Polish Parliamentary Corpus

Maciej Ogrodniczuk
Institute of Computer Science, Polish Academy of Sciences
Warsaw, Poland
maciej.ogrodniczuk@ipipan.waw.pl

Abstract

This paper presents the Polish Parliamentary Corpus (PPC) — a new resource built upon the Polish Sejm Corpus and extended with current Senate proceedings and older (1918–1990) parliamentary transcripts. Corpus texts are automatically annotated with state-of-the-art language tools for Polish, resulting in a multi-layered stand-off sentence- and token-level segmentation, disambiguated morphosyntactic information, syntactic words and groups, named entities and coreference. The corpus is being constantly updated with new data from the current sittings. Currently the PPC is among the largest parliamentary corpora worldwide, amounting to approx. 300M words.

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

1. Introduction

The idea of creating a separate text corpus based on Polish parliamentary data (Sejm and Senate, the lower and upper houses of the Polish parliament) appeared as early as 2010 when a paper outlining the resource was registered to SinFonIJA 3 conference[1]. The text build on the concepts introduced in the National Corpus of Polish (Przepiórkowski et al., 2010, NKJP)[2], pointing out availability of Sejm data in PDF format and suggesting further steps in the process: inclusion of Senate data and audio recordings. The first phase of the work could be completed only with an European CESAR project[3] in 2011, when all then available data was gathered[4] The data was retrieved from internal Sejm databases and compared to previously available NKJP data. The resource has been made available as The Polish Sejm Corpus[5] (Ogrodniczuk, 2012). The texts have been encoded in NKJP-based TEI P5 format and following layers of linguistic annotation available in NKJP: paragraph-, sentence- and token-level segmentation, lemmatization, disambiguated morphosyntactic information, named entities, syntax words and groups. A searchable corpus version has also been 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 (`http://sejm.nlp.ipipan.waw.pl/`).

At the same time the texts were processed, although much more fragmentarily, by many other researchers in Poland. Parliamentary proceedings were included in the major written corpora such as the IPI PAN Corpus (Przepiórkowski, 2004), National Corpus of Polish (Przepiórkowski et al., 2010) with its distributable subcorpus[6], KPWr corpus (Broda et al., 2012) or the internal corpora of the Polish-Japanese Academy of Information Technology.

Over the years many new ideas were put forward calling for the update of the Sejm corpus. Apart from the constant flow of new data, language processing tools of much higher quality have been made available and new parliamentary resources have been produced by the Parliament itself, the most important of which were all transcripts of parliamentary proceedings from 1918–1990[7] digitized by the Sejm Library. Even though they were made available only in the form of images, this was a very important step towards the completion of the resource, now ready to include all 100 years of newest parliamentary history of Poland. Last but not least, the data from Polish Senate, similar in character but originally omitted from the corpus, were ready to be added.

The usage of the current corpus brought another motivation for maintenance of a separate parliamentary resource: the data has been widely popular among representatives of both the humanities and computational linguists[8]. As compared to other domains, the data features a broad spectrum of topics despite its controlled flavour. The usefulness of such setting also seems to be confirmed by several international initiatives related to parliamentary data such as the recent CLARIN-PLUS workshop "Working with Parliamentary Records" in Sofia[9] or user involvement queries summarized at the CLARIN conference in Budapest.

All these intermediary steps paved the way for the current version of the corpus which we call the Polish Parliamentary Corpus (PPC). The next sections of the paper describe its contents and construction principles.

[1] See `http://www.ung.si/~jezik/sinfon_3/program.html`.

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

[3] Central and South-East European Resources, a CIP – ICT PSP grant 271022, February 2011 – January 2013.

[4] Sittings from terms of office 1–6 (1991–2011) and questions from terms of office 3–6 (1997–2011).

[5] See `http://clip.ipipan.waw.pl/PSC`

[6] See `http://zil.ipipan.waw.pl/DistrNKJP`

[7] See Parlamentaria website: `https://bs.sejm.gov.pl/F/?func=file&file_name=find-nowe&local_base=ars01`)

[8] See e.g. (Przybyła and Teisseyre, 2014; Marasek et al., 2015; Pęzik, 2015; Szela, 2016).

