sciendo
DOI 10.2478/jazcas-2019-0061

FROM THE NATIONAL CORPUS OF POLISH TO THE POLISH CORPUS INFRASTRUCTURE

MACIEJ OGRODNICZUK[1] – RAFAŁ L. GÓRSKI[2] –
MAREK ŁAZIŃSKI[3] – PIOTR PĘZIK[4]
[1]Institute of Computer Science, Polish Academy of Sciences, Poland
[2]Institute of Polish Language, Polish Academy of Sciences, Poland
[3]Institute of Polish Language, University of Warsaw, Poland
[4]Institute of English Studies, University of Łódź, Poland

OGRODNICZUK, Maciej – GÓRSKI, L. Rafał – ŁAZIŃSKI, Marek – PĘZIK, Piotr: From the National Corpus of Polish to the Polish Corpus Infrastructure. Journal of Linguistics, 2019, Vol. 70, No 2, pp. 315 – 323.

Abstract: The National Corpus of Polish emerged as a cumulative result of many years of work on large reference corpora by computer scientists and linguists in Poland. While its impact on research in linguistics, humanities and language technology is unquestionable and highly significant, the construction of the national corpus was halted in 2011. In the paper we call for activating the research community and funding institutions around the construction of a corpus infrastructure with the national corpus at its heart. It is claimed that on the verge of an artificial intelligence revolution the envisaged Polish Corpus Infrastructure would provide reliable language data, combine available resources and allow easy integration of new ones.

Keywords: corpus linguistics, corpus lexicography, dialect corpora

1 THE NATIONAL CORPUS OF POLISH IN THE CONTEXT OF POLISH CORPUS RESEARCH

The first edition of the National Corpus of Polish (Pol. Narodowy Korpus Języka Polskiego – NKJP; [19]) has found extremely diverse scientific and technological applications. NKJP is still the main reference corpus in lexicography (see e.g. Żmigrodzki et al. [26]), applied linguistics and psycholinguistics (e.g. Riegel et al. [20]) and language modeling (e.g. Mykowiecka et al. [12]). It has been used to boost the accuracy of natural language processing on various tasks, and to develop many tools and resources for Polish such as the Concraft disambiguating tagger (Waszczuk [22]), Polish Dependency Corpus (Wróblewska [25]), Polish Coreference Corpus (Ogrodniczuk et al. [13]), Hask collocation databases (Pęzik [15]), SEJF phraseological dictionary (Czerepowicka [1]) or Walenty valence dictionary (Hajnicz et al. [6]). The National Corpus is cited as the basic resource of linguistic research in hundreds of publications. The NKJP search engines serve more than one million distinct corpus user queries every year, 11% of which originate from outside of Poland. The corpus is used both by national research infrastructures (e.g. CLARIN-PL[1]) and in international projects (e.g. PARSEME[2]).

Parallel to NKJP, a number of independent reference corpora of Polish exist, spanning the period from the early days of the language to the modern era, including the corpus of pre-1500 Old Polish texts (Twardzik and Górski [21]), Electronic Corpus of 17th and 18th century Polish texts (Pol. short Korpus Barokowy, hence KORBA; Gruszczyński et al. [5]), the corpus of the 19th century Polish texts (Derwojedowa et al. [3], Kieraś and Woliński [8]) and the MoncoPL monitor corpus of web-based Polish (Pęzik [18]). However, each of these resources resulted from a separate project and operates independently using custom-made standards of presenting linguistic information in a variety of user interfaces. This fragmentation has naturally given rise to the idea of linking all related corpora through a common federated infrastructure as recently discussed in papers outlining the development of NKJP (Ogrodniczuk et al. [14]) or proposing the Diachronic Corpus of Polish (Król et al. [10]). Similarly, the first steps towards a common representation format for the planned diachronic corpus were recently completed in the Chronoflex project[3] aimed at providing a formal model of Polish inflection to represent historical changes in this area. There are also new developments in the area of open-source corpus search solutions, such as the MTAS-based corpus search engine[4], which has surpassed the capabilities of the Poliqarp engine (Janus and Przepiórkowski [7]) and was successfully deployed as the main search engine for the Corpus of the 19th century Polish[5], KORBA[6] and NKJP1M, the 1-million-token manually annotated subcorpus of NKJP[7].

