EURALEX XIX

Congress of the European Association for Lexicography

Lexicography for inclusion

7-11 September 2021

Ramada Plaza Thraki
Alexandroupolis, Greece

www.euralex2020.gr

Proceedings Book
Volume 1

Edited by Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

EURALEX Proceedings

ISSN 2521-7100

ISBN 978-618-85138-1-5

Edited by: Zoe Gavriilidou, Maria Mitsiaki, Asimakis Fliatouras

English Language Proofreading: Lydia Mitits and Spyridon Kiosses

Technical Editor: Kyriakos Zagliveris

2020 Edition

The Electronic Dictionary of the 17th- and 18th-century Polish - Towards the Open Formula Asset of the Historical Vocabulary

Bronikowska R.[1], Majdak M.[1], Wieczorek A.[1], Żółtak M.[2]

[1] *Institute of Polish Language, Polish Academy of Sciences*
[2] *Austrian Centre for Digital Humanities and Cultural Heritage*
renata.bronikowska@ijp.pan.pl, magdalena.majdak@ijp.pan.pl, aleksandra.wieczorek@ijp.pan.pl, mateusz.zoltak@oeaw.ac.at

Abstract

The paper discusses the *Electronic Dictionary of the 17th- and 18th-century Polish* (abbreviated e-SXVII), an important resource for the study of language, history and culture of the period. After several dozen years of gathering material, conceptual work, and after the publication of five fascicles, the print dictionary project was discontinued in favour of a digital version. The work has since accelerated significantly, although development is still ongoing. The paper focuses on new aspects of the methodology stemming from the open form of the dictionary. The innovation offers significant benefits both for the editors and the users. This formula also allows for e-SXVII to be integrated with other electronic language resources, like corpora and digital libraries – a feature currently under intense development.

Keywords: electronic dictionaries; integration of linguistic resources; Middle Polish; historical vocabulary

1 Introduction

The 17th and 18th centuries were an important period in the development of the Polish language, when several notable grammatical categories took shape, while others disappeared; vocabulary borrowed greatly from other languages (including Latin, German, French, Turkish and others), whereas style and syntax continued to evolve. Texts published in that period constitute an important source for research into history, culture, literature, history of science, and other fields, and their complete comprehension requires a dictionary. The *Electronic Dictionary of the 17th- and 18th-century Polish* (*Elektroniczny słownik języka polskiego XVII i XVIII wieku*; hereafter e-SXVII)[1] is the first lexicon in Poland to focus on these two centuries and the first ever purpose-built electronic dictionary of Polish.[2] The decision to develop and publish the dictionary entirely in electronic form had significant impact both on the development work, including many solutions within the subject matter, and on its availability and ease-of-access. This approach has the advantage of enabling integration with other electronic resources for historical and modern Polish, such as corpora and digital libraries.

2 Polish Dictionaries Noting Historical Vocabulary

The vocabulary of past periods in the history of the Polish language is presented in several large lexicographic works (printed, then digitised). The lexical layer of the oldest works (until 1500) is registered in the *Dictionary of Old Polish* (*Słownik staropolski*, SStp). The vocabulary of 16th-century works is described in the still-unfinished *Dictionary of 16th-century Polish* (*Słownik polszczyzny XVI w.*, SXVI). The *Dictionary of Polish Language*, edited by W. Doroszewski (*Słownik języka polskiego*, SJPDor) in the second half of the 20th century, was intended mostly as a dictionary of modern Polish, yet it reaches back to mid-18th century. This paper discusses a dictionary intended to bridge the gap and augment the documentation of historical vocabulary with resources for the 17th and 18th centuries.

3 Evolution of Dictionary Development – the Print to Digital Transition[3]

Work on a dictionary of 17th- and 18th-century Polish[4] started in 1954. Its result was a card index of ca. 2.8 million paper slips listing use cases for ca. 80 thousand lexemes excerpted from 275 texts (hereafter referred to as KXVII).[5] The list of sources is supposed to be as representative as feasible and include not only literary works (in the 17th and 18th centuries written mostly in verse), but also functional texts, such as administrative files, inventories, specialist manuals, dictionaries, phrasebooks, personal documents, letters, and others. The first volume of the *Dictionary of the Polish*

[1] Links for electronic resources mentioned in the paper can be found in the References section at the end.
[2] This is understood as a dictionary that was meant to appear in an electronic version from the very beginning and as such has developed new, more adequate methodology. Digital forms of print dictionaries of Polish (especially modern Polish) have been available before. Although e-SXVII is a continuation of a traditional dictionary (i.e. one made with print in mind), its concept and form are now so different from the original as to consider it a separate work (more on this subject later on).
[3] This paper focuses on the description of the electronic version of the dictionary of 17th- and 18th-century Polish. The history of the print version can be found in the works of Siekierska (1998), Majdak (2012, 2018), and others.
[4] Originally the dictionary was intended to cover the 17th century and the first half of the 18th century.
[5] This archive was digitised in 2011 and is now available online (see KXVII).

