Chapter 29
Language Report Polish

Maciej Ogrodniczuk, Piotr Pęzik, Marek Łaziński, and Marcin Miłkowski

Abstract The quality of language technology (LT) for Polish has greatly improved recently, influenced by three independent trends. The first one is Poland-specific and concerns the increase in national funding of both scientific and R&D projects, resulting in the construction of The National Corpus of Polish and the development of the CLARIN-PL and DARIAH-PL infrastructures. Two other trends are global: the development of language resources (LRs) and tools by private companies and of course, the deep learning revolution which has led to enormous improvements in the state-of-the-art in all fields of language processing.

1 The Polish Language

Polish is a Slavic language of the Lechitic group, written in Latin script. It is the most spoken West Slavic language in the world. It is the official language of the Republic of Poland and since 2004, the sixth largest official language of the European Union. It is spoken by 10% of EU citizens: about 40 million native speakers and 10 million second language speakers worldwide. In Poland, it is the common spoken and written language and the native language of the vast majority of the population.

Polish exhibits some specific characteristics (Pisarek 2007), which contribute to the richness of the language but present a challenge for computational processing. Word order is relatively free, which is used mostly to stress the importance of information rather than simply following grammatical rules.

Maciej Ogrodniczuk
Inst. of Comp. Science, Polish Academy of Sciences, Poland, maciej.ogrodniczuk@ipipan.waw.pl

Piotr Pęzik
University of Łódź, Poland, piotr.pezik@uni.lodz.pl

Marek Łaziński
University of Warsaw, Poland, m.lazinski@uw.edu.pl

Marcin Miłkowski
Inst. of Philosophy and Sociology, Polish Academy of Sciences, Poland, mmilkows@ifispan.edu.pl

G. Rehm, A. Way (eds.), *European Language Equality*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-28819-7_29

Polish is relatively morphologically rich, which means that for roughly 180,000 base forms of words, almost 4 million inflected word forms exist. The inflection paradigms are complex, and even their exact number is a matter of dispute, as single exceptions might even be thought to create a new paradigm. Even native speakers have problems with properly inflecting many words, and most speakers of Polish as a second language never completely master the complexities of the inflectional system. Polish syntax is similar to its neighbouring Slavic languages with a tendency to analyse constructions seen in gender marking, forms of address and the use of infinitive and impersonal constructions.

Polish is currently highly influenced by English, one of the biggest sources of neologisms and calques, in particular in science and technology. The number of words loaned from English into Polish is, however, much lower than in Dutch or German because of the problem with inflecting some words as well as differences in pronunciation systems. Other recent changes are the appearance of more direct forms of address and simplification of the traditional inflection patterns.

2 Technologies and Resources for Polish

The level of technology support for Polish is similar to that of many other official EU languages, with several available resources[1] and basic text processing tools obtaining satisfactory accuracy scores.[2] The current landscape of Polish language processing has been shaped by the following developments (see Ogrodniczuk et al. 2022; Miłkowski 2012): 1. The construction of the National Corpus of Polish[3] (NKJP; Przepiórkowski et al. 2012), a reference corpus containing over 1.5 billion words sampled from diverse sources such as classical literature, daily newspapers, specialist periodicals and journals, transcripts of conversations, and a variety of short-lived online texts, balanced with respect to gender, age and regional distribution of samples. The availability of the corpus, and particularly its manually annotated 1-million word sub-corpus, available under a CC-BY-licence, has boosted both research in the humanities as well as the development of many NLP tools. Since the completion of the NKJP in 2011, other reference corpora have been used to represent recent developments in Polish. The most significant examples are the MoncoPL monitoring corpus (Pęzik 2020) and the Corpus of the 2010s.[4] 2. The development of the CLARIN-PL[5] and DARIAH-PL[6] infrastructures, led to the development of many resources and tools such as Słowosieć, the Polish WordNet[7] (Dziob et al. 2019), Ko-

