Polish Round Table Corpus

Maciej Ogrodniczuk[1], Ryszard Tuora[1] and Beata Wójtowicz[1,2]
[1]Institute of Computer Science, Polish Academy of Sciences;
[2]University of Warsaw;
maciej.ogrodniczuk@ipipan.waw.pl; ryszardtuora@gmail.com;
b.wojtowicz@uw.edu.pl

Abstract

The paper describes the process of preparation of the Polish Round Table Corpus (Pol. *Korpus Okrągłego Stołu*), a new resource documenting negotiations taking place in 1989 between the representatives of the communist government of the People's Republic of Poland and the Solidarity opposition. The process consisted of OCR of graphical transcripts of the talks stored in the form of parliament-like stenographic transcripts, carrying out their manual correction and making them available for search in a concordancer currently used for standard parliamentary transcripts.

Keywords: parliamentary data, Polish Round Table negotiations, contemporary history

1. Introduction

In 1988, against the backdrop of a growing wave of strikes and social protests, the authorities in communist People's Republic of Poland entered into negotiations with a section of the opposition (Solidarity movement, led by Lech Wałęsa) to resolve a simmering political conflict. Their final phase was the so-called 'Round Table', held in 1989 between 6 February and 5 April, with representatives of the Catholic Church acting as mediators. These talks marked the beginning of major political changes in Poland and accelerated the collapse of the entire communist bloc in Europe which makes them an important event in the recent history.

Round tables were about building a community of all people being equal. During the meeting, three main negotiating committees (the so-called tables) were established. The first was devoted to discussing the issue of trade union pluralism, the second one dealt with problems of economy and social policy, while the third team focused on the issue of political reforms. In addition to the committees, sub-committees (the so-called sub-tables) were also created. They were engaged in agriculture, mining, law and court reforms, associations and local governments, youth, mass media, housing, science, education and technical progress, health, ecology, wage and income indexation. A total of eleven sub-teams worked simultaneously headed by the country's main political leaders of that time. Several hundred people (participants, experts and observers) took part in the deliberations of all the teams, sub-teams and working groups (Polak and Galij-Skarbińska, 2021).

Although the Round Table negotiations were not part of the official parliamentary debate, they were documented in a form identical to the Polish parliamentary transcripts and are officially available on the Sejm website[1] as graphic PDF documents, without the text layer. This motivated us to make them available for searching in the concordance similarly (though separately from) the Polish Parliamentary Corpus (Ogrodniczuk, 2012, 2018)[2].

2. Data Preparation

2.1. Original Data Format

The original dataset consists of 96 documents contained in nearly 14,500 (A4) pages. Each (sub)table produced 1 to 13 meeting transcripts written on a typewriter (see Fig. 1). The documents vary in size from couple of dozen to 270 pages. They also vary in quality, due to unequal print visibility, writing errors, and handwritten notes that make the document less readable.

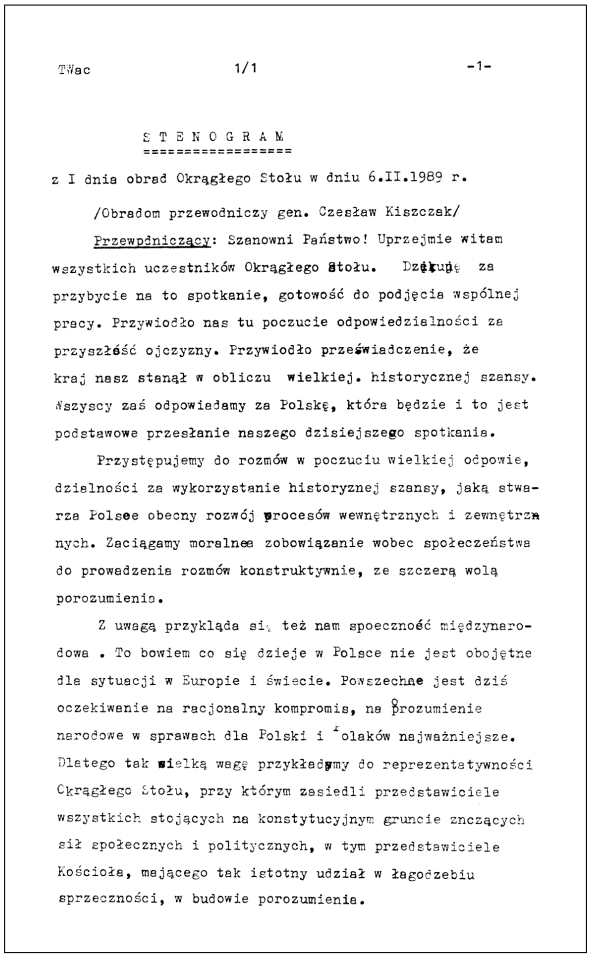
The documents follow a fairly consistent format for specifying the metadata, speakers' name, or interruptions, compatible with the one used while recording parliamentary sessions.

