*Proceedings of ParlaCLARIN II Workshop*, pages 1–4
Language Resources and Evaluation Conference (LREC 2020), Marseille, 11–16 May 2020


# New Developments in the Polish Parliamentary Corpus

**Maciej Ogrodniczuk, Bartłomiej Nitoń**

Institute of Computer Science, Polish Academy of Sciences
Warsaw, Poland
maciej.ogrodniczuk@ipipan.waw.pl, bartek.niton@gmail.com

**Abstract**

This short paper presents the current (as of February 2020) state of preparation of the Polish Parliamentary Corpus (PPC) — an extensive collection of transcripts of Polish parliamentary proceedings dating from 1919 to present. The most evident developments as compared to the 2018 version is harmonization of metadata, standardization of document identifiers, uploading contents of all documents and metadata to the database (to enable easier modification, maintenance and future development of the corpus), linking utterances to the political ontology, linking corpus texts to source data and processing historical documents.

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

## 1. Introduction

The Polish Parliamentary Corpus[1] (Ogrodniczuk, 2018) is a collection of proceedings of Polish parliament dating from 1919 to present. It includes transcripts of Sejm sittings (including Legislative Sejm and State National Council), Sejm committee sittings from 1993, Sejm interpellations and questions from 1997, Senate sittings from 1922–1939 and 1989 to present[2] and Senate committee sittings from 2015. The collection is consequently updated with the most current data acquired from the Sejm and the Senate web portals. Currently the size of the textual data in the corpus amounts to over 340 thousand documents and almost 750 million tokens.

The data features annotation following the National Corpus of Polish (Przepiórkowski et al., 2012, NKJP)[3] (Przepiórkowski et al., 2012) TEI P5 XML format and conventions. Paragraph-, sentence- and token-level segmentation, lemmatization and morphosyntactic description was automatically produced with Morfeusz2 (Kieraś and Woliński, 2017) and disambiguated with Concraft2 (Waszczuk et al., 2018). The named entity layer was produced with Liner2 (Marcińczuk et al., 2013) and the dependency annotation layer with COMBO (Rybak and Wróblewska, 2018).

## 2. Corpus improvements

Apart from the main improvement consisting in adding new data (see Table 1 for detailed statistics) several improvements have been made in the corpus.

**Harmonization of metadata** The basic list of metadata for all document types (plenary sittings, committee sittings and questions) was set to comprise document title, publisher (Sejm or Senate), political system (Second Polish Republic — 1918–1939, Polish People's Republic — 1945–1989, the transition period with the Contract Sejm — 1989–1991 and the current Third Republic — from 1991 to present day), chamber (Sejm, Senate or the National Council), term of office, document type and the major date of the source.

Assignment of historical documents to the term of office was also adjusted, the information on the regime and chamber has been added, document names have been standardized and several naming errors corrected. Missing information on speakers has been filled in and the corpus header has been updated.

**Standardization of document identifiers** The corpus has been divided into 27 periods corresponding to the terms of office of chambers in three different political systems of Poland in the last 100 years (see rows of Table 1).

All identifiers of documents have been standardized reflecting the logical structure of the system:

```
191922- sjm -ppxxx- 00002 - 01
   ↓     ↓     ↓       ↓     ↓
period chamber type sitting/number day/part
```

**Database development** The contents of all documents and metadata have been uploaded to a specifically developed database to enable easier modification, maintenance and future development of the corpus.

The current size of the corpus amounts to 749M segments with detailed distribution over houses, periods, and document types presented in Table 1. Apart from the stenographic records of plenary sittings (261M segments) and committee sittings (288M segments), the corpus contains 199M segments of interpellations and questions.

**Linking utterances to the political ontology** The Polish Political Ontology[4] (PPO) is an RDF resource created in 2015 and modelling the Polish political scene of the period 1989–2014. It includes significant actors based in Polish political and other public institutions, including members

[1] Pol. Korpus Dyskursu Parlamentarnego, see `clip.ipipan.waw.pl/PPC`.

