Ana-Maria Gînsac/Mihai-Alex Moruz/Mădălina Ungureanu

THE FIRST ROMANIAN DICTIONARIES (17TH CENTURY). DIGITAL ALIGNED CORPUS

Abstract This paper presents the project "The first Romanian bilingual dictionaries (17th century). Digitally annotated and aligned corpus" (eRomLex) which deals with the editing of the first bilingual Romanian dictionaries. The aim of the project is to compile an electronic corpus comprising six Slavonic-Romanian lexicons dating from the 17th century, based on their relatedness and the fact that they follow a common model in order to highlight the characteristics of this lexicographical network (the affiliations between the lexicons, the way they relate to the source, the innovations towards it, their potential uses) and to facilitate the access to their content. A digital edition allows exhaustive data extraction and comparison and link with other digitized resources for old Romanian or Church Slavonic, including dictionaries. After presenting the corpus, we point to the necessary stages in achieving this project, the techniques used to access the material and the challenges and obstacles we encountered along the way. We describe how the corpus was created, stored, indexed and can be searched over; we will also present and discuss some statistical analyses highlighting relations between the Romanian lexicons and their Slavonic-Ruthenian source.

Keywords Romanian lexicography; 17th century; Church Slavonic; bilingual dictionaries in electronic format

1. Introduction

Our study focuses on the first Romanian bilingual dictionaries (six Slavonic-Romanian dictionaries dating from the 17th century) and their digital editing. Starting from the context of their elaboration, we present the stages of the digitization project, also pointing at how this edition can be exploited and integrated into other language resource projects dedicated to Romanian.

1.1 Context

The first half of the 17th century in Eastern Europe saw the growing expansion of Catholicism and Protestant currents in an area dominated by Orthodoxy. This expansion triggered various reactions from the Orthodox clergy. One of the most notable such reactions came from the Metropolitan of Kiev, Petru Movilă (1597–1647), who translated and elaborated works related to worship and dogma. The issue of a series of Church Slavonic linguistic instruments, among which Pamvo Berynda's Slavonic-Ruthenian Lexicon (1627) and Meletie Smotrițki's Grammar (1619), was also associated with the context of this counter-reformation of Orthodoxy.

Petru Movilă's descent from a noble Romanian family played a key part in the close cultural relations between Kiev and the Romanian Principalities (Moldova and Wallachia), and thus his activity had a considerable influence on the cultural movement in the Romanian Principalities. The editing and copying of Slavonic-Romanian bilingual lexicons, having as a model Berynda's lexicon, could be associated to the same context. They are the most consistent part of the Romanian lexicography before the 18th century (for an overview, see Seche 1966, pp. 9–11).

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

1.2 Corpus

The six Slavonic-Romanian lexicons compiled in the second half of the 17th century having as a model Pamvo Berynda's Lexicon (= Lex.Ber.) are edited within the eRomLex project (see infra, 2); they are large (Lex.Ber. has more than 7,000 entries and the inventories of the Romanian dictionaries are equally extended, except for Lex.Mard., which records 4,574 entries), are complete (a letter is missing from one of them and some pages from another), all preserved in manuscripts. Four of these are kept at the Romanian Academy Library in Bucharest and two can be found in Russian Libraries (in Moscow and Sankt Petersburg). They all date from the same era, come from the same area (today's Wallachia) and seem (probably with one exception) related to each other not only by the common Slavonic model, but as modified copies of a unique Romanian model, either lost or still undiscovered (see Felea 2021). Lex.St. and Lex.3473 are part of miscellanies that comprise a Romanian version of Meletie Smotrițki's grammar; this fact also gives an idea of the purpose of their compiling. These lexicons are: the Lexicon from Rom. ms. no. 1348, Library of the Romanian Academy, Bucharest (1-84v) (furthermore Lex.1348); the Lexicon from Rom. ms. no. 3473, Library of the Romanian Academy, Bucharest (files 1-369) (= Lex.3473); the Lexicon of Mardarie Cozianul, Rom. ms. no. 450, Library of the Romanian Academy, Bucharest, 1649 (= Lex.Mard.); the Lexicon from Moscow, Russian State Archive of Old Documents, Fond 188, Оп. 1. Ч. 2., p. 491 (= Lex.Mosc.); the Lexicon from Petersburg, the National Library of Russia in Sankt Petersburg (notice n° Q.XVI.5 – Славяно-молдавскийсловарь) (= Lex.Pet.); The Lexicon from Rom. ms. no. 312, Library of the Romanian Academy, Bucharest (41r-216v) (= Lex.St.). For a more thorough description of the lexicons, see Gînsac/Ungureanu (2018, pp. 850–853). The term *corpus*, as used throughout this paper, is not used in the more common sense given in corpus linguistics, but in the broader sense of collections of writings having something in common. In the specific context of dictionaries, most of the text included does not contain proper sentences, but rather disparate words or word lists, given as glosses for Slavonic headwords. Because of this, we consider that the absolute number of words is less relevant than the number of entries (see section 3).

