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

Paper Quotation Slips to the Electronic Dictionary of the 17th- and 18th-Century Polish - Digital Index and its Integration with the Dictionary

Bilińska-Brynk J.[1], Rodek E.[2]

[1] *University of Warsaw*
[2] *Polish Academy of Sciences*

Abstract

The paper presents the results of experimental paper quotation slips' tagging that was conducted to investigate the possibility of electronic indexing of scanned paper quotation slips constituting a citation archive (a card-index) for the *Dictionary of the 17th- and 1st half of the 18th-Century Polish* (e-SXVII https://sxvii.pl).

The paper citation archive consists of more than 3 million paper quotation slips posing an exemplification of ca. 116,000 of words, which means 86,000 dictionary entries – all of them placed in 836 boxes. There is the need for integration of the archive and the lexicographic panel in order to accelerate the lexicographic work and eliminate human-related mistakes. The test allowed the authors to determine the project priorities, main methodological problems and to decide on future project proceedings. The presented case study may be interesting for other lexicographic teams facing the same problems and looking for an efficient, cheap and quick solution to the problem of using such an abundance of available data.

Keywords: quotation slips, historical dictionary, indexing, card-index, citations archive

1 The need for lexicographic tools modernisation

The Electronic Dictionary of the 17th- and 18th-Century Polish (e-SXVII https://sxvii.pl) records vocabulary that comes from Polish baroque and the early Age of Enlightenment, which is known as the Middle Polish language period (cf. Bronikowska et al. 2020). The dictionary history goes back to 1954, when *Polish Language of 17th- and 18th-Century Research Group* was established at the Institute of Polish Language Polish Academy of Sciences and dictionary data excerption started (Majdak 2012 [21/01/2020]; 2018: 177-178; Siekierska & Sokołowska 1999: IV [19/03/2020]). Citations archive containing lexical data with its context usage was preserved on paper quotation slips, filed and stored in boxes in alphabetical order[1].

Initially, a full excerption was held, i.e. the whole lexical resources from 422 marked sources from the selected age were analysed; however, at the beginning of the 1990s, when there were already 83 thousand of words recorded in the paper catalogue, the decision was taken about starting a non-full excerption (selective, called reasoned) (Siekierska & Sokołowska 1999: VI [19/03/2020]). Finally, in the paper citation archive (KXVII) there are stored 2.8 million entry cards that confirm the existence of 116 thousand words (about 86 thousand entries) attested in 275 texts (various volumes) that functioned in 1601-1750 (Siekierska 1998: 84-86). Apart from that, there also originated a paper catalogue and an index of proper names (ca. 11 thousand entries), a paper catalogue and an index of foreign words (ca. 7 thousand entries), a paper catalogue of entries for Michał Abraham Troc's[2] dictionaries. All of them are now rarely used by lexicographers mainly because of the lack of electronic editions.

Since the moment of huge transformation of the dictionary form and its development into an electronic one (Bronikowska et al. 2020; Gruszczyński 2005) there have also been essential changes in its production process. There emerged the need for developing tools enabling online work but basing on foregoing sources. Therefore, the whole electronic index of the card catalogue that is now essential for dictionary entry editors was prepared. Unfortunately, the index is only a list of headwords recorded in the catalogue; hence, it contains only the basic information about the existence of the word, but does not make it possible to quickly reach the quotation slips and verify their contents.

The current analogue form of the card catalogue makes it gradually useless, and uncomfortable – to say the least – both for the usage and preparation, as well as archaic. It is possible to use either paper card catalogue stored in the dictionary editorial office or its digitised version available as scans in DjVu format on the Digital Repository of Sciences Institutes server (RCIN; https://rcin.org.pl/dlibra/publication/20029). The scans are ordered by cards distribution in boxes, i.e. mainly in alphabetical order (with exception of phonetic variant forms filed together with the main entry), not numbered and only divided into sections corresponding to the boxes volume, i.e. containing ca. 3,5 thousand quotation slips. Searching through such a collection is quite troublesome. Moreover, it is essential to remember that there is the need for maintenance of the server containing the scanned catalogue, which is not a property of the Polish Language Institute Polish Academy of