[2] The gap results from the fact that the Senate was abolished by the authorities of the Polish People's Republic and re-established after the reinstatement of democracy after the collapse of the communist government.

[3] Pol. Narodowy Korpus Języka Polskiego, see `http://nkjp.pl`.

[4] `http://zil.ipipan.waw.pl/PolishPoliticalOntology`

| | | | Sittings | | Committees | | Interpellations | |
|---|---|---|---|---|---|---|---|---|
| **System** | **Years** | **Period** | **docs** | **segments** | **docs** | **segments** | **docs** | **segments** |
| | 1919–1922 | Legislative Sejm | 312 | 6 945 162 | – | – | – | – |
| | 1922–1927 | 1st term of office | 277 | 7 338 355 | – | – | – | – |
| Second | 1928–1930 | 2nd | 58 | 2 139 835 | – | – | – | – |
| Polish | 1930–1935 | 3rd | 72 | 2 404 267 | – | – | – | – |
| Republic | 1935–1938 | 4th | 73 | 2 133 181 | – | – | – | – |
| | 1938–1939 | 5th | 23 | 610 455 | – | – | – | – |
| | 1943–1947 | State National Council | 6 | 234 441 | – | – | – | – |
| | 1947–1952 | Legislative Sejm | 107 | 2 575 136 | – | – | – | – |
| | 1952–1956 | 1st term of office | 39 | 1 172 333 | – | – | – | – |
| | 1957–1961 | 2nd | 59 | 2 502 936 | – | – | – | – |
| | 1961–1965 | 3rd | 32 | 1 388 862 | – | – | – | – |
| People's | 1965–1969 | 4th | 23 | 1 163 336 | – | – | – | – |
| Poland | 1969–1972 | 5th | 17 | 526 277 | – | – | – | – |
| | 1972–1976 | 6th | 32 | 1 176 712 | – | – | – | – |
| | 1976–1980 | 7th | 29 | 918 993 | – | – | – | – |
| | 1980–1985 | 8th | 70 | 3 377 139 | – | – | – | – |
| | 1985–1989 | 9th | 45 | 2 641 788 | – | – | – | – |
| | 1989–1991 | 10th | 77 | 6 674 111 | – | – | – | – |
| | 1991–1993 | 1st term of office | 142 | 7 739 147 | – | – | – | – |
| | 1993–1997 | 2nd | 317 | 22 134 682 | 3 858 | 41 756 476 | – | – |
| Third | 1997–2001 | 3rd | 320 | 24 138 142 | 4 691 | 42 510 604 | 23 507 | 12 101 453 |
| Polish | 2001–2005 | 4th | 337 | 28 743 846 | 4 945 | 49 302 521 | 30 986 | 17 519 177 |
| Republic | 2005–2007 | 5th | 148 | 11 737 186 | 2 359 | 18 970 036 | 26 689 | 14 777 377 |
| | 2007–2011 | 6th | 298 | 22 415 708 | 5 565 | 44 363 063 | 59 353 | 36 412 001 |
| | 2011–2015 | 7th | 292 | 20 765 505 | 5 126 | 38 541 083 | 85 679 | 61 565 989 |
| | 2015–2019 | 8th | 239 | 19 131 000 | 4 561 | 36 708 873 | 79 194 | 56 720 590 |