1.3. Current state of knowledge

1.3.1 Concerning the research on the lexicons

This group of lexicons has been studied rather sporadically; the only edited lexicon is the oldest one, Lex.Mard. (Crețu 1900); the others have been analysed focusing on small samples. Crețu (1900), in his introductory study, provides a brief description of the whole group; Bogdan (1891) focuses on the description of Lex.Pet., for which he also establishes the source; Ciobanu (1914) deals in the same manner with Lex.Mosc. This issue has been revisited recently (Gînsac/Ungureanu 2018; Felea 2021), and the comparative editing of the six lexicons was proposed as an objective of the eRomLex project: "The first bilingual Romanian dictionaries (the 17th century). Digitally annotated and aligned corpus".[1]

1.3.2. Electronic lexicographic corpora

The earliest efforts towards the digitization of Romanian lexicographic resources targeted modern dictionaries. The Romanian Language Thesaurus (Romanian Academy, 1906–2010,

[1] See Gînsac/Moruz/Ungureanu 2021; http://www.scriptadacoromanica.ro/bin/view/eRomLex/.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

19 volumes) is available within the eDTLR project (Romanian Language Dictionary in Electronic Format) and provides multiple interrogation criteria based on: word, form, etymology, frequency, age, dialectal area, stable combinations, compound words, authors, period (Cristea et al. 2011). The main objective of the more recent CLRE project (Essential Romanian Lexicographic Corpus) is the alignment at the entry-level of the most important dictionaries – old and new, general or specific – of the Romanian language (for further details, see Clim 2015, pp. 101–104; https://clre.solirom.ro/).

The Multilingual Buda Lexicon (1825), considered the first modern normative dictionary of the Romanian language, was edited and processed in electronic format between 2011 and 2013 (available at: https://doi.org/10.26424/lexiconuldelabuda). The dictionary can be consulted according to several criteria: lemma, language (Romanian, Latin, Hungarian and German), grammatical form, semantic-stylistic value, etymology, idioms, phrases, quotations etc.

In South-Eastern Europe there are some multilingual lexicographical resources in digital format, of great importance for the Romanian language, which used the Cyrillic alphabet until the 19th century. F. Miklosich's Lexicon Palaeoslovenico-Graeco-Latinum is regarded as the most relevant dictionary in Slavic studies, and its digital edition allows interrogation by words, word parts and grammatical categories (Miklosich 1865).

The Old Church Slavonic Dictionary is available in digital format for Bulgarian (Totomanova 2021), allowing word-based interrogation (see https://histdict.uni-sofia.bg/oldbgdict/oldbg_search/). This dictionary is part of a digital platform for Bulgarian language and literature containing: Unicode fonts, diachronic corpora, historical dictionaries equipped with tools for writing and editing the entries, grammatical dictionaries, prototypical search engine, and virtual keyboard (Ganeva 2018, p. 117; see also Tasovac 2020).