[1] More on the history of e-SXVII (Gruszczyński 2005; Majdak 2018; Siekierska & Sokołowska 1999 [19/03/2020]).
[2] M.A. Troc was the author of three dictionaries: *Nouveau dictionnaire françois, allemand et polonois*, v.1 (1744) and v.2 (1747), Lipsk, and *Nowy dykcyonarz to iest mownik polsko-francusko-niemiecki*, v. 3 (1764), Lipsk. All of them have been incorporated into the source canon for e-SXVII.

Sciences, which is the owner of the KXVII catalogue[3].
Furthermore, making KXVII available in a new digital form is becoming more and more necessary as there are no more djview plugins in web browsers in the form they used to be in the past and this makes opening scans stored as DjVu files more difficult, non-user friendly and troublesome or even almost impossible for some users.
When the Electronic Corpus of the 17th- and 18th-Century Polish Texts (to 1772) (KorBa; https://korba.edu.pl/) was launched, it became another significant citation source for dictionary entries and, as it is accessible online, it developed into the main source of material exemplification for the entries[4] because the editors do not use now the paper KXVII as frequently as before.
It must not be, however, forgotten that there are quotations in the paper catalogue from the sources that are available neither in KorBa nor in online digital libraries, nor as paper books in the e-SXVII editorial office. It is a major impediment for the editors as they can observe a given word only in the quotation slip without being able to check the broader word context. Frequently, those are the only records of the words or variant forms of the basic words. Moreover, KorBa does not collect many texts that constituted the basis of the first dictionary material excerption and are recorded in KXVII[5].
In conclusion, it has to be admitted that there is a major need to balance the meaning of both citation data sources – the KXVII paper catalogue and the KorBa corpus – and transform an analogue citation archive into a functional tool for the e-SXVII editors.

2 Analogue quotation slips in modern lexicography

There are multiple approaches to the retrodigitisation of the dictionaries. There is a possibility of scanning the dictionary, performing OCR, data encoding and data enrichment (Kallas J. et al 2020: 26). In the e-SXVII, which is a digitally born dictionary, we aim to integrate the dictionary with the data collected in the paper form. Therefore, we need to scan it and encode the data.
The idea of digitising and making paper quotation slips available to the users is not new. There have already been projects involving such actions, e.g. *Dictionary of Old Norse Prose* (ONP) (cf. Johannsson 2019). However, there was another approach to the issue of citations and their results and goals are different from the goals set in KXVII project.
In case of *a Dictionary of Old Norse Prose*, it was decided that the citation archive would be scanned for the needs of the integration with a dictionary so the editors started with finishing the headword list they already had and then scanned the quotation slips. This part of work has already been done when it comes to KXVII, but it was not aimed from the very beginning at making the scans a part of the dictionary itself so the scans were neither stored nor named in a proper form for the processing that is now necessary. After making the headword list and scanning the quotation slips in the ONP project, the editors processed the quotation slips multiple times and used database to store them, including tagging the scans. The process took ca. 3 years (Johannsson 2019: 254). Then the dictionary assistants keyed in relevant citations into the database. "The edition and the citations are linked by the sigla in the database (...)" (Johannsson 2019: 255). At the end of the process the editors had completely processed scans and information stored in quotation slips that they would further use. However, for the needs of our project, where there is a shortage of financial sources and time, there is a need to shorten the process of making quotation slips handier for the editors and useful for the users. Therefore, we would like to make only some of the work similar to that done in the ONP project and we need our own database and interface to integrate it with our dictionary interface.