[9] See `http://www.clarin.eu/event/2017/clarin-plus-workshop-working-parliamentary-records`.

2. Corpus Data and Format

All data available in the Polish Sejm Corpus (stenographic records from 1991–2011, interpellations and questions from 1997–2011) has been included in the resulting resource. The Senate data from 1991–2011 has been retrieved from the distributable subcorpus of the National Corpus of Polish; later data has been newly harvested from the Senate website. The texts of interpellations and questions not available in the Polish Sejm Corpus have been acquired from Sejm website[10] using the P4 toolset by Daniel Janus.

However, the most interesting portion of Polish parliamentary transcripts dating to 1918–1990 which has never been made available before has been added to the resource only recently. The digitized data in the form of image-based PDF files has been retrieved from Parlamentaria website, passed through FineReader OCR tool and manually verified by human proof-readers.

The corpus follows the XML TEI P5-based annotation model put forward by the National Corpus of Polish[11] — a *de facto* standard for encoding and documenting Polish linguistic data. The format assumes stand-off linguistic annotation distributed over various layers represented in separate files:

- `header.xml`, covering detailed metadata of the sitting (sitting number/day, list of speakers etc.)
- `text_structure.xml`, the structure of the sitting split into utterances of the MPs grouped into continuous statements, created with dedicated scripts
- `ann_segmentation.xml.gz`, sentence- and token-level segmentation, created with Morfeusz SGJP (Woliński, 2006)
- `ann_morphosyntax.xml.gz`, disambiguated morphosyntactic annotation and lemma information, created with Morfeusz SGJP and Toygger tagger (Krasnowska-Kieraś, 2017)
- `ann_words.xml.gz`, syntactic words, created with Spejd (Buczyński and Przepiórkowski, 2009) shallow parser
- `ann_groups.xml.gz`, syntactic groups, created with Spejd
- `ann_named.xml.gz`, named entities, created with NERF (Savary et al., 2010)
- `ann_coreference.xml.gz`, mentions and coreference annotation, created with the newest neural system (Nitoń et al., 2018).

For detailed description of the format structure see (Ogrodniczuk, 2012).

[10] See `http://www.sejm.gov.pl`.

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

3. Corpus Statistics and Availability

The current size of the corpus amounts to 194M segments with detailed distribution over houses and periods presented in Table 1. Apart from the stenographic records the corpus contains 104M segments of interpellations and questions.

Several interfaces have been made available to access corpus data such as a familiar Poliqarp-based (Janus and Przepiórkowski, 2006) interface at `http://sejm.nlp.ipipan.waw.pl/` (see Figure 1) or a more utterance-centric Smyrna-based (Janus, 2015) interface at `http://smyrna.sejm.nlp.ipipan.waw.pl/` (see Figure 2). Poliqarp binary package has also been made available to facilitate offline statistical queries, currently available only in the desktop version of the search engine.

4. Current and Future Work

Providing the corpus data together with some basic search capabilities is just the first step in the long process of making the data usable by representatives of the digital humanities, the most interested in the parliamentary resources. Apart from natural directions of development of the corpus (improving search and presentation, adding more annotation layers, richer metadata, audio/video linking) three of them are particularly worthy of note.

The first of them is inclusion of the remaining parliamentary data, now available only in the paper form: the first of them being the proceedings of committees, both standing and select. Their processing has already started and will be continued until the end of 2018.

The linguistic engineering tasks seem equally important. First type of them relates to improved processing of data with newest tools. Due to the deep neural network revolution new processing applications are being made available every year with improved accuracy reached on many levels of linguistic annotation[12] In case of large automatically tagged corpora even small progress results in much better quality of data. Another group of NLP tasks concerns the fact that the data span a long period of time with at least two important changes in the Polish orthography (a major reform in 1936 and some minor adaptations after 1956). Obviously, the accuracy of processing of such data with modern tools drops significantly with orthographic alterations which calls for development of separate tools for different historical periods. Such experiments for morphological analysis have already been carried out, cf. e.g. (Kieraś et al., 2017).