A similar digital resource for Slavonic is Gorazd: An Old Church Digital Hub (http://www.gorazd.org/?q=en/node/12), a database developed by the Institute of Slavic Studies (The Czech Republic Science Academy), which includes three modern dictionaries of Church Slavonic (Knoll 2021). The database is accessible via the "Gulliver" interface translated into Czech and English. The interface can be interrogated by: dictionary, word, grammatical ending, type of entry (main, cross-reference, exhausted), texts quoted within the entries.

2. Project presentation

2.1. Corpus creation and storage

The first stage in creating of the corpus consisted of obtaining the lexicons in an editable format. To this end, we tested Transkribus, an automatic handwriting recognition programme (see https://readcoop.eu/transkribus/). However, the specific format of the lexicons (written on columns), particular handwriting (in some cases quite irregular) made the use of Transkribus relatively inefficient (see Fig. 1; manual correction would have been extremely time consuming), therefore the editable format was obtained manually.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

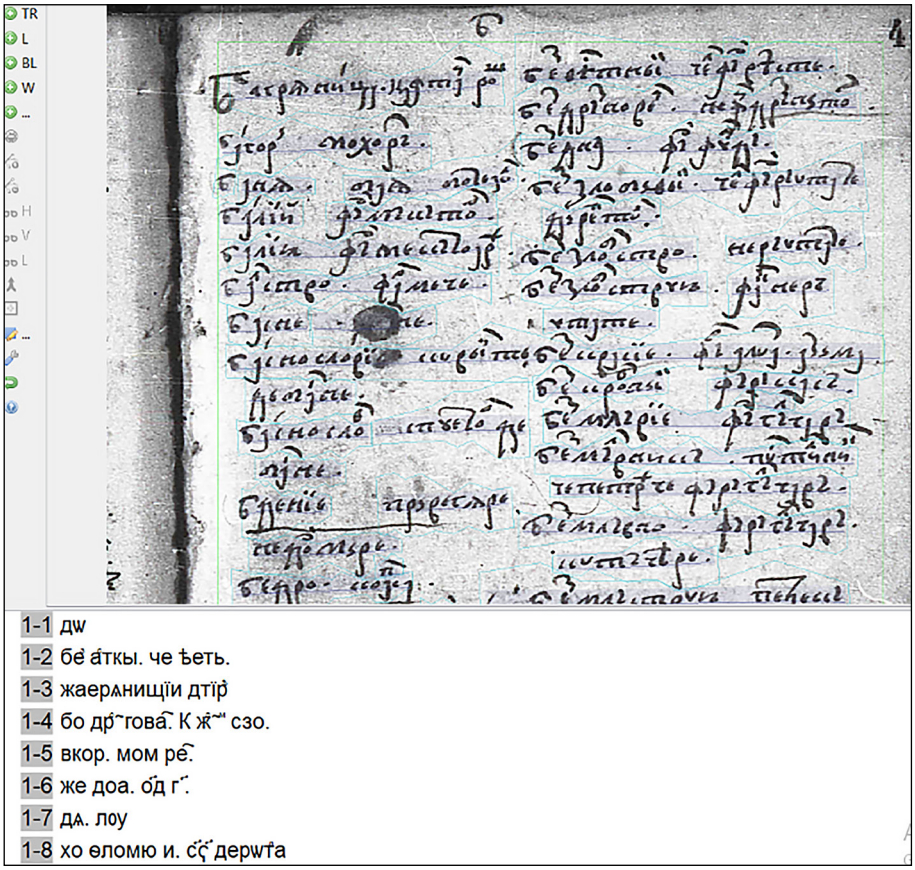

Fig. 1: Automatic text recognition using Traskribus on part of Lex.St.

Each entry was filed in a standard Word form, allowing users to automatically extract distinct information (headword, definition, and location: row, column, entry number for recording the sequence of entries, comments).

2.2 Indexing

From the standard form discussed above, we have automatically extracted the entries in a structured format to be imported into the platform. To this end, we have transformed the Word form to an XML format (MS Word XML) through Visual Basic scripts. This XML format, however, is complex and large amounts of formatting information are not relevant to the extraction process. We have thus transformed the Word XML format to a more simplified version, which contains only the necessary information. In order to better control the flow of imported entries (so as to spot potential errors in the processing chain, for example), we chose to perform the processing at the letter and volume level: we only import the entries beginning with the same letter from all the lexicons.