[1] http://clip.ipipan.waw.pl/LRT

[2] http://clip.ipipan.waw.pl/benchmarks

[3] http://nkjp.pl

[4] http://korpus-dekady.ipipan.waw.pl

[5] https://clarin-pl.eu, http://clarin.biz

[6] https://dariah.pl, https://lab.dariah.pl

[7] http://plwordnet.pwr.wroc.pl/wordnet/

rpusomat, a corpus creation tool[8] (Kieraś and Kobyliński 2021), COMBO, a neural tagger, lemmatiser and dependency parser[9] (Klimaszewski and Wróblewska 2021), or SpokesPL, a search engine for Polish conversational data.[10] 3. External funding in the form of grants, both European (Horizon 2020, Connecting Europe Facility) or national, distributed by the National Science Centre and National Centre for Research and Development, have allowed many research institutions and companies to increase the budgets of research projects by an order of magnitude, and thus react to commercial demands for speech recognition or dialogue systems. As a result, their NLP products are characterised by state-of-the art performance. 4. The PolEval evaluation campaign for NLP tools for Polish[11] started in 2017 as a practical exercise intended to advance the state-of-the-art with a series of tasks in which submitted tools compete against one another. This contest has brought the NLP community together and resulted in the development, enhancement and public release of reference datasets for tasks such as sentiment analysis, speech recognition and machine translation. 5. The latest Transformer models (HerBERT[12] and plT5[13]) trained by researchers from the company Allegro and the Institute of Computer Science of the Polish Academy of Sciences, based on several large corpora of Polish, including NKJP. Making these models freely available for the community has facilitated enormous progress. 6. Increased accessibility of multimodal spoken corpora and speech databases such as a large annotated corpus of phone-based customer support dialogues,[14] which boosts the development of goal-oriented chatbots and helps Polish ASR engines to be on par with solutions by global service providers. Nonetheless, many complex and labour-intensive resources such as audio-video corpora and corpora with discourse structure and semantic annotations are practically unavailable.

3 Recommendations and Next Steps

The national Polish AI strategy (Council of Ministers 2020) mentions the development of LT as a short-term goal, supported by national grants for projects related to Polish language processing based on world-leading algorithms. Notably, the document mentions the importance of language data: the need for the elimination of legal barriers to the exploration of language text corpora under copyright protection and awarding projects that make architecture, trained models and training data sets available for common use. This assumption is in line with findings from the Polish NLP community as well as international trends. What needs to be added to this plan is ac-

[8] https://korpusomat.pl

[9] https://github.com/360er0/COMBO

[10] http://spokes.clarin-pl.eu

[11] http://poleval.pl

[12] https://huggingface.co/allegro/herbert-large-cased

[13] https://huggingface.co/allegro/plt5-large

[14] http://pelcra.pl/new/diabiz

cess to common (national or European) computing power to boost the development and optimization of standard language models and secure stable funding for crucial LRs such as the National Corpus of Polish or the Great Dictionary of Polish.

However, there is also a new dimension of this plan, created by the Russian invasion of Ukraine. With 3 million Ukrainian refugees in Poland in 2022, bilingual public administration has become an important new role for the Polish LT community, and is boosting the development of bilingual Polish-Ukrainian resources and tools. On the European level, this new situation calls for the embracing of Ukrainian as one of the languages officially supported by the EU.

References

Council of Ministers (2020). *Polityka dla rozwoju sztucznej inteligencji w Polsce od roku 2020 – The Policy for the development of AI in Poland from 2020*. https://www.gov.pl/web/ai/polityka-dla-rozwoju-sztucznej-inteligencji-w-polsce-od-roku-2020.

Dziob, Agnieszka, Maciej Piasecki, and Ewa Rudnicka (2019). "plWordNet 4.1 – a Linguistically Motivated, Corpus-based Bilingual Resource". In: *Proceedings of the 10th Global Wordnet Conference*. Global Wordnet Association, pp. 353–362.

Kieraś, Witold and Łukasz Kobyliński (2021). "Korpusomat – stan obecny i przyszłość projektu". In: *Język Polski* CI.2, pp. 49–58. DOI: 10.31286/JP.101.2.4.

Klimaszewski, Mateusz and Alina Wróblewska (2021). "COMBO: State-of-the-Art Morphosyntactic Analysis". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. ACL, pp. 50–62.

Miłkowski, Marcin (2012). *Język polski w erze cyfrowej – The Polish Language in the Digital Age*. META-NET White Paper Series: Europe's Languages in the Digital Age. Heidelberg etc.: Springer. http://www.meta-net.eu/whitepapers/volumes/polish.

Ogrodniczuk, Maciej, Piotr Pęzik, Marek Łaziński, and Marcin Miłkowski (2022). *Deliverable D1.27 Report on the Polish Language*. European Language Equality (ELE); EU project no. LC-01641480 – 101018166. https://european-language-equality.eu/reports/language-report-polish.pdf.

Pęzik, Piotr (2020). "Budowa i zastosowania korpusu monitorującego MoncoPL". In: *Forum Lingwistyczne* 7, pp. 133–150. DOI: 10.31261/FL.2020.07.11.

Pisarek, Walery (2007). *The Polish Language*. Warsaw: The Council for the Polish Language.

Przepiórkowski, Adam, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk, eds. (2012). *Narodowy Korpus Języka Polskiego*. Warsaw: PWN. http://nkjp.pl/settings/papers/NKJP_ksiazka.pdf.