All these attempts aimed at unifying existing corpus resources and tools into a common infrastructure intended for synchronic and diachronic research on the Polish language. In the next sections of this paper we elaborate on this concept, outlining plans for the development of a distributed corpus framework under the umbrella name of *The Polish Corpus Infrastructure;* Pol. *Polska Infrastruktura Korpusowa* – PIK. The framework is planned to create a unified platform for corpus-based studies of Polish and establish standards for the collection, processing and distribution of Polish corpus resources.

[1] https://clarin-pl.eu

[2] https://typo.uni-konstanz.de/parseme

[3] http://zil.ipipan.waw.pl/Chronofleks

[4] https://meertensinstituut.github.io/mtas/index.html

[5] http://korpus19.nlp.ipipan.waw.pl

[6] https://korba.edu.pl

[7] http://nkjp.nlp.ipipan.waw.pl

2 MOTIVATION FOR THE POLISH CORPUS INFRASTRUCTURE

The Polish Corpus Infrastructure is planned as a unique project playing a key role in further progress of research on the Polish language, both in linguistics (or more generally in the humanities) and the technology. The National Corpus of Polish was completed in 2011. It emerged as an effect of collaboration of four teams, which – prior to joining their efforts – had worked on their own corpora in the spirit of competition rather than cooperation. NKJP revealed the potential of synergy. The project which we describe in this article will cover a broader group of undertakings.

Although NKJP was one of the largest reference corpora available when it was compiled, it is a medium-sized corpus by modern standards. Moreover, the corpus is to some extent outdated, at least as a source of lexical data, which limits its applications in lexicography, but also in natural language processing: many proper and common names names vital to language processing (i.e. Emmanuel Macron, Donald Trump, Brexit, Instagram, fejk/fake, fanpage, or even selfie) are absent in NKJP or occur only in outdated contexts. Straightforward consequence of such a state of affairs is increasing error of statistic language models — for example speech recognition — only because they are based on outdated linguistic data. Finally, the spoken data do not meet modern requirements, e.g. the quality of recordings is often very low, in many cases it is impossible to consult the voice, not to mention research in phonetics.

Moreover, a number of corpus related initiatives have emerged since 2011: a number of historical corpora have been compiled, covering the 16th century (Institute for Literary Research, Polish Academy of Sciences), 1600–1770 (Institute of Polish Language, Polish Academy of Sciences), and 1820–1918 (University of Warsaw). Prior to NKJP a corpus of Medieval Polish was prepared at the Institute of Polish Language, Polish Academy of Sciences. These corpora cover a large portion of historical Polish, however, there remain some gaps. What is more important, having been compiled by diverse researchers, they are not entirely compatible. Recently, there have been attempts to unify such corpora and establish a common format of metadata and a tagset based on a unified theoretical approach (Król et al. [10]). Morphosyntactic taggers for post-medieval Polish have also been created (e.g. Waszczuk et al. [23]).

As for the development of spoken Polish corpora, the original datasets of spoken-conversational language included in NKJP were expanded, time-aligned with the original recordings and exposed though the Spokes search engine (Pęzik [16]). In 2019, a large corpus documenting the dialect of Spisz, comprising ca. 2 million running words of transcripts of spoken language was launched (Grochola-Szczepanek et al. [4]) and a corpus of Corpus of Polish Teenage Talk was compiled in 2014[8]. Despite these efforts, the level of representation of spoken Polish in the form of multimedia databases leaves much to be desired.

[8] `http://www.laboratoriumjezykowe.uw.edu.pl`

With regard to parallel corpora, major parallel corpora compiled after 2011 include a Polish-Russian[9], Polish-German, and Polish-English (Paralela; Pęzik [17]). Additionally a Polish component of the International Comparable Corpus was developed (Kirk et al. [9]).