Fig. 1 illustrates well the quality of the transcript; already on its first page the number of problems of various kind is very high:

- 9 words with overwritten wrong characters
- one case of missing hyphenation (`odpowie, działności`)

[1] https://www.sejm.gov.pl/sejm7.nsf/stenOkrStol.xsp
[2] https://kdp.nlp.ipipan.waw.pl/

ParlaCLARIN IV Workshop on Creating, Analysing, and Increasing Accessibility of Parliamentary Corpora, pages 43–47



TWac 1/1 -1-

S T E N O G R A M

z I dnia obrad Okrągłego Stołu w dniu 6.II.1989 r.

/Obradom przewodniczy gen. Czesław Kiszczak/

Przewpdniczący: Szanowni Państwo! Uprzejmie witam wszystkich uczestników Okrągłego Stołu. Dziękuję za przybycie na to spotkanie, gotowość do podjęcia wspólnej pracy. Przywiodło nas tu poczucie odpowiedzialności za przyszłość ojczyzny. Przywiodło przeświadczenie, że kraj nasz stanął w obliczu wielkiej. historycznej szansy. Wszyscy zaś odpowiadamy za Polskę, która będzie i to jest podstawowe przesłanie naszego dzisiejszego spotkania.

Przystępujemy do rozmów w poczuciu wielkiej odpowie, dzialności za wykorzystanie historycznej szansy, jaką stwarza Polsee obecny rozwój procesów wewnętrznych i zewnętrznych. Zaciągamy moralnee zobowiązanie wobec społeczeństwa do prowadzenia rozmów konstruktywnie, ze szczerą wolą porozumienia.

Z uwagą przykląda się też nam spoeczność międzynarodowa . To bowiem co się dzieje w Polsce nie jest obojętne dla sytuacji w Europie i świecie. Powszechne jest dziś oczekiwanie na racjonalny kompromis, na porozumienie narodowe w sprawach dla Polski i Polaków najważniejsze. Dlatego tak wielką wagę przykładamy do reprezentatywności Okrągłego Stołu, przy którym zasiedli przedstawiciele wszystkich stojących na konstytucyjnym gruncie znczących sił społecznych i politycznych, w tym przedstawiciele Kościoła, mającego tak istotny udział w łagodzebiu sprzeczności, w budowie porozumienia.

Figure 1: The first page of the transcript of the inaugural meeting of the Round Table on February 6, 1989.

- almost unreadable end character in a word (`się`)
- 6 words with various uncorrected typos (`historyznej` → `historycznej`, `moralnee` → `moralne`, `przykląda` → `przygląda`, `spoeczność` → `społeczność`, `znczących` → `znaczących`, `łagodzebiu` → `łagodzeniu`)
- one correction made by adding the missing character over the word (`prozumienie` → `porozumienie`)
- one character typed over the line (`Polaków`)
- one case of wrong punctuation (dot in place of a comma: `wielkiej. historycznej szansy`)

All writing flaws and text imperfections had a negative impact on the quality of the OCR process. Therefore an additional phase of manual correction had to be introduced.

2.2. Data Conversion and Annotation

The transcripts were OCR-ed with ABBYY FineReader 12 under manual supervision, and initially reviewed by an annotator (all errors noted down in the previous section were successfully corrected in this process). After this phase, the documents were converted to HTML format, to preserve their structural features. Subsequently, the data was cleaned and homogenized using semi-manual techniques (e.g. regular expressions). Additionally normalization of speakers was applied to around 200 most frequent vocal participants (originally multiple aliases[3] per person were used).

The data was then processed using the pl_nask model[4] for spaCy (Honnibal et al.), to provide POS tagging, lemmatization, dependency parsing and named entity recognition.

2.3. Data Statistics

The corpus consists of 96 documents, which amount to 3 272 149 tokens, 162 595 sentences, 67 185 paragraphs and 23 437 speeches.