Last but not least, such a diverse dataset (as far as the topic and date span is concerned) would benefit from presentation interface resembling Google Books Ngram Viewer, capable of visualising term frequency differences. A similar solution has recently been adopted for Chronopress (Pawłowski, 2016), a corpus of post-war Polish press (until 1962).

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.

[12] See e.g. Toygger, a new tagger of Polish (Krasnowska-Kieraś, 2017).

Period	Years	Sejm	Sittings	Segments	Senate	Sittings	Segments
Second	1919–1922	Legislative Sejm	342	208 524	–	–	–
Polish	1922–1927	1st term of office	340	533 109	1st term	157	153 980
Republic	1928–1930	2nd	86	111 894	2nd	31	5 790
	1930–1935	3rd	148	139 166	3rd	80	123 651
	1935–1938	4th	90	141 387	4th	54	56 191
	1938–1939	5th	31	44 592	5th	20	38 776
	1943–1947	State National Council	11	11 671	–	–	–
People's	1947–1952	Legislative Sejm	108	514 572	–	–	–
Poland	1952–1956	1st term of office	39	76 244	–	–	–
	1957–1961	2nd	59	115 563	–	–	–
	1961–1965	3rd	32	62 799	–	–	–
	1965–1969	4th	23	45 233	–	–	–
	1969–1972	5th	19	33 492	–	–	–
	1972–1976	6th	32	63 155	–	–	–
	1976–1980	7th	29	57 352	–	–	–
	1980–1985	8th	79	132 845	–	–	–
	1985–1989	9th	50	89 436	–	–	–
Third	1989–1991	10th	79	157 225	1st term	61	111 450
Polish	1991–1993	1st term of office	45	7 803 935	2nd	38	1 461 165
Republic	1993–1997	2nd	115	22 299 861	3rd	102	5 057 468
	1997–2001	3rd	119	24 313 939	4th	90	8 261 548
	2001–2005	4th	109	28 986 555	5th	88	6 489 812
	2005–2007	5th	48	11 833 471	6th	39	3 573 955
	2007–2011	6th	100	22 682 341	7th	83	8 827 024
	2011–2015	7th	102	22 587 764	8th	82	7 110 114
	2015–	8th	54	5 905 461	9th	53	3 504 637