| | | | Sittings | | Committees | |
|---|---|---|---|---|---|---|
| **System** | **Years** | **Period** | **documents** | **segments** | **documents** | **segments** |
| | 1922–1927 | 1st term | 96 | 1 979 541 | – | – |
| | 1928–1930 | 2nd | 3 | 171 345 | – | – |
| Second Polish Republic | 1930–1935 | 3rd | 64 | 1 804 635 | – | – |
| | 1935–1938 | 4th | 29 | 724 687 | – | – |
| | 1938–1939 | 5th | 20 | 347 430 | – | – |
| | 1989–1991 | 1st term | 60 | 3 170 293 | – | – |
| | 1991–1993 | 2nd | 48 | 1 459 440 | – | – |
| | 1993–1997 | 3rd | 125 | 5 051 677 | – | – |
| Third | 1997–2001 | 4th | 187 | 8 255 897 | – | – |
| Polish | 2001–2005 | 5th | 175 | 6 485 347 | – | – |
| Republic | 2005–2007 | 6th | 74 | 3 571 293 | – | – |
| | 2007–2011 | 7th | 167 | 8 819 116 | – | – |
| | 2011–2015 | 8th | 159 | 7 100 841 | – | – |
| | 2015–2019 | 9th | 204 | 9 554 544 | 2 156 | 15 645 801 |
| | 2019– | 10th | 9 | 412 279 | 82 | 505 991 |

Table 1: Statistics of the Polish Parliamentary Corpus (2020)

of government and the parliament. Specifically, it contains information about the MPs (their gender, functions, terms of office, political affiliation) and political parties.

The corpus data, previously marked with speaker names only, was linked to the PPO by extending the `particDesc` section in TEI header files (`header.xml`) of individual documents of the corpus. Links were represented as pointers (`ptr` elements) to functions in PPO (see Fig. 2.).

**Linking corpus texts to source data** Corpus data have been updated with links to the original materials which were used as source of text, i.e.:

- websites from which the text of individual documents has been extracted
- websites from which the metadata for the document concerned has been extracted
- records of meetings in PDF format.

In order to prevent a possible loss of access to the source files (e.g. due to changes in parliamentary services) the source files were additionally downloaded to store their copies locally.

The process has been completed with a number of Internet robots browsing respective websites, separately for docu-

```
<teiHeader ...>
  ...
  <profileDesc>
    <particDesc>
      ...
      <person xml:id="PrezesRadyMinistrowDonaldTusk" role="speaker">
        <persName>The Prime Minister Donald Tusk</persName>
        <linkGrp type="function">
          <ptr target="http://legis.nlp.ipipan.waw.pl/onto/ppo.owl
                       #Donald_Franciszek_Tusk__Sejm6"/>
          <ptr target="http://legis.nlp.ipipan.waw.pl/onto/ppo.owl
                       #Donald_Tusk_2051"/>
          <ptr target="http://legis.nlp.ipipan.waw.pl/onto/ppo.owl
                       #Donald_Tusk_280"/>
       </linkGrp>
      </person>
    </particDesc>
  </profileDesc>
</teiHeader>
```

Figure 1: Representation of pointers to Polish Political Ontology in TEI header

ments between 1919 and 1997 as well as terms of office 2–6, 7–8 and 9 (due to changes in IT systems used in these periods). The processing consisted of keeping the URL address of the document source, the URL address of a file containing the content of the document, usually in PDF format and the address of the page containing document metadata.

**Processing historical documents** Due to changes in Polish ortography in 1936 modern tools are not always very successful with processing older data. To overcome this problem, a transcriber for historical documents and a customized version of morphological analyzer have been included in the process of linguistic analysis of 1027 documents between 1919 and 1939.

The processing pipeline consists of:

1. a rule transcriber[5] with a set of rules for nineteenth-century language[6] (Kieraś et al., 2017) (the original text is preserved in the database)
2. Morfeusz2 morphological analyzer using SGJP dictionary extended with vocabulary of the 19th century (Kieraś and Woliński, 2018) but with a set of tags consistent with contemporary vocabulary
3. Concraft2 tagger (no additional modifications)
4. Liner2 (no additional modifications);
5. COMBO (no additional modifications).

## 3. MTAS-based search engine

The previous searchable version of the corpus was made available as Poliqarp (Janus and Przepiórkowski, 2006) search engine binary (to be run on user's computer) and a Poliqarp-powered simple online search engine was available to facilitate search in a familiar NKJP-like interface. Still, one of the major faults of Poliqarp was inability to combine search over different annotation layers.