3 The experimental indexing

In order to transform card-index KXVII into a form that would make working with it comfortable and efficient, there was designed a project of indexing quotation slips and incorporating them into the respective entries in e-SXVII in both editor and user interfaces. Therefore, five randomly chosen card boxes were indexed in aim to optimise the future indexing of the catalogue. These were card boxes containing headwords starting with the letter from the middle of the alphabet (ca. 17,5 thousand paper quotation slips which means 0,625% of the collection). It was done by means of djview4poliqarp software (Bień 2016 [16/03/2020]). The software was initially developed to enable searching digitised texts in the DJVu format and was later supplemented with an indexing function. In the assignment we used the latter functionality. Although it turned out that the software was not efficient and suitable enough for the planned work, the test allowed us to estimate the time and cost of preparing a new index to the citation archive and also to determine the project priorities, main methodological problems and to decide on future project proceedings. Please compare the figure 1 with the example of a scanned paper quotation slip and a test index in the djview4poliqarp software. On the left there is a scan and on the right there is an index where the first and the last scan of the scans' scope containing examples of the headword are tagged. After the word there is a number informing whether it is a first scan or the last one and how many scans are in the scope and the brackets inform whether it is a beginning of the scope (left square bracket) or the end (right square bracket). Therefore eg. *marszczyć się 0020]* (English *to pucker, cockle*) means that there are 20 scans in the scope and it is the last scan.

[3] RCIN was established as a scientific institutes consortium with EU and Polish government financial sources, but it is now supported by institutes themselves (https://rcin.org.pl/dlibra/text?id=Projekty [14/07/2020]). Therefore, there is neither stability nor guarantee that the collected resources will be available in the future as it depends on the current economic and political situation.
[4] The first KorBa edition was published in 2013-2018 and since 2019 the second edition has been prepared. It is now being expanded (it will contain 25 million tokens) and the time horizon was moved until 1800, which is 50 years longer than in the card index (https://korba.edu.pl/overview [14/07/2020]).
[5] In the second edition of KorBa there will be a large amount of texts that are important lexicographic sources, but which were not incorporated into the first edition of the corpus.

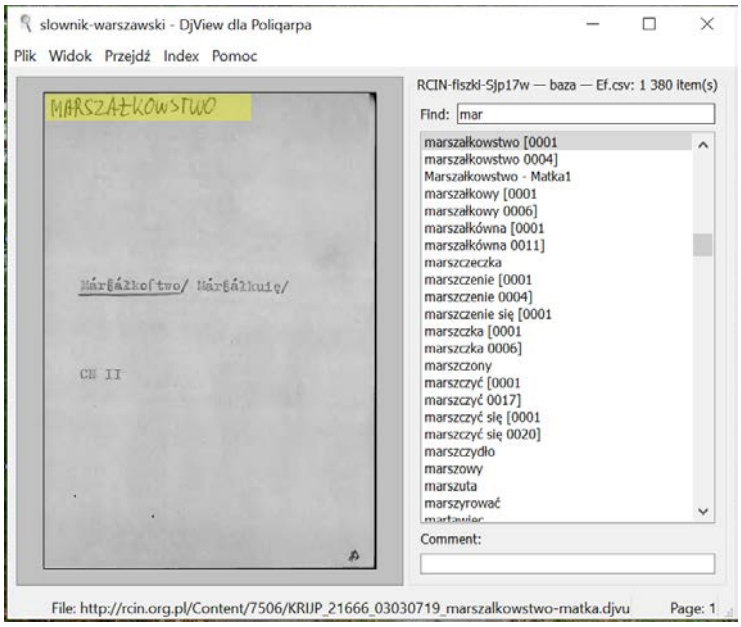

Figure 1. A screehshot of the djview4poliqarp software, scanned paper quotation slip *marszałkostwo* (English *marshalcy*) and the test card index

The software allows, among other things, to tag scans (in our case scanned quotation slips) with e.g. the entry name that this card relates to. As there are multiple quotation slips for every headword (sometimes even hundreds), it was decided that only the first scan and the last scan were tagged, which made it possible to divide the scans into groups relating to one word and number them. We would like the software to number automatically also the rest of the scans within the scope which would allow the editor to orientate themselves in the selection place. It should be possible thanks to the card meter and special tags.