Language of the 17th and First Half of the 18th Century (*Słownik języka polskiego XVII i pierwszej połowy XVIII wieku*, SXVII) appeared in print as five fascicles, published between 1999 and 2004. Predictions at the time indicated that if that pace was maintained, the work would be completed in approximately 100 years (per Gruszczyński 2005: 48). The search for ways to expedite development prompted the 2004 decision to move to a digital form. In light of its evolution, this new version of the dictionary was renamed the *Electronic Dictionary of the 17th- and 18th-century Polish* (*Elektroniczny słownik języka polskiego XVII i XVIII wieku*, e-SXVII).
The changes made in 2004 have to be considered revolutionary for the era, even if today they appear natural, necessary, and obvious. An internal network connected to the internet was created, the usage of computers as just text editors was abandoned in favour of working within a dictionary editor's platform, and a webpage was developed for the dictionary (cf. Majdak 2018: 178). Moreover, the postulated conversion of as many sources as possible into electronic form also started to become reality, giving rise to the *Electronic Corpus of 17th- and 18th-century Polish Texts (up to 1772)* (*Elektroniczny Korpus Tekstów Polskich z XVII i XVIII w. (do 1772 roku)*, KorBa; see p. 6).

4 The e-SXVII IT System

The dictionary authors resolved to offer e-SXVII to the users in the form of a webpage. In order to simplify the architecture of the IT system, the dictionary editor's interface was also developed in the form of a web application.
As for the general concept, the IT system used in the dictionary is simply a content management system (CMS) with dictionary entries in place of posts. A characteristic feature of the e-SXVII system is very precise modelling of the complex structure of a dictionary entry, both in the dictionary editor's interface and within the relational database storing the dictionary data. This solution not only forces the dictionary editors into a unified system of data entry, but also ensures clean separation between the data and the presentation. This, in turn, permits for quick changes to the way e-SXVII is presented to the user (as only the display template is changed, without having to alter the content of the entries), as well as very precise content searches or easy integration with other computer systems. Since frameworks for content management systems with such complex data structures were still in their infancy in 2004[6], the e-SXVII CMS was written from scratch in PHP, with Postgresql as the relational data base.

5 Changes to Dictionary Editor Workflow

The present form of the dictionary uses more advanced techniques both in its development and in resolving search queries. The open formula permits constant improvements to the lexicographic description and removes the need for printing errata and supplements, while online presentation removes limitations on illustrative material.
A very important feature for a dictionary editor is the ability to alter existing entries. In particular, this raises a possibility of making changes to the methodological principles of the lexicographic description (where applicable), while preserving the macrostructural integrity of the dictionary.
A good example here are the entries for numerals. Originally, they were separated into several grammatical categories on a largely semantic basis, as is the traditional Polish practice. However, electronic resources and tools, as well as newer print works, are abandoning this classification in favour of one based on the grammatical properties of numerals. Therefore, a decision was recently made to introduce changes to categories assigned to numerals in the dictionary: they are now a single part of speech with additional descriptors listing inflectional categories typical to the specific word (e.g. the numeral *dwa* "two" inflects for case and gender, while *drugi* "second" inflects for case, gender, and number). These changes have affected over twenty extant entries.

6 User Amenities

The open form of the dictionary allows for the publication of material that is still undergoing development. The users also gain new features they can use while searching for entries or their elements. At the same time, they must accept the possibility that the dictionary editors will introduce changes to finished entries or the dictionary structure as a whole.

6.1 Access to Unfinished Entries

The dictionary development team distinguishes three development stages of an entry: "stub" (only contains basic information about the grammar and perhaps a single quotation); "entry in preparation," and "entry fully developed" (approved by editor in chief). The users can access an entry at any point of its development. The entry view displays information about its status and date of the most recent modification.