Many of these projects employed a shared set of tools, e.g. a morphological analyzer Morfeusz (Woliński [24]) which is used in virtually all Polish corpora.

Perhaps most importantly, Polish is still one of the few large European languages with an outdated national corpus. There is little doubt that the proposed infrastructure could bridge a number of gaps in the availability and interoperability of corpus resources, thus advancing research on Polish — one of the biggest Slavic languages — in the forthcoming decades of the digital era.

3 THE OUTLINE OF THE INFRASTRUCTURE

The National Corpus of Polish will serve as the core of the proposed Polish Corpus Infrastructure and a representative and up-to-date resource whose main part covers present-day Polish starting from the year 1945. The updated corpus will have a large gender-, channel- and register-balanced component. Additionally, the reference corpus of modern Polish will be federated with various existing corpora of older Polish and Polish dialects, parallel corpora, and model training sub-corpora annotated on different semantic and syntactic levels. Federated collection of corpora will require locating individual parts of the infrastructure at individual branches of the consortium and at affiliated institutions.

One of the main goals of the infrastructure is to provide a proper level of representation of Polish for the purposes of linguistic research and language technologies in the era of big data. Reference corpora fulfilling this criterion require constant updating in order to efficiently contribute to the enhancement of linguistic technologies this facilitating the monitoring of current trends in lexical and syntactic change. For example, every five years the Institute of the Czech National Corpus issues a 100-million-word balanced sample comprising texts published during the most recent five-year period. At the same time, the corpus is supplemented with post-1990 journalistic texts, currently totalling one billion words. Similarly, distributional models of Polish, which are a basic resource in recent approaches to natural language processing can only perform optimally if they are based on a regularly updated reference corpus. A variety of language resources and technologies originally based on the 2011 edition of NKJP may soon become critically obsolete if the national corpus is not regularly updated. Also, the continuous emergence of neologisms, neosemanticisms and other aspects of language change dynamics may soon severely limit potential of NKJP in research in lexicography and linguistics.

[9] http://pol-ros.polon.uw.edu.pl

Apart from simply supplementing NKJP with new data, another important goal of the proposed infrastructure is to insure the proper quality of data collected. Although in the age of the Internet, acquiring large bodies of textual data is becoming increasingly easier, the proper sanitation, balancing and classification of such texts calls for a rigorous method. There are clear benefits of a systematic and controlled approach to continuous development of reference corpus resources, as opposed to their largely uncontrolled and thus biased compilation from ad hoc internet sources.

The planned infrastructure will also ensure interoperability of tools and enable their adjustment to various types of linguistic data.

4 IMPLEMENTATION SCOPE

The PIK infrastructure will deliver a number of technical outcomes in the form of data exchange standards, reference data sets, federated corpus search and monitoring services as well as advanced corpus exploration tools. The four major technical work packages planned in the project are described below.

1. The first major set of technical tasks will focus on implementing multimodal metadata and linguistic data exchange formats for the Polish language. It will involve developing a principled approach to storing texts, including historical and dialectal ones, with potentially rich bibliographic, sociolinguistic, morphosyntactic, syntactic, semantic and discursive annotations, as well as methods for multimedia data representation: aligning texts with their sources (spoken data, video, scans) and methods of searching and managing source corpus files. This step is also necessary for integrating different formats developed within current corpus projects which will feed into the new infrastructure.
2. The second area of technical work will involve extending the balanced segment of the National Corpus of Polish with contemporary texts published after 2011. Mechanisms of continuous acquisition of densely sampled web-based corpus data will be created and deployed to monitor regular fluctuations in lexical frequencies and long-term dynamics of language change. Continuous acquisition of spoken data will have to be addressed as a separate challenge.
3. A separate work package will be aimed at establishing a federation of Polish corpora in order to provide programmatic access to existing third-party contemporary, diachronic, dialectal, spoken and parallel corpora. Mechanisms for simultaneous federated search will be implemented with special consideration of user interface experience and programmatic access, facilitating the use of the corpus infrastructure for both researchers and non-specialist users.