4 The results

Thanks to the experimental tagging there were determined the priorities of the future project:

- Maximum comfort for the index user, which de facto means collecting as much data in one place as possible.
- Annotating person's work optimisation, i.e.:
    - minimising the time needed to annotate a scan;
    - avoiding situations where a annotator would need to make their own decisions;
    - minimising the possibility of errors.

It sometimes happens that the quotation on the slip does not apply to the headword and illustrates a different word from the one written on the card. However, the headword is shown in the previous index in the editor interface and, therefore, the new index should not be ordered newly and the tagger should not be given the permission to encode the proper headwords. Only the entry editor should verify the quotation slip content while working on the entry. In this way the person tagging the archive does not need to have any special scientific skills to perform the task, which would make it possible to outsource tagging and thereby relieve the e-SXVII editorial board work.
The main problems that occurred was how to tag quotation slips in two specific situations:

1) Presentation of two headwords on one quotation slip that disturb the alphabetical order. Phonetic and morphological variants that are illustrated in the quotation were additionally lemmatised (cards are dually lemmatised) and incorporated into the card scope without preserving the alphabetical order. Therefore, there is no possibility to prepare a linear index, e.g. within the scope of the cards containing quotations for the entry *maska* (English *a mask*), one can find also not ordered cards with the variant form *maszka* and also cards with two headwords signed as *maska, maszka*. Compare examples in the figure 2.

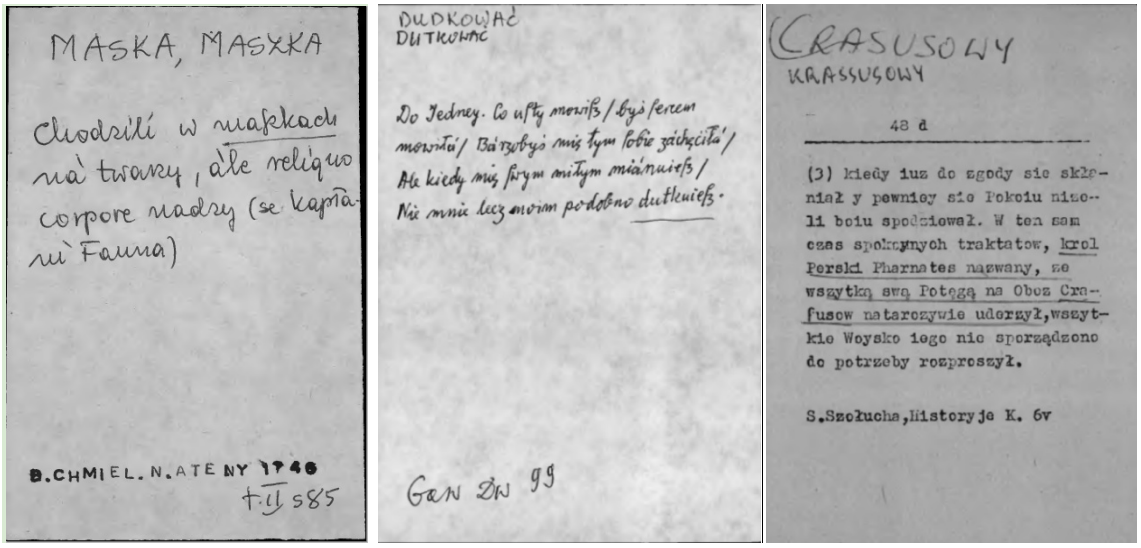

Figure 2. Examples of paper quotation slips with two headwords

2) Accidentally misplaced individual quotation slips.