The most frequent speaker designation is 'Chairman', decoded in the initial section of the transcript. It is also very common that speeches are not attributed to anyone (see Table 1).

Speaker	Speeches
Chairman	7507
Missing	2997
Jerzy Kołodziejski	744
Władysław Baka	610
Łukasz Balcer	407
Chairwoman	375
Stefan Kozłowski	325
Jan Brol	322
Adam Strzembosz	310
Alojzy Pietrzyk	258
Voice from the audience	244
Witold Trzeciakowski	242
Rajmund Moric	211
Tadeusz Mazowiecki	202
Bronisław Geremek	189

Table 1: Most frequent speakers.

[3] For instance: Aleksander Kwaśniewski, the then minister-without-portfolio in the People's republic of Poland figures in text as 'minister Kwaśniewski', 'colleague Kwaśniewski', 'deputee Kwaśniewski' etc. It was not possible to provide full normalization of speakers, as in some cases (e.g. when two participants share a surname, or a private person is speaking, with no full name given) attributions are ambiguous.

[4] https://huggingface.co/ipipan/pl_nask

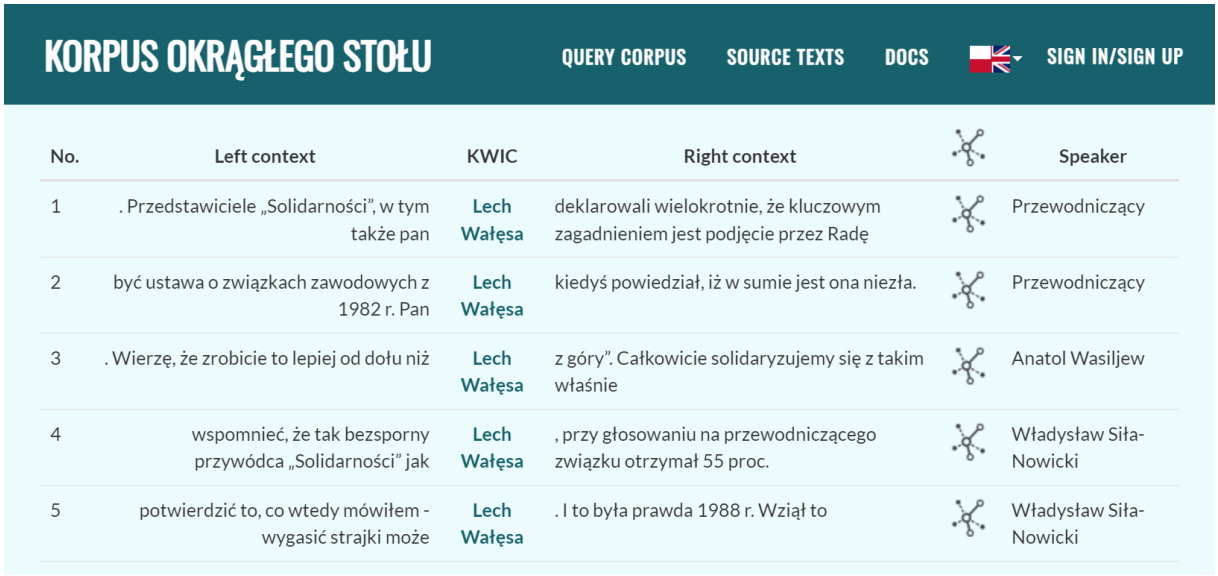


Figure 2: A KWIC index offered by Korpusomat.

The statistics of speeches within specific committees and sub-committees (see Table 2) illustrates the importance of the topics discussed.

Committee	Speeches
Economy and Social Policy	4874
Law and Court Reform	2756
Ecology	2650
Health	1899
Union Pluralism	1987
Political Reform	1758
Mining	1602
Associations and Local Gov.	1504
Agriculture	1071
Housing Policy	1005
Science, Education and Technical Progress	850
Youth Affairs	584
Mass Media	449
Wage and Income Indexation	391
Plenary Sessions	57

Table 2: Statistics of speeches within specific committees.

3. Searching the Corpus

Finally the documents were indexed in KORPUSOMAT (Kieraś et al., 2018; Saputa et al., 2023) — an established Web application for accessing and working with corpus data (see Figure 2). The transcripts are searchable, using both the annotation layers, and metatextual information (i.e. speaker names, or metadata such as committee name).