Table 1: Statistics of the Polish Parliamentary Corpus

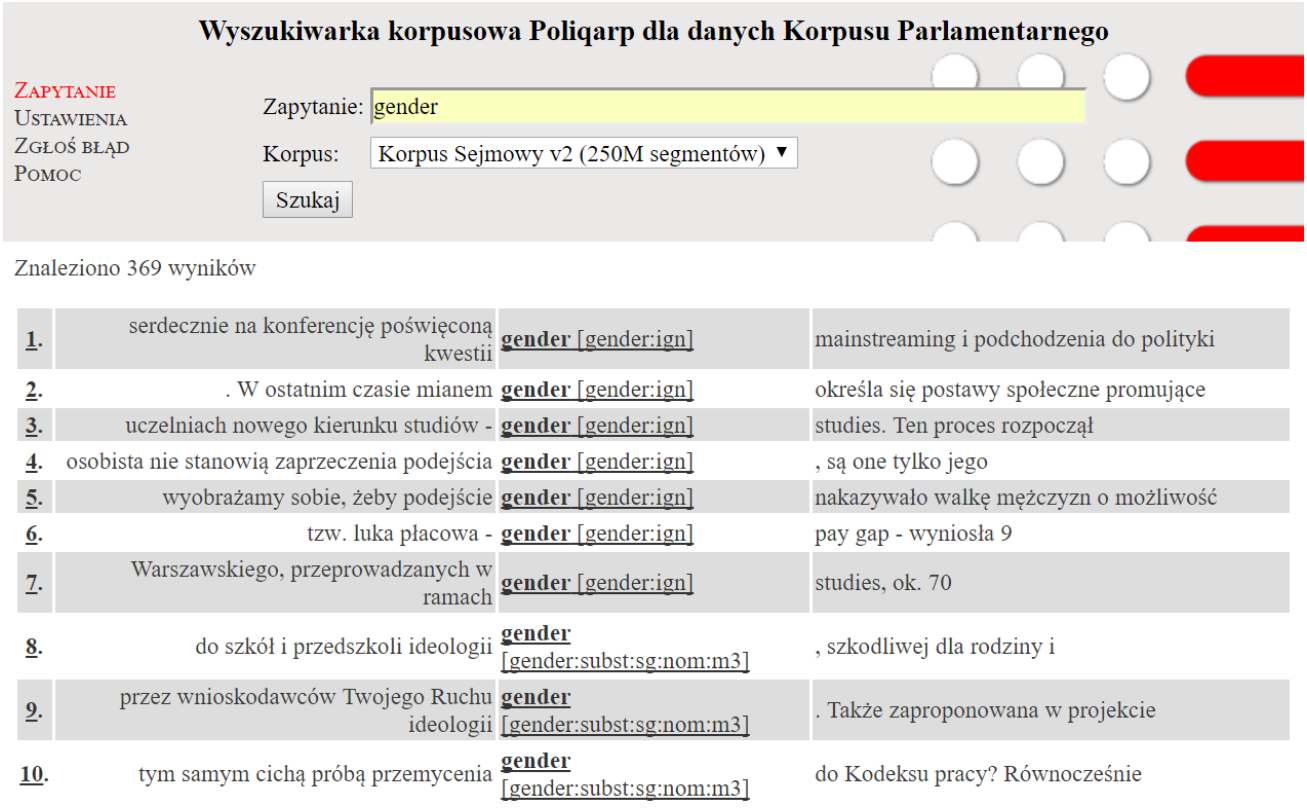

Figure 1: NKJP-based Poliqarp search in the corpus

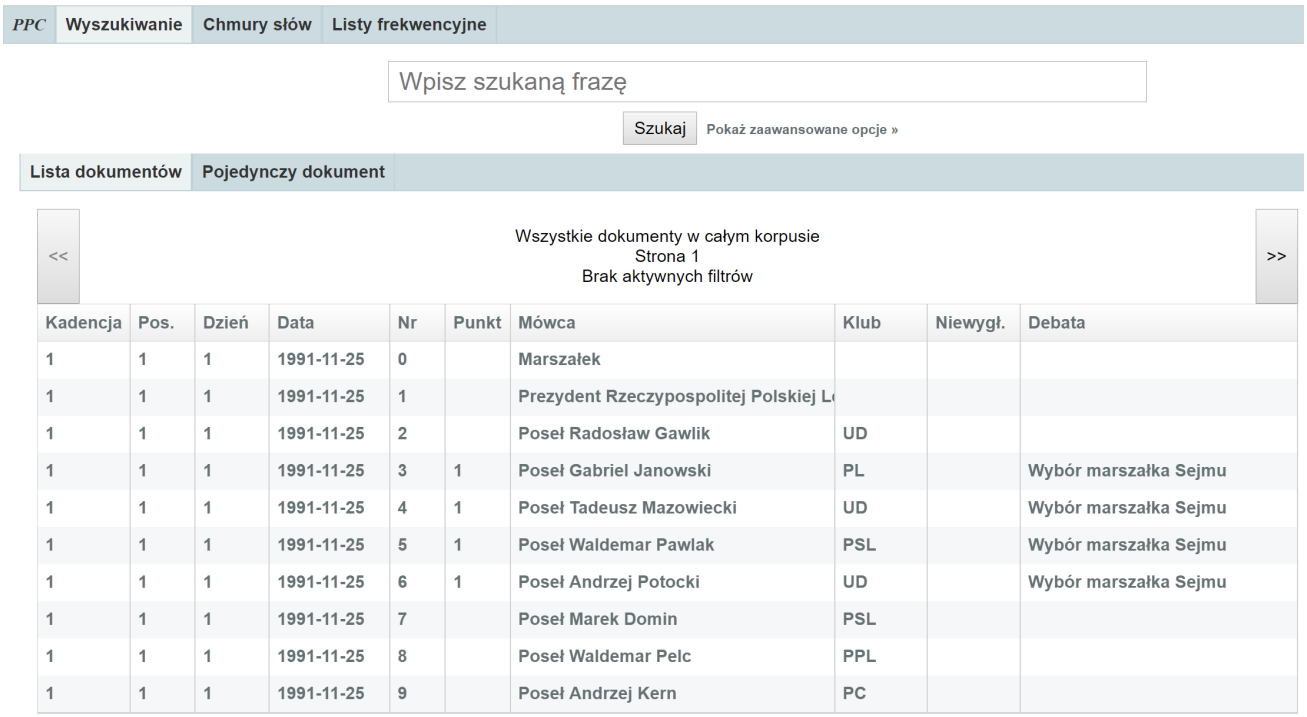

Figure 2: Smyrna-based search in the corpus

Bibliographical References

Broda, B., Marcińczuk, M., Maziarz, M., Radziszewski, A., and Wardyński, A. (2012). KPWr: Towards a Free Corpus of Polish. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pages 3218–3222, Istanbul, Turkey. European Language Resources Association (ELRA).

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

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, *The proceedings of Practical Applications of Linguistic Corpora 2005*, Frankfurt am Main. Peter Lang.

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, *Proceedings of the Twentieth International Conference Text, Speech, and Dialogue (TSD 2017)*, volume 10415 of *Lecture Notes in Computer Science*, pages 308–316. Springer International Publishing.

Krasnowska-Kieraś, K. (2017). Morphosyntactic disambiguation for Polish with bi-LSTM neural networks. In Zygmunt Vetulani et al., editors, *Proceedings of the Eighth Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 367–371, Poznań, Poland. Fundacja Uniwersytetu im. Adama Mickiewicza w Poznaniu.

Marasek, K., Koržinek, D., and Brocki, Ł. (2015). System for Automatic Transcription of Sessions of the Polish Senate. *Archives of Acoustics*, 39(4).

Nitoń, B., Morawiecki, P., and Ogrodniczuk, M. (2018). Deep Neural Networks for Coreference Resolution for Polish. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018*, Miyazaki, Japan. European Language Resources Association (ELRA).

Ogrodniczuk, M. (2012). The Polish Sejm Corpus. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eighth International Conference on Language Resources and Evaluation, LREC 2012*, pages 2219–2223, Istanbul, Turkey. European Language Resources Association (ELRA).