Both problems are fixable by the modification of the software search engine query syntax, e.g. by enabling searching for any character string instead of conducting an alphabetical search as is being done now. Then there would be no searching problem with cards tagged 'X a. Y' ('X or Y', where X and Y mean headwords).
It also seems impossible to accelerate the work using the OCR software as the headwords are handwritten and there have been multiple quotation slips editors. Moreover, the vocabulary is historical both when considering the forms themselves and spelling (e.g. *dedukcyja,* English *a deduction*, now spelled as *dedukcja*) and there also sometimes occur quotation slips that contain postprepositional phrases as their headwords (e.g. *po frantowsku*[6]) that are now lemmatised in a different way. Furthermore, the tagging person could tag the quotation slips according to the current e-SXVII headwords preparation guidelines, e.g. passive adjectival participles (*frasowany,* English *worried*) are lemmatised as infinitives (*frasować,* English *to worry*) and only when there is no record of the personal verb form should they be lemmatised as participle forms. Thus, in the new electronic index there would sometimes be a different headword from the headword written on a paper quotation slip.
The analysis allowed to estimate the time needed for the project and indirectly also its cost. When using the chosen software in its current form with the already designed editorial guidelines, one needs ca. 3 hours to tag one box of quotation slips, which means 2,508 hours necessary to perform the task (836 boxes) by one person, which results in ca. 16 months of one full-time working person. However, it is considered impossible to work efficiently 8 hours per day doing such a task; hence, the numbers would certainly increase.
The above-described possible changes to the djview4poliqarp software could improve and accelerate the tagger's job and also eliminate the main query problem. It would considerably decrease costs of indexing the archive and speed it up. The estimated time per box would be two hours, i.e. ca. 10 months of one-man job.
However, the software is standalone, which means that one has to install it on a computer and then, if there are many people working on the same project, which would be expected in the described case, there is a need to merge their work afterwards (in CSV files). It leads to other possible errors and even to data loss. In this case it would also be more demanding and time consuming to manage the project as well as to control contractors' work than while using an online application programme.
It seems that developing such software for the needs of the project would be the best solution to the problems described. It would make it possible to do the task remotely, control the tagger(s) in the real time, facilitate assigning material to the contractors and, thanks to the necessary backup option, avoid the danger of possible data loss. It would shorten the estimated indexing time to ca. 1.5 hours per box (ca. 8 months per one person), which is around half of the time needed when using the current software option. It should also be possible to integrate the new index with e-SXVII, which is one of the basic assumptions of the task.

5 Integrating KXVII with e-SXVII

We strive to transform KXVII into a tool for e-SXVII editors that will be functional and user-friendly. Presently, the dictionary is in the phase of creating the so-called stub entries, i.e. entries containing at least one recorded form and illustrated in the quotation grammar form (cf. Bronikowska et al. 2020). The editors can use one of the citation sources and more often they opt for KorBa than KXVII. Moreover, KorBa has already been integrated with e-SXVII (both with the

[6] It is a combination of the preposition "PO" with an adjective in Dative in the so-called short historic form. However, as the Polish short adjective forms were already linguistic relicts in the 17th-18th. c. we treat them in e-SXVII as morphological variant forms and record them in dual entries, e.g. BEZPIECZNY, BEZPIECZEN. But in case of adjectives ending in -ski we do not recreate a possible short form, but we only record long ones.

editor's interface and user's interface, cf. Bronikowska et al. 2016 [23/01/2020]; Bronikowska et al. 2020). However, it has to be noted that a new KXVII index integrated with the scanned quotation slips (as distinct from a present index being only an entry list) and its implementation in the editor's interface will be crucial in the next stages of e-SXVII development.
The entry editor should have access to each and every archived quotation in a nick of time and to be able to preview the scanned quotation slips in the entry edition part of the interface. Basic functionalities would be: information about the number of quotation slips relating to one headword, reference to the place of the slip in the collection, marking/tagging scanned quotation slips as already processed. It is crucial to give a possibility to change the order of the cards and grouping them, e.g. depending on the meaning while editing the entry. It could be useful if the editor could also reject wrongly matched, false or blank quotation slips.
KXVII is, on the one hand, a lexicographic tool and should be integrated with the e-SXVII editor's interface, but, on the other hand, being an archive of quotations coming from the works written in 1600-1750, it can be interesting for the researchers working on the language of this period and, therefore, it could also be available for e-SXVII users.
Admittedly, the process of data excerption took 40 years and during this time there were various decisions made when solving some problems and the data is not fully consistent. For instance, there are paper quotation slips with only a headword and reference information[7] of the quotation without the quotation itself. On some slips there are also illegible (or difficult to read) handwritten notes concerning grammar forms or meanings made by numerous editors. The abbreviations used originally have also changed as the growing number of dictionary sources required doing so in a more systematic way. However, lists of abbreviations are already prepared in both forms allowing users to decode the meanings and available to the e-SXVII users. The KXVII scans are already available in the public domain so it is possible to use it now, but it would be quicker if they were incorporated into the e-SXVII entries.
A close communication between quotation slips and dictionary entries, their incorporation into the entries, especially germ ones, would enable showing the whole collected but still not processed material. Presently, at the e-SXVII website there is a button "More quotations in Baroque Corpus", which automatically generates an appropriate query to the corpus and the results are shown on the corpus website in the new web browser tab or window. The integration of the KXVII with the dictionary user interface could be done in a similar way enabling the user to get access to all the available quotations not only those incorporated into an entry.