6.2 Active Elements of the Entry

Some elements of an entry function as hyperlinks to other entries, allowing the user to easily find dictionary information supplementing what is shown in the entry they are viewing. For example, some definitions include words that are currently archaic and no longer understood, and so they feature links to definitions of those words (e.g. the definition of the word *arkabuźnik* "heavy cavalryman armed with an → arquebus" includes a link to the entry for *arkabuz* "arquebus").

6.3 User-friendly Entry View

An entry appears in a truncated version by default, only displaying the entry name, abbreviated grammatical notes and phonetic variants (if any). Further sections of the entry, such as "Grammatical forms," "Etymology," or "Meanings,"

[6] The oldest frameworks of that type to remain popular to this day, such as Django or Ruby on Rails, were not created until 2005.

appear as drop-down boxes (see Fig. 1). The user may choose to view any of them in more detail (see Fig. 2). By hovering the cursor over an abbreviation of a given source, the user may see its full citation. It is also possible to display an entry in a printable version that resembles a traditional dictionary entry.

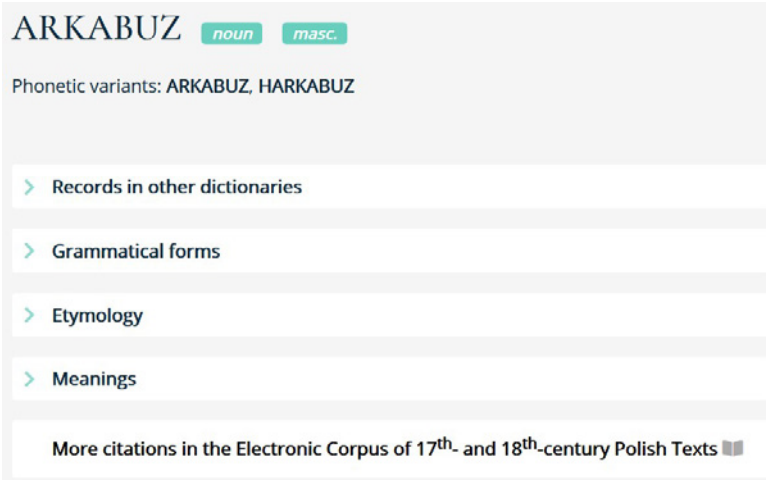

Figure 1: Truncated form of an entry.

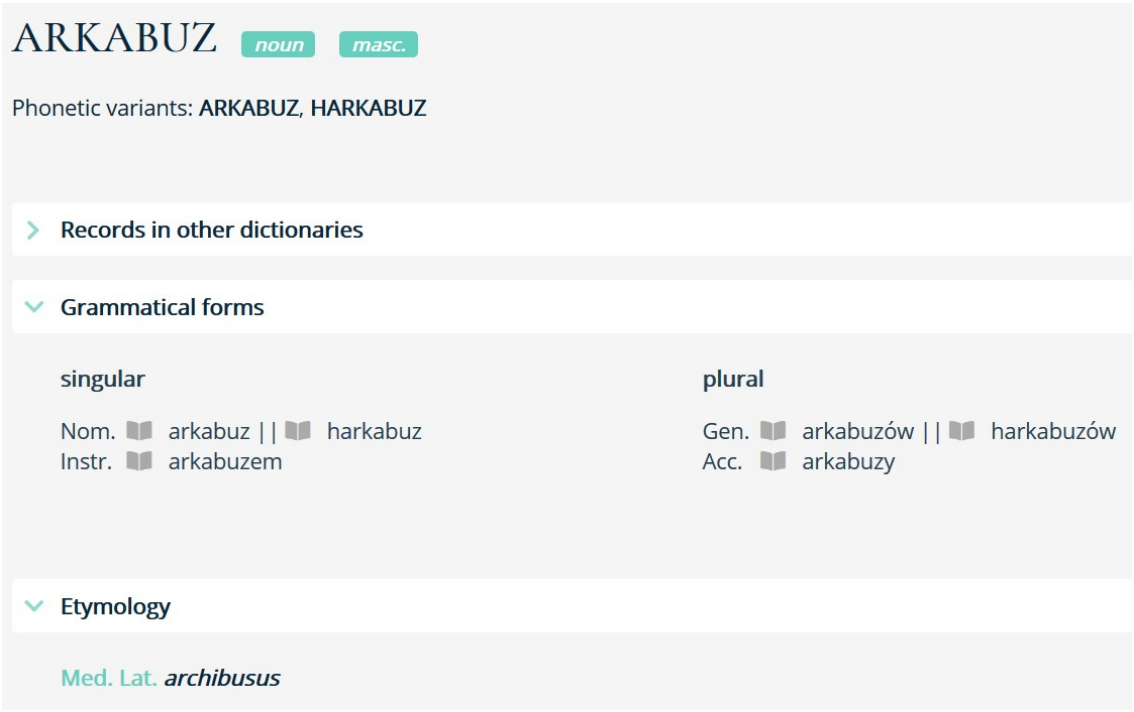

Figure 2: An entry with unrolled drop-downs for “Grammatical forms” and “Etymology”.