Przepiórkowski, A., Górski, R. L., Łaziński, M., and Pęzik, P. (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.

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

Przybyła, P. and Teisseyre, P. (2014). Analysing utterances in Polish parliament to predict speaker's background. *Journal of Quantitative Linguistics*, 21(4):350–376.

Pęzik, P. (2015). Spokes – a search and exploration service for conversational corpus data. In *Selected Papers from CLARIN 2014*, Linköping Electronic Conference Proceedings, pages 99–109. Linköping University Electronic Press, Linköpings universitet.

Savary, A., Waszczuk, J., and Przepiórkowski, A. (2010). Towards the Annotation of Named Entities in the National Corpus of Polish. In Nicoletta Calzolari, et al.,

editors, *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC 2010)*, pages 3622–3629, Valletta. European Language Resources Association.

Szela, M. (2016). O wykorzystaniu angielsko-polskiego korpusu równoległego tekstów prawnych w badaniu cech języka tekstów tłumaczonych. In Agnieszka Gruszczyńska, Ewa; Leńko-Szymańska, editor, *Polsko-językowe korpusy równoległe*, pages 210–226. Instytut Lingwistyki Stosowanej, Warszawa.

Woliński, M. (2006). Morfeusz — a practical tool for the morphological analysis of Polish. In Mieczysław A. Kłopotek, et al., editors, *Proceedings of the International Intelligent Information Systems: Intelligent Information Processing and Web Mining'06 Conference*, pages 511–520, Wisła, Poland, June.

Language Resource References

Janus, D. (2015). *Smyrna: prosty konkordancer obsługujący język polski*. Available: `http://smyrna.danieljanus.pl`.

Pawłowski, A. (2016). *ChronoPress – Chronologica Corpus*. Copyright CC BY-NC-SA 3.0 (Attribution-NonCommercial-ShareAlike 3.0 Unported). Available in CLARIN-PL digital repository: `http://hdl.handle.net/11321/260`.