To overcome this flaw, a new framework for building search engines was created based on MTAS (Brouwer et al., 2017), a stable and reliable solution for multi-layered linguistic search, currently also used for other corpora of Polish[7]. MTAS offers rich search functions, using regular expressions, filtering results using metadata or merging of analytical layers.

Figure 3. presents a sample search result linking the morphological analysis layer with named entity layer: proper names identical with common names can be easily filtered.

## 4. Current and future work

The processing of the corpus is ongoing on many levels, starting with adding new historical data (transcripts of committee meetings before 1989).

Several 'administrative' tasks are also envisaged, starting from processing of corpus data with new versions of linguistic tools made available in the recent months. They are e.g. newest version of Morfeusz2, Concraft2 or COMBO parser, providing dependency trees.

Even though the manual correction of OCR-ed data has been successful, there are still numerous typos in this data, mostly due to poor quality of originals before 1989. To overcome this problem, new methods for automated discovery of errors in the texts will be developed, such as investigation of words unrecognized by the morphological analyser or detection of non-standard character ngrams. Related to this task is implementation of mechanisms that trigger linguistic analysis and re-indexation of corrected data after changes have been approved by an authorized user.

[5] https://bitbucket.org/jsbien/pol

[6] http://chronofleks.nlp.ipipan.waw.pl/static/files/reguly_xixw.zip

[7] See e.g. the 1 million subcorpus of NKJP (http://nkjp.nlp.ipipan.waw.pl/), the Electronic Corpus of 17th and 18th century Polish Texts (http://korba.edu.pl/) Corpus of 19th Century Polish (http://korpus19.nlp.ipipan.waw.pl/) or the Polish Coreference Corpus http://pcc.nlp.ipipan.waw.pl/