4. The range of dedicated tools for exploring and analyzing the core resources will be considerably larger than the federated search functionalities. The fourth technical work package of the planned infrastructure will provide corpus as well as search and analysis tools for exploring phraseology, differences in stylistic distribution, generating frequency lists, user-defined filtering of concordances, text profile analysis and creation of virtual collections from the index of reference corpus data.

5 AVAILABILITY OF THE PROPOSED INFRASTRUCTURE

In general, the proposed corpus infrastructure will be widely available for scientific research and technological applications. At the same time one should be aware of the copyright limitations on the distributability of the texts included in corpora. A number of infrastructure access models and scenarios are therefore planned to deal with this problem:

- *Access for end-users by Internet tools & services*. End-users will be able to use infrastructure through Web-based applications, i.e. corpus browsing systems and multimedia databases. This model of access will work particularly well for researchers in linguistics and humanities, because their requirements are fairly predictable.
- *Remote access for programmers*. Web applications will be accessible as programming interfaces (API) to facilitate large scale processing of data and development of client applications by the user community.
- *Full access to annotated subcorpora of samples*. Manually annotated subcorpora and training resources/data, indispensable for further progress of language processing, will be accessible in full under open-access licenses. A good example of such a resource is the 1-million-token subcorpus of NKJP which has been widely used in many NLP systems/applications for Polish.
- *Full access to public domain resources*. Whenever possible, resources acquired from public domain repositories, together with manual and automatic annotation will be fully available for offline use.
- *Full access to statistical and distributional models and other derivatives*. Statistical and distributional models, which cannot be used to reconstruct source texts, will be accessible under open licenses. An example of such data are n-gram models of language used in speech recognition systems or vector representations of words (i.e. word embeddings) used currently in many tasks in NLP, text classification, etc.
- *Custom-made models*. Researchers with special requirement will be able to order custom-made statistical models such as distributional language models computed with special tools and parameters. The operators of the infrastructure will make such models available under open licenses to other users.

We believe that the creation of the proposed corpus infrastructure will enable two-way cooperation with users and researchers. On the one hand, it will allow simple and effective inclusion of collections created for special purposes such as corpora of students' and teenagers' language, writers' idiolects, learner language and translation. On the other, it will be used through dedicated tools in research and teaching.

6 CONCLUSIONS

To sum up, it has to be stated clearly that Polish needs a large, balanced, representative national corpus. An up-to-date reference corpus is an indispensable resource for any modern language. The original National Corpus of Polish was innovative and it was even considered as an exemplary reference corpus for other languages in 2011, but for the last five years we have been getting more and more questions from our colleagues abroad about the current state of NKJP. Even a chronological update of the original corpus with samples of registers (dialects, registers, teen talk) and parallel sub-corpora would not be enough for today's challenges. The National Corpus of Polish which is truly a part of the European corpus landscape should be characterized by unrestricted availability for scientific research and innovative technical applications.

Our project addresses these expectations, but it needs funding to be realized. Two institutes of the Polish Academy of Sciences (Institute of Computer Science and Institute of Polish Language), University of Łódź and University of Warsaw have submitted an application to include the Polish Corpora Infrastructure in the The Polish Roadmap for Research Infrastructures developed by the Polish Ministry of Science and Higher Education. The inclusion of PIK in this framework would create an opportunity to obtain permanent funding which is a *sine qua non* of the initiative in the form described in this paper.

