The Corpus of Contemporary Polish: 2011-2020 Decade and Beyond

Witold Kieraś, Małgorzata Marciniak, Katarzyna Krasnowska-Kieraś, Marcin Woliński
Institute of Computer Science, Polish Academy of Sciences
Jana Kazimierza 5, 01-248 Warszawa, Poland
{w.kieras, m.marciniak, k.krasnowska-kieras,m.wolinski}@ipipan.waw.pl

Abstract

The aim of this poster is to present the Contemporary Corpus of Polish (KWJP), a new reference resource spanning the period of 2011–2020. The KWJP complements the now discontinued National Corpus of Polish project (NKJP, Przepiórkowski et al. 2012) by providing up-to-date linguistic data. It comprises a 100M-token balanced sub-corpus alongside a larger 1.5B-token unbalanced (opportunistic) component, consisting of books and periodicals not included in the balanced part. While the corpus contains almost exclusively copyrighted material and is therefore accessible only via a web-based search engine, a representative 0.5M-token sample has been published as open-source data. Details of the resource description that fall beyond the scope of this abstract can be found in a paper accepted for the main LREC 2026 conference (Kieraś et al., 2026).

While the KWJP was conceived as a successor to the NKJP for the subsequent decade, it differs from its predecessor in several key respects. A primary distinction lies in its temporal scope: whereas the NKJP aimed to represent written Polish from the early 20th century onwards, the KWJP focuses exclusively on a single decade of contemporary texts. Given the emergence of numerous specialized resources—such as the Corpus of Parliamentary Discourse (Ogrodniczuk, 2018), the MoncoPL web monitoring corpus (Pęzik, 2020), and various spoken language corpora (Pęzik, 2015)—the necessity for a general reference corpus to cover every possible genre has diminished. Consequently, the KWJP focuses strictly on edited texts, namely books (both fiction and non-fiction) and a broad selection of national and regional periodicals.

Consequently, the KWJP's text type classification and the proportions represented in the corpus have been significantly simplified, now consisting of three categories: fiction, non-fiction, and journalism. Fiction (30% of the balanced corpus) primarily comprises literary books (novels and short story collections) across various genres, along with a limited selection of literary periodicals that publish predominantly short stories. The non-fiction category (35%) encompasses a broad range of texts, including journalistic books, diaries, biographies, travel guides, popular science, and scholarly works, as well as official documents—all of which were distributed across several distinct labels in the NKJP. Unlike the NKJP, thematic magazines are also classified as non-fiction. Finally, the journalism genre (35%) consists exclusively of traditional news and public affairs press, mainly national and regional daily and weekly newspapers, supplemented by a selection of monthly, bi-monthly, and annual periodicals.

Distribution channels are categorized into two primary types: books and press. Only a negligible portion of the texts (0.3%) is assigned to the internet channel, representing a sample of court rulings from various judicial instances—a specific type of official document. In all other cases, classification into the book or press channel follows standard library identification schemes, specifically ISBNs for books and ISSNs for the press, even for publications existing solely in electronic format. Books comprise approximately 55% of the balanced corpus, while the press accounts for the remaining 45% (daily newspapers: 19%; weekly magazines: 12%; monthly periodicals: 9.5%; other: 5.5%).

The KWJP features rich, multi-layer annotation, generally adhering to the NKJP annotation scheme. Rules for segmentation (tokenization) are followed directly. Regarding morphosyntactic tagging, the tagset has been aligned with the latest version of the Morfeusz morphological analyzer for Polish (Kieraś and Woliński, 2017; Woliński, 2014). Additionally, the KWJP introduces two new annotation layers: named entities (NE) and syntactic structures. Named entity annotation strictly follows the schema used in the one-million-word manually annotated subcorpus of the NKJP (NKJP1M, Przepiórkowski et al. 2012). As opposed to the NKJP, the automatic NE layer in the KWJP covers the entire resource. The syntactic layer is entirely new compared to the NKJP and comprises hybrid tree structures that combine both dependency and constituency relations (Krasnowska-Kieraś and Woliński, 2024). All annotation layers are accessi-

Proceedings of the 12th Workshop on Challenges in the Management of Large Corpora (CMLC-12) @ LREC 2026, pages 78–79. Palma, Mallorca. May 11, 2026.

ble (to a certain extent) via corpus queries (CQL).

The KWJP project aims to establish a team and infrastructure for the long-term development of the resource. The first update is scheduled for 2026 and will introduce a sub-corpus covering the 2021–2025 period. We intend to maintain a regular five-year update cycle, similar to the SYN corpora series developed by the Czech National Corpus team (Křen et al., 2016). The estimated minimum size of the update is 50 million tokens, keeping the same proportions as the original 2011–2020 corpus. Simultaneously, technical work is underway to provide a new, more efficient search engine that will support the development of web-based applications and the gathering of statistics for linguistic research.

1. Bibliographical References

Witold Kieraś, Małgorzata Marciniak, Marcin Woliński, Katarzyna Krasnowska-Kieraś, and Marek Łaziński. 2026. The Corpus of Contemporary Polish — a New Reference Corpus with Rich Syntactic Annotations. In *Proceedings of LREC 2026*, Palma, Spain.

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

Katarzyna Krasnowska-Kieraś and Marcin Woliński. 2024. Parsing headed constituencies. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 12633–12643, Turin, Italy. ELRA and ICCL.

Michal Křen, Václav Cvrček, Tomáš Čapka, Anna Čermáková, Milena Hnátková, Lucie Chlumská, Tomáš Jelínek, Dominika Kováříková, Vladimír Petkevič, Pavel Procházka, Hana Skoumalová, Michal Škrabal, Petr Truneček, Pavel Vondřička, and Adrian Jan Zasina. 2016. SYN2015: Representative corpus of contemporary written Czech. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2522–2528, Portorož, Slovenia. European Language Resources Association (ELRA).

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

Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, editors. 2012. *Narodowy Korpus Języka Polskiego*. Wydawnictwo Naukowe PWN, Warsaw.

Piotr Pęzik. 2015. Spokes – a search and exploration service for conversational corpus data. In *Selected Papers from the CLARIN 2014 Conference, October 24-25, 2014, Soesterberg, The Netherlands*, Linköping Electronic Conference Proceedings, pages 99–109. Linköping University Electronic Press, Linköpings universitet.

Piotr Pęzik. 2020. Budowa i zastosowania korpusu monitorującego MoncoPL. *Forum Lingwistyczne*, 7(7):133–150.

Marcin Woliński. 2014. Morfeusz reloaded. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation, LREC 2014*, pages 1106–1111, Reykjavík, Iceland. European Language Resources Association (ELRA).