains automatically annotated transcripts of Sejm sessions saved in TEI P5 format as described in the previous section. The corpus data is freely available for download in the form of compressed files corresponding to all six modern-time terms of office at `http://clip.ipipan.waw.pl/PSC`.

3.1. Basic Statistics

One of the first experiments carried out with the Sejm data was confronting them with the balanced 220-million subcorpus of NKJP, which allowed basic frequency comparisons. Although the subject needs further studies, certain initial observations can be made even at this stage. One of the most interesting findings is that even such simple indicators as the number of unique segments or lemmata (see Tab. 1) show that the richness of general language represented by the balanced corpus is incomparable to that of the parliamentary language (note e.g. over 6:1 unique lemma ratio as compared to the 2:1 total segment ratio).

	Balanced NKJP	**PSC**
Segments	219 946 994	113 536 955
Unique analyses[9]	2 341 623	718 267
Unique segments	1 795 722	427 598
Unique lemmata	1 188 737	189 321
Unique MSD tags	812	898

Table 1: Basic statistics of the Polish Sejm Corpus as compared to 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%

Table 2: Average frequency of the 10 most frequent POS tags in the Polish Sejm Corpus and the balanced subcorpus of the NKJP

This result cannot be explained with any substantial upset of the word category balance since Tab. 2 clearly shows that the POS ratio is nearly the same in both samples.

3.2. The Audio/Video Sample

However audio/video recording of Sejm sittings in digital form started in 1993, it is only recently when the first attempts of making it available for the general public began. The older parts of data, recorded in Betacam format, are still to be converted into newer formats and aligned with texts, but the most current footage is already prepared for integration. One of the biggest advantages of this move is the possibility of using the audio/text for training statistical speech-to-text engines, the present quality of which is still not satisfactory for Polish, mostly due to lack of training resources.

Recently audio/video samples of one sitting day has been made available to corpus creators. Audio material is stored in 64 kb/s mp3 files while corresponding video material in 280 kb/s mov files (320 x 240). The location of audio/video data in the current version of the corpus is referenced from the text header in `<encodingDesc>`. At later stages, SMIL (Synchronized Multimedia Integration Language) descriptions that assign identifiers to various temporal spans of sound are to be considered.

4. Current and Future Work

The Polish Sejm Corpus was made available in 2011 in the first batch of resources populating META-SHARE, the open digital exchange facility provided by META-NET. Currently a number of extensions to the present version are being prepared (among them, a general search interface

[9] Unique (segments, lemma, tag) triples as compared to unique orthographic forms of segments, counted in the next row.

based on Poliqarp[9] (Janus and Przepiórkowski, 2006) – a universal concordancer for large corpora which would facilitate access to data in a standard query-based manner.

The most obvious direction is addition of data as they appear with the next terms of Sejm. It can be either achieved by publishing subsequent editions of the corpus, preferably at the term turn (every 4 years) or by integrating data harvesting procedures into the processing chain to retrieve sitting transcripts on the fly. This method would allow for creation of a „live corpus", constantly updated with new data. Additional content (parliamentary questions or Sejm committee meeting transcripts) which partially has already been handed over to the corpus creators is already scheduled to be included in the next batches of resources (July 2012 and January 2013); negotiations with the Polish Senate are also being conducted.

Moreover, related information interesting for the general public is planned to be extracted from the data using linguistic technologies. Current plans include development of queries to retrieve "phrase of the sitting" (the most frequent nominal phrase that describes current interests of the MPs), characteristics of language use by individual MPs or political parties they represent or collating utterances with detailed information about the speakers (their age, region, education etc.), available in external Sejm metabases.

5. 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.

The author of the text would like to express his gratitude towards the authorities of Polish Sejm for the long-lasting co-operation, with special acknowledgements to Minister Lech Czapla, Chief Chancellor of Sejm, Mr. Zbigniew Jabłoński, the Director of Computer Science Department, Mr. Władysław Baksza from the Legislative Office, Mrs. Agata Sawicka-Wójcik and other employees of the Computer Science Department.

6. References

Szymon Acedański. 2010. A Morphosyntactic Brill Tagger for Inflectional Languages. In Hrafn Loftsson, Eiríkur Rögnvaldsson, and Sigrún Helgadóttir, editors, *Advances in Natural Language Processing*, volume 6233 of *Lecture Notes in Computer Science*, pages 3–14. Springer.

Aleksander Buczyński and Adam Przepiórkowski. 2009. Spejd: A Shallow Processing and Morphological Disambiguation Tool. *Human Language Technology: Challenges of the Information Society. Vol. 5603*, pages 131–141.

Daniel Janus and Adam Przepiórkowski. 2006. Poliqarp 1.0: Some technical aspects of a linguistic search engine for large corpora. In Jacek Waliński, Krzysztof Kredens, and Stanisław Goźdź-Roszkowski, editors, *The proceedings of Practical Applications of Linguistic Corpora 2005*, Frankfurt am Main. Peter Lang.

Maciej Ogrodniczuk. 2000. Wykorzystanie SGML i TEI do zapisu polskich danych lingwistycznych (Using SGML and TEI for representation of Polish linguistic information, in Polish). Master's thesis, Institute of Computer Science, Faculty of Mathematics, Informatics and Mechanics, University of Warsaw.

Adam Przepiórkowski and Piotr Bański. 2009a. Which XML standards for multilevel corpus annotation? In *Proceedings of the 4th Language & Technology Conference*, pages 245–250, Poznań, Poland.

Adam Przepiórkowski and Piotr Bański. 2009b. XML text interchange format in the National Corpus of Polish. In Stanisław Goźdź-Roszkowski, editor, *The proceedings of Practical Applications in Language and Computers PALC 2009*, Frankfurt am Main. Peter Lang.

Adam Przepiórkowski and Marcin Woliński. 2003. A Flexemic Tagset for Polish. In *Proceedings of* Morphological Processing of Slavic Languages*, EACL 2003*.

Adam Przepiórkowski, Rafał L. Górski, Marek Łaziński, and Piotr Pęzik. 2010. Recent developments in the National Corpus of Polish. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation, LREC 2010*, Valletta, Malta. ELRA.

Adam Przepiórkowski. 2004. *The IPI PAN Corpus: Preliminary version*. Institute of Computer Science, Polish Academy of Sciences, Warsaw.

Adam Przepiórkowski. 2009. TEI P5 as an XML standard for treebank encoding. In Marco Passarotti, Adam Przepiórkowski, Savina Raynaud, and Frank Van Eynde, editors, *Proceedings of the Eighth International Workshop on Treebanks and Linguistic Theories (TLT 8)*, pages 149–160, Milan, Italy.

Agata Savary, Jakub Waszczuk, and Adam Przepiórkowski. 2010. Towards the Annotation of Named Entities in the National Corpus of Polish. In *Proceedings of the 7th International Conference on Language Resources and Evaluation, LREC 2010*, Valletta, Malta. ELRA.

Marcin Woliński. 2006. Morfeusz — a practical tool for the morphological analysis of Polish. In Mieczysław A. Kłopotek, Sławomir T. Wierzchoń, and Krzysztof Trojanowski, editors, *Proceedings of the International Intelligent Information Systems: Intelligent Information Processing and Web Mining'06 Conference*, pages 511–520, Wisła, Poland, June.

[9] See `http://poliqarp.sourceforge.net` and (Przepiórkowski, 2004) for detailed information on Poliqarp syntax.