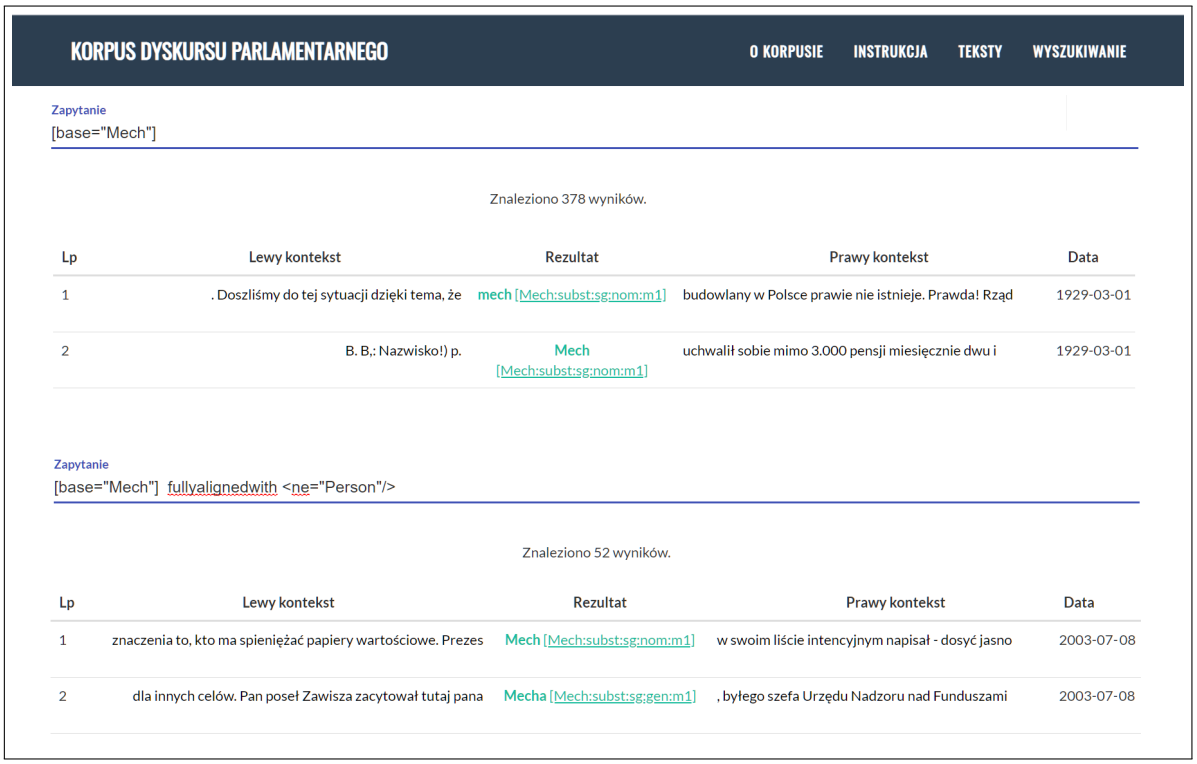


Figure 2: Sample search result in the corpus

## Acknowledgements

The work reported here was financed as part of the investment in the CLARIN-PL research infrastructure funded by the Polish Ministry of Science and Higher Education.

## Bibliographical references

Brouwer, M., Brugman, H., and Kemps-Snijders, M. (2017). MTAS: A Solr/Lucene based Multi Tier Annotation Search solution. In *Selected papers from the CLARIN Annual Conference 2016. Linköping Electronic Conference Proceedings 136*, pages 19–37. Linköping University Electronic Press.

Janus, D. and Przepiórkowski, A. (2006). Poliqarp 1.0: Some technical aspects of a linguistic search engine for large corpora. In Jacek Waliński, et al., editors, *Proceedings of Practical Applications of Linguistic Corpora 2005 conference*, Frankfurt am Main. Peter Lang.

Kieraś, W. and Woliński, M. (2017). Morfeusz 2 – analizator i generator fleksyjny dla języka polskiego. *Język Polski*, XCVII(1):75–83.

Kieraś, W. and Woliński, M. (2018). Manually annotated corpus of Polish texts published between 1830 and 1918. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 3854–3859, Paris, France. European Language Resources Association (ELRA).

Kieraś, W., Komosińska, D., Modrzejewski, E., and Woliński, M. (2017). Morphosyntactic annotation of historical texts. The making of the baroque corpus of Polish. In Kamil Ekštein et al., editors, *Text, Speech, and Dialogue 20th International Conference, TSD 2017, Prague, Czech Republic, August 27-31, 2017, Proceedings*, number 10415 in Lecture Notes in Computer Science, pages 308–316. Springer International Publishing.

Marcińczuk, M., Kocoń, J., and Janicki, M. (2013). Liner2 — A Customizable Framework for Proper Names Recognition for Polish. In Robert Bembenik, et al., editors, *Intelligent Tools for Building a Scientific Information Platform*, volume 467 of *Studies in Computational Intelligence*, pages 231–253. Springer-Verlag, Cham, Heidelberg, New York, Dordrecht, London.

Ogrodniczuk, M. (2018). Polish Parliamentary Corpus. In Darja Fišer, et al., editors, *Proceedings of the LREC 2018 Workshop* ParlaCLARIN: Creating and Using Parliamentary Corpora, pages 15–19, Paris, France. European Language Resources Association (ELRA).

Przepiórkowski, A., Bańko, M., Górski, R. L., and Lewandowska-Tomaszczyk, B. (2012). *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warszawa.

Rybak, P. and Wróblewska, A. (2018). Semi-supervised Neural System for Tagging, Parsing and Lematization. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 45–54, Brussels, Belgium, October. Association for Computational Linguistics.

Waszczuk, J., Kieraś, W., and Woliński, M. (2018). Morphosyntactic disambiguation and segmentation for historical Polish with graph-based conditional random fields. In Petr Sojka, et al., editors, *Text, Speech, and Dialogue: 21st International Conference, TSD 2018, Brno, Czech Republic, September 11-14, 2018, Proceedings*, number 11107 in Lecture Notes in Artificial Intelligence, pages 188–196. Springer-Verlag.