Marek Łaziński

Polish language resources 2021[1]

Abstract

This paper presents digital resources and language technology in Polish. The Polish LT landscape comprises the National Corpus of Polish with 1.5 billion words, a monitor corpus Monco with 7.7 billion words, several parallel corpora including Polish texts, the Polish WordNet with 600 thousand lexical relations, tools for building and maintaining corpora, taggers, lemmatizers and dependency parsers.

Digital language resources and language technology in Poland

Polish has been present on the web for years. In 2020, the number of .pl domains reached almost 2.5 million, in 2021 the number of internet users in Poland added up to 28.8 million, i.e., 87% of the population. In 2022 Polish Wikipedia ranked 11th in terms of the number of articles (currently over 1.5 million). In 2020, 77 percent of Poles used Facebook and 60 percent were Messenger users.

Since the 1990s several written corpora of contemporary Polish have been created, starting with the National Corpus of Polish: nkjp.pl. Constructed in 2007–2011 by the Institute of Computer Sciences Polish Academy of Sciences, the Institute of Polish Language PAS, the University of Łódź, and the Polish Scientific Publishers PWN, the corpus comprises over 1.5 billion words, with 250 million in the balanced part covering texts from 1918 to 2010. All texts are annotated morphosyntactically, 1 million words in a sub-corpus have been fully annotated manually. Two search programs give access to sophisticated morphosyntactic concordance queries and to a collocations search (Przepiórkowski et al. 2012). The continuation of the National Corpus of Polish is the Corpus of the Decade, a project in progress (http://korpus-dekady.ipipan.waw.pl).

There are many parallel corpora with Polish: Polish-English (http://paralela.clarin-pl.eu), Polish-German (http://diaspol.uw.edu.pl/polniem/), and others. Written corpora of historical Polish are also being actively developed. The largest monitor corpus of Polish is Monco PL (monco.frazeo.pl) with over 7.7 billion words and a collocation search (Pęzik 2020). The recently released ELEXIS Polish Web corpus is currently the largest corpus, with over 12 billion tokens.

All of the corpora mentioned above are freely searchable but due to copyright issues they cannot be freely downloaded and further used for language technology

[1] Based on Ogrodniczuk/Łaziński/Miłkowski/Pęzik (in print).

processing. Some small corpora, such as the Polish Corpus at Wrocław University of Technology, Open Subtitles (film subtitles in Polish), Wolne Lektury (Free Lecture) are freely distributable but not balanced and not up-to-date. The ELEXIS corpus is freely downloadable for research purposes because it contains only public web documents but it is not balanced either. A list of over 200 resources and tools for Polish can be found at: http://clip.ipipan.waw.pl/LRT.

The National Corpus of Polish is a basic resource for research in the humanities and the testbed for developing many language technology tools, including the first of their kind for Polish: morphological analyzers, disambiguating taggers or named entity recognizers.

Apart from the National Corpus of Polish, another project which has significantly changed the state of Polish language technology is CLARIN-PL, the Polish part of the pan-European Common Language Resources & Technology Infrastructure aimed at researchers in the humanities and social sciences. The co-operation of many research institutions led to the development of many language technology resources and tools such as:

- Słowosieć, the Polish WordNet, a relational lexico-semantic dictionary of Polish with almost 200 thousand lexemes and 600 thousand lexical relations (Dziob et al. 2019),
- Korpusomat, a corpus creation tool for non-technical users: https://korpusomat.pl/ (Kieraś/Kobyliński 2021),
- COMBO, a neural tagger, lemmatizer and dependency parser (Rybak/Wróblewska 2018),
- SpokesPL – a search engine for Polish conversational data: http://spokes.clarin-pl.eu/.

The development of language technology in Poland is based on four pillars:

- Research labs and groups mainly located at universities and the institutes of the Polish Academy of Sciences,
- Government-based institutions and ministries responsible for drafting strategic documents,
- Companies, both the big international players as well as mid-size companies and startups,
- Independent researchers, without any formal affiliation, often forming informal research groups gathered around meetups.

Linking Language Technology and Natural Language Processing to Artificial Intelligence has already happened in Poland with the advent of deep neural network powered solutions but its consequences are more far reaching than we can imagine. However, even when the technology seems mature enough, its absorption by larger public institutions and companies is proceeding much more slowly.

The availability of deep neural network powered frameworks has moved the focus from tools to resources. Therefore, an awareness of the value of data is still increasing. This includes opening up public data and eliminating legal barriers to the exploration of Polish data under copyright protection.

A crucial, and maybe the most important, factor in the development of Polish Language Technology is the support of the national research community with international cooperation. Polish Language Technology research has already benefited from numerous pan-European initiatives such as ELRC (European Language Resource Coordination, https://lr-coordination.eu/), ELG (European Language Grid, https://www.european-language-grid.eu/) and ELE (European Language Equality, https://european-language-equality.eu/), research infrastructures such as CLARIN and DARIAH, COST Actions and CEF projects. This trend must continue to strengthen the European research community.

References

Dziob, A./Piasecki, M./Rudnicka, E. (2019): plWordNet 4.1 – a linguistically motivated, corpus-based bilingual resource. In: *Proceedings of the 10th Global Wordnet Conference*. Wroclaw, 353–362. https://aclanthology.org/2019.gwc-1.45.

Kieraś, W./Kobyliński, Ł. (2021): Korpusomat – stan obecny i przyszłość projektu. In: *Język Polski* CI (2), 49–58.

Ogrodniczuk, M./Pęzik, P./Łaziński, M./Miłkowski, M. (in print): *European Language Equality. D.1.217 Report on the Polish Language*. European Language Resource Coordination.

Pęzik, P. (2020): Budowa i zastosowania korpusu monitorującego MoncoPL. In: *Forum Lingwistyczne* (7), 133–150. http://doi.org/10.31261/FL.2020.07.11.

Przepiórkowski, A./Bańko, M./Górski, R. L./Lewandowska-Tomaszczyk, B. (eds.) (2012): *Narodowy Korpus Języka Polskiego*. Warsaw.

Rybak, P./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*. Brussels, 45–54. http://www.aclweb.org/anthology/K18-2004.

Bibliographical information

This text is part of the book:

Željko Jozić/Sabine Kirchmeier (eds.) (2022): The role of national language institutions in the digital age. Contributions to the EFNIL Conference 2021 in Cavtat. Budapest: Nyelvtudományi Kutatóközpont/Hungarian Research Centre for Linguistics.

This electronic PDF version of the text is accessible through the EFNIL website at:
http://www.efnil.org