6.4 Search Features

The dictionary search engine enables the viewer to find not only the entry, but also all of its phonetic variants, placed on the list of results alongside the entries themselves. Thus, it will return the entry for *arkabuz* even if the *harkabuz* “harquebus” form is typed into the search box. Subentries are searched alongside the main entries. The search engine also provides more advanced queries, and therefore various ways of filtering the entries. For example, a user can search for entries within a particular part of speech, with a given etymology, or a specific qualifier. It is also possible to search for entries with specific grammatical forms (e.g. dual number forms, no longer present in modern Polish).

6.5 Queries

The “Queries” function provides additional ways of searching the entries. It enables the user to access quotations of interest to them without having to open an entry. In this way, one may search for quotes from a specific source or ones that illustrate specific word combinations, set phrases, proverbs, or metaphors. A link underneath a quotation allows the user to quickly access the entry featuring it.

7 Integration with Other Resources

Several other electronic resources for 17th- and 18th-century texts began their development concurrently with e-SXVII. Originally, they functioned independently from the e-SXVII computer system; however, integrating those resources eventually became the natural next step in the development of the dictionary.

The most important resource created with the aim (although not exclusive) to improve the work of e-SXVII lexicographers is the *Electronic Corpus of 17th- and 18th-century Polish Texts* (hereafter KorBa).[7] The corpus, available online, currently includes 13.5 million tokens (and is expected to reach 25 million by 2023). The texts contained within it, presented as transliterations and transcriptions, have been annotated structurally (marked up for document structure), linguistically (marked up for foreign fragments), and morphosyntactically (lemmatised, marked up for parts of speech, and denoted for values of appropriate grammatical categories).[8] Each text was also provided with extensive metadata. The corpus is searched through the MTAS search engine (Brouwer, Brugman & Kemps-Snijders 2017), which utilises the CQL query language.

The creation of the corpus was a major aid for the dictionary editors. Firstly, it greatly increased the breadth of sources for dictionary quotations, as KorBa also includes material that was not used in the paper card index. The comparison of the frequency lists of lemmas present in the corpus to the KXVII index showed that KorBa contains many words absent from the dictionary card index. The corpus therefore allows for a large expansion of the dictionary's entry network. Secondly, the dictionary editors have gained easy access to sources – by using the corpus search engine, they can find forms of any lexeme or stable phrase that interests them and then copy an appropriate quotation to the dictionary form.

The ongoing developmental work on the dictionary and corpus is meant to integrate these resources, both from the point of view of the users and of the dictionary editors.

The first stage of integration has granted the users the ability to automatically issue queries about instances of a given token in the corpus. The links to the corpus search engine are available next to each inflectional form of the lexeme covered in a given entry, as well as below every entry article (in which case the search targets all inflectional forms of the lexeme).[9]

This solution, intended for the users of the dictionary, has also found use in the work of the dictionary editors, for whom it facilitates searching for appropriate quotations for dictionary entries. However, it was clear since the outset that functions that would connect KorBa to the dictionary editor's interface directly are needed. The first of these is the possibility of automatically downloading inflectional forms found in KorBa by clicking a button in the corpus editor platform. This selects all inflectional forms of the selected lexeme and displays them as hints on the platform. Grammatical markers used in the corpus are automatically translated into grammatical terms used in the dictionary. Each form proposed by the system must be approved by a dictionary editor before it is entered into the paradigm presented in the entry. Work is currently underway on automatic sampling of citations from KorBa marked with a source abbreviation and page number for placing in the appropriate section of the dictionary platform.

Further features are also planned to better utilise another resource gathering 17th- and 18th-century source texts: the digitalised card index of the dictionary. Experiments investigating the possibility of using djview4poliqarp (Bień 2016) for indexing the electronic version of KXVII are ongoing. Each card record will be assigned a label (in practice equivalent to the entry title), enabling editors to automatically find any and all cards to be used for the development of an entry (for further information, see Bilińska & Rodek 2020).