After the conversion to a structured format, the entries were automatically aligned. This was done manually by pointing to the equivalent Lex.Ber. entry (if extant) during the transcription process. To reduce alignment error rate, we have built, on the basis of the Lex.Ber. equivalent, a simplified form of the word by removing of accents, replacing of superscripts with standard letters, and replacing equivalent letters with the same surface representation (e. g., the letters ї ї ı и й I И are transformed into и). Replacement of letters is done according to a correspondence matrix, developed iteratively, which, at the time of writing, contains

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

more than 200 replacement pairs. This simplified form presents multiple advantages: alignment is more precise, as spelling differences are flattened; words which are not found in Lex.Ber. can be aligned as well; searching is easier and more intuitive over simplified forms, since user cannot know what writing peculiarities a given word might have. The aligned entries are then stored as one object, with multiple equivalent entries from different lexicons, in a proprietary XML format. This format can be easily transformed to XML TEI or PDF by means of XSL transformations.

For storing the entries in an indexed form, and for efficient and intuitive searching and viewing, we have chosen the XWiki[2] technology. This allows for collaborative editing of entry sets, fast and parameterizable searching (via scripting and plugins), layered access and restricting permissions so as to avoid accidental modifications.

2.3 Searching the corpus

In XWiki, each entry is a Wiki page, which is automatically indexed in the platform database, and can be efficiently interrogated using much faster tools than XQuery, XWiki Query Language, which is significantly more efficient as it is derived from SQL and runs on a specific instance of a relational database. For better accuracy of search, we have also lemmatized and performed morphological analysis on the glosses, using the tool described in (Patraș/Pavel/Haja 2007).

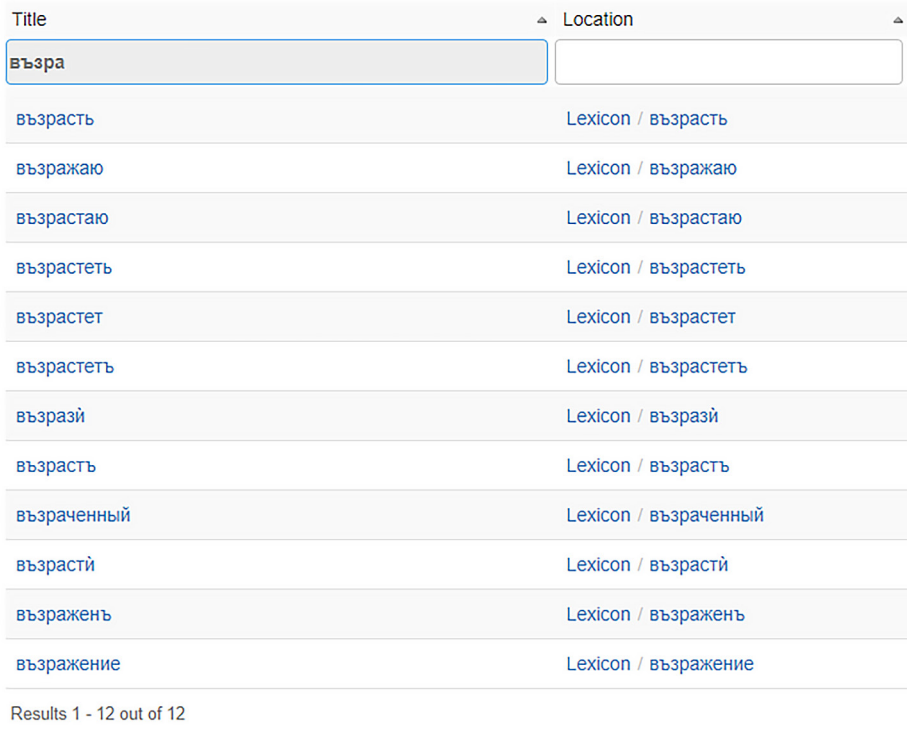

Fig. 2: Search interface

[2] www.xwiki.org.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

To prototype the query module and to test its efficiency, we have provided, for the project members, a multi-criterial search interface prototype; this allows for word or part of word searches, and, for testing purposes, page author and date of modification. This prototype will be expanded into the full search interface for the corpus. Figure 2 shows the search interface prototype.