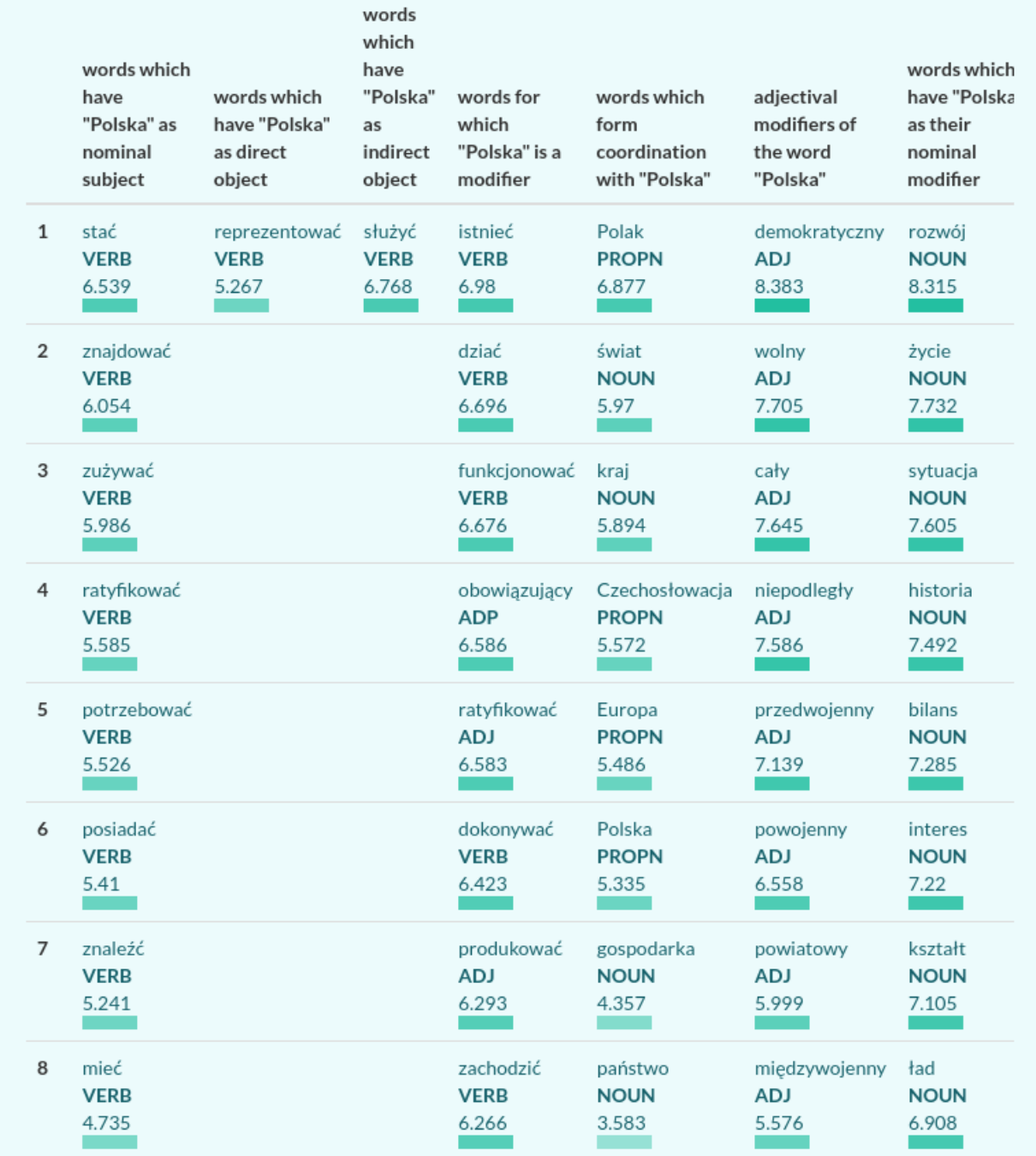
Additionally, the 'word profile' functionality was employed, which allows to visualize how a particular word is used in the corpus (see Figure 3) by surveying regularities in grammatical connections it enters into with other words.