Even though it goes without saying that the development of the infrastructure will pose several scientific and technical challenges, a few of them are worth stating. The most important one seems to be related to the federated data model, requiring comparison of data obtained from many dispersed data resources and its proper evaluation, which is a research problem in its own right. Another problem is related to building complex language models adequate for highly inflectional languages. One more challenge of this kind is building a federal access system to the corpus infrastructure that meets the requirements of security and efficiency of data access. Last but not least, we envisage a logistic challenge resulting from the need to obtain consent from the copyright holders of new data, which requires convincing them that the project will bring substantial benefits to their businesses. Moreover, due to the concentration on the media market a lack of a consent of a large market player causes a significant loss of available texts. Acquiring spoken texts, which are very important for linguistic research, is an expensive and logistically complex undertaking.

However significant they may seem, these challenges must be overcome. Corpora have become a basic resource for linguistic research and language technology development. A language without an up-to-date reference corpus has limited perspectives for consideration in international research projects and language technology enterprises. It is high time that we released a new edition of the National Corpus of Polish with its full infrastructure to the public.

References

[1] Czerepowicka M. (2014). SEJF – Słownik elektroniczny jednostek frazeologicznych. Język Polski XCIV (2), pages 116–129.

[2] Čermák, F. (1997). Czech National Corpus: A case in many contexts. International Journal of Corpus Linguistics 2 (2), pages 181–197.

[3] Derwojedowa M., Kieraś W., Skowrońska D., and Wołosz R. (2014). Korpus polszczyzny XIX wieku — od mikrokorpusu do korpusu średniej wielkości. Prace Filologiczne LXV, pages 251–256.

[4] Grochola-Szczepanek H., Górski R. L., von Waldenfels R., and Woźniak M. (2019). Korpus języka mówionego mieszkańców Spisza. LingVaria LV (1), pages 165–180.

[5] Gruszczyński W., Adamiec D., and Ogrodniczuk M. (2013). Elektroniczny korpus tekstów polskich z XVII i XVIII w. (do 1772 r.) Polonica XXXIII, pages 311–318.

[6] Hajnicz E., Patejuk A., Przepiórkowski A., and Woliński M. (2016). Walenty: słownik walencyjny języka polskiego z bogatym komponentem frazeologicznym. In K. Skwarska and E. Kaczmarska (eds.) Výzkum slovesné valence ve slovanských zemích, pages 71–102. Prague, Czech Republic, Slovanský ústav AV ČR.

[7] Janus D., and Przepiórkowski A. (2007). Poliqarp: An open source corpus indexer and search engine with syntactic extensions. In Proceedings of the ACL 2007 Demo and Poster Sessions, pages 85–88, Prague, Czech Republic.

[8] Kieraś W., and Woliński M. (2018). Manually annotated corpus of Polish texts published between 1830 and 1918. In N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, and T. Tokunaga (eds.) Proceedings of the 11th International Conference on Language Resources and Evaluation (LREC 2018), pages 3854–3859, Paris, France: European Language Resources Association.

[9] Kirk J., Čermáková A., Ebeling S. O., Ebeling J., Kren M., Aijmer K., Benko V., Garabík R., Górski R. L., Jantunen J., Kupietz M., Simkova M., Schmidt T., and Wicher O. (2018). Introducing the International Comparable Corpus. In S. Granger, M–A. Lefer and L. Aguiar de Souza Penha Marion (eds.) Book of Abstracts: Using Corpora in Contrastive and Translation Studies Conference (5th edition). CECL Papers, Louvain-la-Neuve.

[10] Król M., Derwojedowa M., Górski R. L., Gruszczyński W., Opaliński K. W., Potoniec P., Woliński M., Kieraś W., and Eder M. (2019). Narodowy Korpus Diachroniczny Polszczyzny. Projekt. Język Polski XCXIX (1), pages 92–101.

[11] Łaziński M. (2018). Nowe zjawiska w języku młodzieży. Gramatyka slangu. In B. Pędzich, M. Wanot-Miśtura, and D. Zdunkiewicz-Jedynak (eds.) Tyle się we mnie słów zebrało. Szkice o języku i tekstach, pages 339–356. Warsaw, Poland.

