The use of DIY corpus in the research on Anglicisms in specialized Polish

Marcin Zabawa
(University of Silesia in Katowice)

Keywords: DIY corpus, specialized corpus, corpus size, data selection, borrowings.

The aim of the paper is to discuss the use of one's own corpus (do-it-yourself corpus) in the research on Anglicisms. To be more specific, the paper is set in the English-Polish context and will explore the usefulness of DIY corpora in the analysis of English lexical borrowings in specialized Polish (on the example of the semantic area of computers and the Internet). While general corpora of Polish do exist, viz. NKJP (Przepiórkowski et al. 2012) and MoncoPL (Pęzik 2020), they are not very useful for the study of English loans in specialized Polish. Three problems seem to be predominant: the lack of specialized texts in the corpus, the lack of access to full texts and the lack of knowledge of forms that would act as queries. Both NKJP and MoncoPL can be very useful in the lexical research provided the form of the word is known. In the study on newest lexical borrowings, the forms are, by definition, not known beforehand. Therefore, the first step, i.e. identification of Anglicisms, appears the most demanding. There were some (only partially successful) attempts at an automated extraction of lexical Anglicisms in some Romance and Germanic languages (Mańczak-Wohlfeld and Witalisz 2019), but not (yet) in Polish that I am aware of.

The situation is even more problematic in the case of specialized Polish. The general corpora, as was explained above, are not very useful; consequently, the best solution is to build one's own corpus. One of the main advantages of DIY corpus is the access to full texts (indispensable for loan word identification and very useful for contextual study). The main part of the paper will thus be devoted to the description of my own corpus of specialized Polish, composed of short Internet texts (entries) collected from selected Internet forums on computers and the Internet. The following points will be discussed at some length: the process of data selection (based on a size of a given forum), corpus compilation (the corpus has been manually created, i.e. without the use of crawlers), corpus size (the corpus consists of approximately 1,500,000 running words), and the corpus use for the analysis of English lexical loanwords in the Polish semantic field of computers and the Internet (the loanwords have been manually identified, but the possible use of *Korpusomat.eu* (Saputa et al. 2023), i.e. an online tool for corpora creation and analysis, will be addressed as well). In the corpus, approximately 583 types of English lexical loans have been identified (45,239 tokens); detailed results will be presented in the talk. In addition, other areas, both general (e.g. optimal size of DIY corpus, lexical saturation) and research-specific (e.g. problems with variant spellings of loanwords) will be discussed.

References

Mańczak-Wohlfeld, Elżbieta, and Alicja Witalisz (2019), Anglicisms in The National Corpus of Polish: Assets and Limitations of Corpus Tools, *Studies in Polish Linguistics* 14/4, 171–190, doi: 10.4467/23005920SPL.19.019.11337.

Pęzik, Piotr (2020), Budowa i zastosowania korpusu monitorującego MoncoPL [The structure and use of MoncoPL monitor corpus], *Forum Lingwistyczne* 7, 133–150, doi: 10.31261/fl.2020.07.11.

Przepiórkowski, Adam, Mirosław Bańko, Rafał L. Górski, and Barbara Lewandowska-Tomaszczyk (2012), *Narodowy Korpus Języka Polskiego* [National Corpus of Polish], Warszawa: Wydawnictwo Naukowe PWN, https://nkjp.pl/settings/papers/NKJP_ksiazka.pdf.

Saputa, Karol, Aleksandra Tomaszewska, Natalia Zawadzka-Paluektau, Witold Kieraś, and Łukasz Kobyliński (2023), Korpusomat.eu: A multilingual platform for building and analysing linguistic corpora, in J. Mikyška, C. de Mulatier, M. Paszynski, V. V. Krzhizhanovskaya, J. J. Dongarra, and P. M. A. Sloot (eds), (2023), *Computational Science – ICCS 2023. 23rd International Conference, Prague, Czech Republic, July 3–5, 2023, Proceedings, Part II*, Cham: Springer, 230–237.