The future integration of e-SXVII and other linguistic resources for 17th- and 18th-century Polish will also cover the *Digital Library of Polish and Poland-related Ephemeral Prints from the 16th, 17th and 18th Centuries* (*Cyfrowa Biblioteka Druków Ulotnych Polskich i Polski Dotyczących z XVI, XVII i XVIII Wieku*, CBDU). In this case, the dictionary will be used as a source of information about a given historical lexicon – the meaning of the old lexemes taken from dictionary entries will be displayed to the readers of the old texts, helping them understand the content (for more on this subject, see Ogrodniczuk & Gruszczyński 2019).

8 Conclusion

The case of e-SXVII serves as an example of one path of further development for dictionaries originating in the second half of the 20th century (cf. e.g. Johansson & Battista 2014). Initially developed with traditional methods (based on paper card indices, intended for publication in several volumes over the course of years), they begin to be developed and shared through new digital solutions. The first and most natural change is the creation of a dictionary editor's platform and enabling access to entry articles online. The next stage of development stresses the creation of a convenient way of accessing sources (digital libraries of digitised source texts and dictionary archives as well as electronic text corpora). Later on, these electronic resources are subject to integration.

[7] The word *KorBa*, itself an abbreviation of the Polish phrase *korpus barokowy* "Baroque corpus", is an alternative name for the corpus. In its original form, it included texts from a period mostly dominated by the Baroque style in Polish literature (17th century and 18th century up to 1772). Currently, the corpus is expanded to cover the entirety of the 18th century, thus also including Enlightenment texts. However, the *KorBa* name has become recognizable in the Polish lexicographic circles to the point where the decision was made to retain it.

[8] The structural and linguistic annotation was carried out by a group of transcribers, while morphosyntactic annotation was completed automatically, through a range of utilities adapted for analysis of old texts. For more on the subject, as well as on the creation of the corpus and its current form, see Bronikowska *et al.* 2016.

[9] These solutions have been presented in more detail during the 6th eLex Conference, held in 2019 in Sintra, Portugal, as part of a lecture titled *Integration of the Electronic Dictionary of the 17th-18th-c. Polish and the Electronic Corpus of the 17th- and 18th-c. Polish Texts* by R. Bronikowska, Z. Gawłowicz, M. Ogrodniczuk, A. Wieczorek, and M. Żółtak (relevant paper currently in development).

The new technical possibilities have a larger-than-anticipated influence on the bases of these lexicographic analyses. Far-reaching changes to both the working environment and the ultimate visions for the shape of the dictionary resulting from the specifics of its electronic form allow us to conclude that we are looking at a new lexicographic work, independent from the original print dictionary, although drawing readily on its materials and concept.

9 References

Bień, J.S. (2016). Elektroniczne indeksy fiszek słownikowych. In *Kwartalnik Językoznawczy*, 2 (publ. 2018), pp. 16-27. Accessed at: http://pmichal-kwartjez.home.amu.edu.pl/teksty/teksty2016_2_26/Bien.pdf [16/03/2020].

Bilińska, J. & Rodek, E. (2020). Paper Quotation Slips to the Electronic Dictionary of the 17th and 18th Century Polish - Digital Index and its Integration with the Dictionary. In …… (eds.) Proceedings of the XIX EURALEX International Congress: Lexicography for Inclusion. Alexandroupolis, pp…….

Bronikowska, R., Gruszczyński, W., Ogrodniczuk, M. & Woliński, M. (2016). The use of electronic historical dictionary data in corpus design. In *Studies in Polish Linguistics*, vol. 11, issue 2, pp. 47-56. Accessed at: https://doi.org/10.4467/23005920SPL.16.003.4818 [23/01/2020].

Brouwer, M., Brugman, H. & Kemps-Snijders M. (2017). MTAS: A Solr/Lucene based multi tier annotation search solution. In L. Borin (ed.) Selected papers from the CLARIN Annual Conference 2016 (Aix-en-Provence, 26–28 October 2016). Linköping Electronic Conference Proceedings 136: 19–37. Accessed at: http://www.ep.liu.se/ecp/136/002/ecp17136002.pdf [27/10/2019].

Gruszczyński W. (2005). O przyszłości „Słownika języka polskiego XVII i 1. połowy XVIII wieku". In *Poradnik Językowy*, 7, pp. 48–61.

