Enhanced and extended Electronic Corpus of 17th- and 18th-century Polish Texts

Renata Bronikowska, Ewa Rodek & Aleksandra Wieczorek
(Institute of Polish Language Polish Academy of Sciences)

Keywords: historical text corpus, Baroque, Enlightenment, NLP tools, connecting resources

We will present a new version of the Electronic Corpus of the 17th- and 18th-century Polish Texts (shortly referred to as KorBa; www.korba.edu.pl), twice as large as the one available until recently (Gruszczyński *et al.* 2022). The corpus, initially including texts from the Baroque period, was expanded to include texts from the late 18th century belonging to the Enlightenment period. Since these cultural trends have left a clear mark on the language, the new version of KorBa comprises two subcorpora that can be searched separately: Baroque (1601-1740) and Enlightenment (1741-1800). New texts from the 17th and early 18th centuries have also been added to the corpus, selected to ensure greater chronological, geographical, genre and thematic balance. In total, the new KorBa contains nearly 27 million tokens from more than 2,000 texts. An experimental syntactically annotated corpus, consisting of 1,000 sentences has also been developed.

The new version of KorBa has been built using two new tools based on neural networks: a transcriber, designed for automatic transforming transliterated text into modern spelling, and a tagging system KFTT comprising a tokenizer, a morphosyntactic tagger and a lemmatizer (Wróbel 2020). Thanks to its neural architecture, KFTT can handle less common tokenization and spelling found in historical texts with high accuracy. The use of modern technologies allowed to reduce the number of errors occurring during data processing.

Our additional goal was to integrate four sources for research on the Polish language of the 17th and 18th centuries: KorBa, the Electronic Dictionary of the 17th- and 18th-century Polish (e-SXVII), the Digital Library of Polish and Poland-Related News Pamphlets from the 16th to the 18th Century (CBDU) and the Card-index of the Dictionary of the Polish Language of the 17th and First Half of the 18th Century (KXVII) (Bilińska-Brynk, Rodek 2020). For this purpose, the dedicated website Polish Language of the 17th and 18th Centuries (https://polszczyzna17-18.ijp.pan.pl) has been developed, which allows the simultaneous searching of these resources. In addition, connections between individual resources aimed at special purposes have been created (Ogrodniczuk, Gruszczyński 2019, Wieczorek 2021). For instance the connections between the KorBa and e-SXVII websites make it easier for dictionary editors to use the corpus. The links between CBDU and e-SXVII allow explaining archaic words appearing in CBDU texts by referring to the appropriate e-SXVII entries. All these connections are dynamic, which means that each time data is downloaded from the current database of individual resources.

References:

Bilińska-Brynk, Joanna, Rodek, Ewa (2020), Paper Quotation Slips to the Electronic Dictionary of the 17th- and 18th-Century Polish – Digital Index and its Integration with the Dictionary, in Z. Gavriilidou, M. Mitsiaki, and A. Fliatouras (eds), (2020), *Proceedings of the XIX EURALEX Congress: Lexicography for Inclusion*, vol. I, Komotini: Democritus University of Thrace, 465–470.

https://euralex.org/elx_proceedings/Euralex2020-2021/EURALEX2020-2021_ProceedingsBook-Vol1.pdf

Gruszczyński Włodzimierz, Adamiec Dorota, Bronikowska Renata, Kieraś Witold, Modrzejewski Emanuel, Wieczorek Aleksandra, Woliński Marcin (2022), The Electronic Corpus of 17th- and 18th-century Polish Texts, *Language Resources and Evaluation* 56, 309–332. https://link.springer.com/article/10.1007/s10579-021-09549-1

Ogrodniczuk Maciej, Gruszczyński Włodzimierz (2019), Connecting Data for Digital Libraries: The Library, the Dictionary and the Corpus, in A. Jatowt, A. Maeda, and S. Syn (eds), (2019), *Digital Libraries at the Crossroads of Digital Information for the Future. ICADL 2019. Lecture Notes in Computer Science*, vol. 11853, Springer, Cham, 125–138. https://link.springer.com/book/10.1007/978-3-030-34058-2

Wieczorek, Aleksandra (2021), Integracja Elektronicznego słownika języka polskiego XVII i XVIII wieku i Elektronicznego Korpusu Tekstów Polskich z XVII i XVIII Wieku okiem użytkownika i redaktora, in E. Horyń, E. Młynarczyk, and P. Żmigrodzki (eds), (2021), *Język polski – między tradycją a współczesnością. Księga jubileuszowa z okazji stulecia Towarzystwa Miłośników Języka Polskiego*, Kraków: Wydawnictwo Naukowe Uniwersytetu Pedagogicznego, 547–560. https://rep.up.krakow.pl/xmlui/bitstream/handle/11716/10824/PM1030--Jezyk-polski.pdf?sequence=1&isAllowed=y

Wróbel, Krzysztof (2020), KFTT: Polish full neural morphosyntactic tagger, in M. Ogrodniczuk, and Ł. Kobyliński (eds), (2020), *Proceedings of the PolEval 2020 Workshop*, Warszawa: Institute of Computer Sciences, Polish Academy of Sciences, 47–53. http://poleval.pl/files/poleval2020.pdf