Using Digital Tools to Create Modern Multi-Search Engine for Polish Historical Dictionaries

Rodek, Ewa
ewa.rodek@ijp.pan.pl
Institute of Polish Language Polish Academy of Sciences, Poland

Thanks to the increasing resources of digital libraries, diachronic linguists have access to historical lexicographic works. However, due to many factors, including the dispersion of sources and erroneous bibliographic metadata making it difficult to find the most appropriate edition of dictionary for research, the vocabulary recorded in old dictionaries has not been fully revealed. Above all, the lack of a searchable text layer for image files of the aforementioned dictionaries means that the material cannot be conveniently and optimally used by researchers.

Therefore, seeing the need to digitise Polish historical lexicographic works our team of researchers have prepared a project for the deep digitisation of the three most important dictionaries with Polish material for the Middle Polish era: *Thesaurus latino-polono-graecus* (1643) by Grzegorz Knapiusz (https://crispa.uw.edu.pl/object/files/416379/display/Default), *Nowy dykcjonarz to jest mownik polsko-niemiecko-francuski* (1764) by Michał A. Trotz (https://www.dbc.wroc.pl/dlibra/doccontent?id=7138) and *Forytarz języka polskiego* (1674) by Jan Ernesti (online version not available). These works are ground-breaking for this period in the history of the Polish language, with the most mature work and complex structure, and were a point of reference for many later imitators.

This deep digitisation (also called digitization of the third level) will consist of several stages. The first stage of deep digitisation will be the automatic recognition of old prints using the HTR method. This will require specially trained models adapted to multilingual material, including text printed in Latin and Greek fonts. The next step will be xml tagging of the dictionary micro- and macrostructure in the TEI format. The final stage of digitisation work will be morphosyntactic tagging of the Polish vocabulary contained therein. This proceeding will enable data standardisation and obtaining database files, making it possible to search individual headwords as well as fragments of headword articles in the future. On this material we will develop a multi-search engine suitable for research in lexicography, pragmatics or translation studies. We would like this to be the beginning of a larger database of historical Polish lexicons, to which more dictionaries can be added. This also fits in with the concept of a dictionary understood as a corpus (Żmigrodzki 2005).

Moreover, the creation of an API for this service will allow its integration with various resources. As a trial experimental application of this API the project will involve connecting the created lexicon database with the Electronic Dictionary of the Polish Language of the 17th and 18th Centuries (e-SXVII, sxvii.pl, Bronikowska et al. 2020). The integration will include the creation of a list of Polish lexemes contained in the digitised material, which will be used to complete the e-SXVII index. Also we will create a connection with the editorial panel of this dictionary to give the editors the opportunity to supplement e-SXVII with information from digitised sources. We would also like to try out the automatic transfer of some of the entries from the digitised dictionaries to e-SXVII. Other resources with which such a lexical database can be linked are The Electronic Corpus of 17th- and 18th-century Polish Texts (korba.edu.pl) or eLexicon of the Polish Medieval Latin (elexicon.scriptores.pl).

Bibliography

Bronikowska, Renata / Majdak Magdalena / Wieczorek, Aleksandra / Żółtak, Mateusz (2020): “ The Electronic Dictionary of the 17th- and 18th-century Polish - towards the open formula asset of the historical vocabulary ”, in: Gavriilidou, Zoe / Mitsiaki, Maria / Fliatouras, Asimakis (ed.): *Proceedings of the XIX EURALEX Congress: Lexicography for Inclusion* , vol. I, Democritus University of Thrace: 471-475.

Ogrodniczuk, Maciej / Gruszczyński, Włodzimierz (2019): “ Connecting Data for Digital Libraries: The Library, the Dictionary and the Corpus ”, in: Jatowt, Adam / Maeda, Akira / Syn, Sue Yeon (ed.): *Digital Libraries at the Crossroads of Digital Information for the Future. ICADL 2019. Lecture Notes in Computer Science* , vol. 11853. Springer, Cham: 125-138.

Żmigrodzki, Piotr (2005): „Słownik jako korpus tekstów – korpus tekstów jako s#ownik. Perspektywy polskiej leksykografii naukowej”, in: *Poradnik Językowy* 6: 3-14.