Johannsson, E. & Battista, S. (2014). A Dictionary of Old Norse Prose and its Users – Paper vs. Web-based Edition. In A. Abel, C. Vettori & N. Ralli (eds.) Proceedings of the XVI EURALEX International Congress: The User in Focus. Bolzano/Bozen: Institute for Specialised Communication and Multilingualism, pp. 169-179. Accessed at: https://euralex.org/wp-content/themes/euralex/proceedings/Euralex%202014/euralex_2014_010_p_169.pdf [02/03/2020].

Majdak, M. (2012). Słownik języka polskiego XVII i 1. połowy XVIII wieku, Kraków 1996- IJP PAN. In *Poradnik Językowy*, 8, pp. 105-111. Also in M. Bańko, M. Majdak & M. Czeszewski (eds.) Słowniki dawne i współczesne. Internetowy przewodnik edukacyjny. Accessed at: http://leksykografia.uw.edu.pl/slowniki/21/slownik-jezyka-polskiego-xvii-i-1-polowy-xviii-wieku-krakow-1996 [21/01/2020].

Majdak, M. (2018). Elektroniczny słownik języka polskiego XVII i XVIII wieku IJP PAN. In M. Pastuch & M. Siuciak (eds.) Historia języka w XXI wieku. Stan i perspektywy. Katowice: Wydawnictwo Uniwersytetu Śląskiego, pp. 176-182.

Ogrodniczuk M. & Gruszczyński W. (2019). Connecting Data for Digital Libraries: The Library, the Dictionary and the Corpus. In A. Jatowt, A. Maeda, S. Syn (eds.) Digital Libraries at the Crossroads of Digital Information for the Future. ICADL 2019. Lecture Notes in Computer Science, vol. 11853, Springer, Cham, pp. 125-138. Accessed at: https://doi.org/10.1007/978-3-030-34058-2_13 [02/12/2019].

Siekierska, K. (1998). Słownik języka polskiego XVII i 1. połowy XVIII wieku, historia przedsięwzięcia, założenia teoretyczne, plan prac, prognozy na przyszłość. In *Język Polski*, 1-2, pp. 82-90.

10 Language Resource References

CBDU - Cyfrowa Biblioteka Druków Ulotnych Polskich i Polski Dotyczących z XVI, XVII i XVIII Wieku/Digital Library of Polish and Poland-related Ephemeral Prints from the 16th, 17th and 18th Centuries. Accessed at: https://cbdu.ijp.pan.pl/ [15/10/2019].

e-SXVII - Gruszczyński, W. (ed.). Elektroniczny słownik języka polskiego XVII i XVIII wieku/Electronic Dictionary of the 17th- and 18th-century Polish. Accessed at: https://sxvii.pl/ [19/03/2020].

KorBa - Elektroniczny Korpus Tekstów Polskich z XVII i XVIII wieku (do 1772 roku)/Electronic Corpus of 17th- and 18th-century Polish Text (up to 1772). Accessed at: http://www.korba.edu.pl [19/03/2020].

KXVII - Kartoteka Słownika języka polskiego XVII i 1. połowy XVIII wieku/Card-index of the Dictionary of the Polish Language of the 17th and First Half of the 18th Century. Accessed at: https://www.rcin.org.pl/dlibra/publication/20029 [29/03/2020].

SJPDor - Doroszewski, W. (ed.) (1950–1969). *Słownik języka polskiego*. Vol. 1-11. Accessed at: https://sjp.pwn.pl/doroszewski [1/04/2020].

SStp - Urbańczyk, S. (ed.) (1953-2002). *Słownik staropolski*. Vol. 1-11, IJP PAN, Kraków. Accessed at: https://pjs.ijp.pan.pl/sstp.html [1/04/2020].

SXVI - Mayenowa, M. R. & Pepłowski, F. (vol. 1–34), Mrowcewicz, K. & Potoniec, P. (vol. 35–37), Wilczewska, K., Woronczakowa L. *et al.* (vol. 27–37) (eds.). *Słownik polszczyzny XVI wieku*. Wrocław: Ossolineum, 1966-1994, Warszawa: IBL PAN, 1995-. Accessed at: http://spxvi.edu.pl [1/04/2020].

SXVII - Siekierska, K. (ed.) (1999-2004). *Słownik języka polskiego XVII i 1. połowy XVIII wieku*, Vol. 1, Fasc. 1-5, Kraków.

Acknowledgements

Article prepared within the project The extending of the Electronic Corpus of the 17th- and 18th- Century Polish Texts and its