4. Future Work

Even though the data conversion process involved manual interventions at various stages, the data still needs many manual updates. Known types of errors include:

- typos introduced by the stenographer and corresponding to in-vocabulary Polish words (such as `patynie` instead of `pytanie` on page 4 in the first session), undetectable without careful revision of the text
- wrong recognition of mostly Polish characters during the OCR process (such as `sie` instead of `się`, `l` instead of `i`), which are difficult to spot
- obvious slips which are always corrected in the official transcript (e.g. `w sprawach najważniejsze` → `najważniejszych`)
- typographical errors, including editing errors, e.g. introducing unnecessary characters, like extra spaces, in the text
- names of speakers' functions (e.g. "chairman") used in place of their names after the function assignment to the speaker is recorded in the commentary on the earlier part of the transcript (see Fig. 1, line in typewriter brackets directly over the underlined designation).

Despite such errors, the transcripts make a valuable documentation of the Polish bloodless road



	words which have "Polska" as nominal subject	words which have "Polska" as direct object	words which have "Polska" as indirect object	words for which "Polska" is a modifier	words which form coordination with "Polska"	adjectival modifiers of the word "Polska"	words which have "Polska as their nominal modifier
1	stać VERB 6.539	reprezentować VERB 5.267	służyć VERB 6.768	istnieć VERB 6.98	Polak PROPN 6.877	demokratyczny ADJ 8.383	rozwój NOUN 8.315
2	znajdować VERB 6.054			dziać VERB 6.696	świat NOUN 5.97	wolny ADJ 7.705	życie NOUN 7.732
3	zużywać VERB 5.986			funkcjonować VERB 6.676	kraj NOUN 5.894	cały ADJ 7.645	sytuacja NOUN 7.605
4	ratyfikować VERB 5.585			obowiązujący ADP 6.586	Czechosłowacja PROPN 5.572	niepodległy ADJ 7.586	historia NOUN 7.492
5	potrzebować VERB 5.526			ratyfikować ADJ 6.583	Europa PROPN 5.486	przedwojenny ADJ 7.139	bilans NOUN 7.285
6	posiadać VERB 5.41			dokonywać VERB 6.423	Polska PROPN 5.335	powojenny ADJ 6.558	interes NOUN 7.22
7	znaleźć VERB 5.241			produkować ADJ 6.293	gospodarka NOUN 4.357	powiatowy ADJ 5.999	kształt NOUN 7.105
8	mieć VERB 4.735			zachodzić VERB 6.266	państwo NOUN 3.583	międzywojenny ADJ 5.576	ład NOUN 6.908

Figure 3: Word profile generated for the word "Polska" (Poland), which occurs a total of 2030 times. The figures correspond to logDICE values for each collocation.

to democracy and its searchable variant will definitely help the digital humanities researchers in their work.

Acknowledgements

The work was financed by the European Regional Development Fund as a part of the 2014–2020 Smart Growth Operational Programme, CLARIN — Common Language Resources and Technology Infrastructure, project no. POIR.04.02.00–00C002/19[5], the Polish Ministry of Education and Science grant 2022/WK/09 and as part of the investment CLARIN ERIC — European Research Infrastructure Consortium: Common Language Resources and Technology Infrastructure (period: 2024–2026) funded by the Polish Ministry of Science and Higher Education (Programme: "Support for the participation of Polish scientific teams in international research infrastructure projects"), agreement number 2024/WK/01.

[5] https://clarin.biz/

5. Bibliographical References

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python.

Witold Kieraś, Łukasz Kobyliński, and Maciej Ogrodniczuk. 2018. Korpusomat — a tool for creating searchable morphosyntactically tagged corpora. *Computational Methods in Science and Technology*, 24(1):21–27.

Maciej Ogrodniczuk. 2012. The Polish Sejm Corpus. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC 2012)*, pages 2219–2223, Istanbul, Turkey. European Language Resource Association.

Maciej Ogrodniczuk. 2018. Polish Parliamentary Corpus. In *Proceedings of the LREC 2018 Workshop* ParlaCLARIN: Creating and Using Parliamentary Corpora, pages 15–19, Paris, France. European Language Resources Association.

Wojciech Polak and Sylwia Galij-Skarbińska. 2021. The Round Table in 1989 — Consequences and Evaluation. *Polish Political Science Yearbook*, 50:149–156.

Karol Saputa, Aleksandra Tomaszewska, Natalia Zawadzka-Paluektau, Witold Kieraś, and Łukasz Kobyliński. 2023. Korpusomat.eu: A multilingual platform for building and analysing linguistic corpora. In *Computational Science – ICCS 2023. 23rd International Conference, Prague, Czech Republic, July 3–5, 2023, Proceedings, Part II*, number 14074 in Lecture Notes in Computer Science, pages 230–237, Cham. Springer Nature Switzerland.