[12] Mykowiecka A., Marciniak M., and Rychlik P. (2017). Testing word embeddings for Polish. Cognitive Studies / Études Cognitives 17, pages 1–19.

[13] Ogrodniczuk M., Głowińska K., Kopeć M., Savary A., and Zawisławska M. (2013). Polish Coreference Corpus. In Z. Vetulani (ed.), Proceedings of the 6th Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics, pages 494–498, Poznań, Poland: Wydawnictwo Poznańskie, Fundacja Uniwersytetu im. Adama Mickiewicza.

[14] Ogrodniczuk M., Derwojedowa M., Łaziński M., and Pęzik P. (2017). Narodowy Korpus Języka Polskiego – co dalej? Prace Filologiczne, LXXI, pages 237–245.

[15] Pęzik P. (2014). Graph-Based Analysis of Collocational Profiles. In V. Jesenšek and P. Grzybek (eds.) Phraseologie Im Wörterbuch Und Korpus (Phraseology in Dictionaries and Corpora), pages 227–243. ZORA 97. Maribor.

[16] Pęzik P. (2015). Spokes – a Search and Exploration Service for Conversational Corpus Data. In Selected Papers from CLARIN 2014, pages 99–109. Linköping Electronic Conference Proceedings. Linköping University Electronic Press.

[17] Pęzik P. (2016). Exploring Phraseological Equivalence with Paralela. In Polish-Language Parallel Corpora, edited by Ewa Gruszczyńska and Agnieszka Leńko-Szymańska, pages 67–81. Warsaw, Instytut Lingwistyki Stosowanej UW.

[18] Pęzik P. (forthcoming, 2019). Budowa i zastosowania korpusu monitorującego MoncoPL. Forum Lingwistyczne.

[19] Przepiórkowski A., Bańko M., Górski R. L., and Lewandowska-Tomaszczyk B. (eds.) (2012). Narodowy Korpus Języka Polskiego. Warsaw, Wydawnictwo Naukowe PWN.

[20] Riegel M., Wierzba M., Wypych M., Żurawski Ł., Jednoróg K., Grabowska A., and Marchewka A. (2015). Nencki Affective Word List (NAWL): The Cultural Adaptation of the Berlin Affective Word List–Reloaded (BAWL-R) for Polish. Behavior Research Methods 47(4), pages 1222–1236.

[21] Twardzik W., and Górski R. L. (2003). Korpus staropolski Instytutu Języka Polskiego PAN w Krakowie. In S. Gajda (ed.) Językoznawstwo w Polsce. Stan i perspektywy, pages 155–157.

[22] Waszczuk J. (2012). Harnessing the CRF complexity with domain-specific constraints: The case of morphosyntactic tagging of a highly inflected language. In Proceedings of COLING 2012, pages 2789–2804. Mumbai, India.

[23] Waszczuk J., Kieraś W., and Woliński M. (2018). Morphosyntactic disambiguation and segmentation for historical Polish with graph-based conditional random fields. In P. Sojka, A. Horák, I. Kopeček, and K. Pala (eds.) Proceedings of the 21st Text, Speech, and Dialogue International Conference (TSD 2018), Brno, Czech Republic. Lecture Notes in Artificial Intelligence 11107, pages 188–196. Springer-Verlag.

[24] Woliński M. (2014). Morfeusz reloaded. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis (eds.) Proceedings of the 9th International Conference on Language Resources and Evaluation (LREC 2014), pages 1106–1111, Reykjavík, Iceland: European Language Resources Association.

[25] Wróblewska A. (2012). Polish dependency bank. Linguistic Issues in Language Technology 7 (2), pages 1–18.

[26] Żmigrodzki P., Bańko M., Batko-Tokarz B., Bobrowski J., Czelakowska A., Grochowski M., Przybylska R., Waniakowa J., and Węgrzynek K. (eds.) (2018). Wielki słownik języka polskiego PAN. Geneza, koncepcja, zasady opracowania. Kraków, Instytut Języka Polskiego PAN/LIBRON, 264 p.