6 Summary

The quickly changing reality, new technologies development and growing users' expectations also provoke changes in lexicographers' work. There is a need to adjust the tools in order to be able to take advantage of the material efficiently and fully.
The integrated paper citation archive would be especially useful both for the e-SXVII editors and its users, but can also be interesting for other linguists researching the language of that time. As it is historic material, there are e.g. uncontemporised *appelative names* that would be interesting for some researchers.
Developing new indexing software would also enable quick processing of the three smaller abovementioned collections that were produced during the excerption of dictionary sources: the card catalogue and index of proper names, the card catalogue and index of foreign words (7 thousand paper quotation slips) and M.A. Troc's dictionary card catalogue.
We suppose that our approach to the analogue quotation slips may be interesting to other lexicographic teams or even libraries that face the same problems and are looking for an efficient, cheap and quick solution to the issue of using the abundance of available data.

7 References

Bień, J.S. (2016). Elektroniczne indeksy fiszek słownikowych. In *Kwartalnik Językoznawczy*, 2, p. 16-27. Accessed at: https://doi.org/10.14746/kj.2016.2.2 http://pmichal-kwartjez.home.amu.edu.pl/teksty/teksty2016_2_26/Bien.pdf [16/03/2020].

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].

Bronikowska R., Majdak M., Wieczorek A. & Żółtak M. (2020) The Electronic Dictionary of the 17th- and 18th-century Polish. In (eds) *Proceedings of the XIX EURALEX International Congress: Lexicography for Inclusion.*, (in the current volume).

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

Johannsson, E. (2019). Integrating analog citations into an online dictionary. In C. Navarretta, M. Agirrezabal & B. Maegaard (eds.) Proceedings of the Digital Humanities in the Nordic Countries 4th Conference, pp. 250-258.

Kallas J. et al (2020). D1.1 Lexicographic Practices in Europe: A Survey of User Needs. Accessed at: https://elex.is/wp-content/uploads/2020/06/Revised-ELEXIS_D1.1_Lexicographic_Practices_in_Europe_A_Survey_of_User_Needs.pdf [26/07/2020].

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.

[7] A siglum: resource name abbreviation and page number.

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.

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.

Siekierska K. & Sokołowska T. (1999). Historia *Słownika*, [w:] Słownik języka polskiego XVII i 1. połowy XVIII wieku. T. 1 z. 1, Kraków, pp. IV-VIII. Accessed at: https://rcin.org.pl/dlibra/show-content/publication/edition/25137?id=25137 [19.03.2020].

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*. Accessed at: https://rcin.org.pl/dlibra/publication/20029 [6/04/2020].

MED - Middle English Dictionary. Ed. Robert E. Lewis, et al. Ann Arbor: University of Michigan Press, 1952-2001. Online edition in Middle English Compendium. Ed. Frances McSparran, et al. Ann Arbor: University of Michigan Library, 2000-2018. Accessed at: http://quod.lib.umich.edu/m/middle-english-dictionary/ [16/04/2020].

RCIN - Repozytorium Cyfrowe Instytutów Naukowych. Accessed at: https://rcin.org.pl/dlibra/publication/20029 [6/04/2020].