3. Selected results

Since the creation of the lexicon corpus is still a work in progress, we have performed statistical analysis on only part of the corpus. Below, we have given the statistical analysis for the entries under the letter L in all 6 of the lexicons, compared to their source:

- 427 aligned entries; of these, only 173 are found in Lex.Ber.
- Lex.Ber. has 216 entries, which means that 43 of them can't be found in any of the six lexicons
- 248 entries in Lex.Mosc
- 217 in Lex.St., of which 13 are doubles (2 separate entries for the same headword) and 1 is triple
- 255 in Lex 3473, of which 7 are doubles
- 213 in Lex.Pet, of which 11 are doubles
- 228 in Lex.1348, of which 12 are doubles and 1 is triple
- 136 in Lex.Mard, of which 2 doubles.

Further research will focus on the words from Lex.Ber. missing from the Romanian lexicons, on similarities and differencies between the lexicons in terms of entries inventory and definition structures.

4. Future work

In terms of future research, we intend to investigate the manner in which the lexical inventory of the lexicons is found in other documents of the period which are available in electronic format. Also, on the basis of the alignment with Lex.Ber., we can align the lexicons with similar resources of the period in other languages and with Romanian lexicographic electronic corpora (see above, 1.3.2.). The valorization of the eRomLex corpus will enrich the already existing lexicographic corpora with new meanings, variations in form or morphology, attestations. Upon completion of the project, the electronic dictionaries will be made freely available, in an open-source format, at www.scriptadacoromanica.ro. The search functions will allow researchers to identify new morphological forms and earlier attestations or even words which are not yet registered in the Romanian Thesaurus Dictionary. For translation studies, the eRomLex electronic dictionary could be used in those situations where glosses are encyclopaedic, providing samples for the Romanian language and not just single equivalents, and as a bilingual tool for Slavonic translations into Romanian. The dictionary could also be used for studies regarding the history of mentalities or cultural history.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

References

Lex.Ber. = Pamvo Berynda: Леѯиконъ славенороссκїй и именъ тлъкованїе, Kiev, 1627. Edition by V. Nimciuk, 1961. Online: http://litopys.org.ua/berlex/be.htm (last access: 19-05-2022)

Bogdan, I. (1891): Un lexicon slavo-român din secolul XVI. In: Convorbiri literare 25 (3), pp. 193–204

Ciobanu, Ş. (1914): Славяно-румынскій словарь библіотеки Московскаго Общества Исторіи и Древностей No 240. In: Русскій филологическій вестникъ 71 (1), pp. 75–88.

Clim, M. (2015): La lexicografía rumana informatizada: tendencias, obstáculos y logros. In: Domínguez Vázquez, M. J./Gómez Guinovart, X./Valcárcel Riveiro, C. (eds.): Lexicografia de las lenguas románicas. Aproximaciones a la lexicografía moderna y contrastiva, vol. II. Berlin/München/Boston, pp. 95–110.

Crețu, G. (1900): Mardarie Cozianul. Lexicon slavo-românesc și tîlcuirea numelor din 1649. Bucharest.

Cristea, D./Haja, G./Moruz, A./Răschip, M./Pătrașcu, M. (2011): Partial statistics at the end of the eDTLR project–the Thesaurus Dictionary of the Romanian Language in electronic form. In: Zafiu, R./Uşurelu, C./Oprea, H. Bogdan (eds.), Romanian language. Hypostasis of linguistic variation. Acts of the 10th Colloquium of the Chair of Romanian Language, Bucharest, 3–4 December. Bucharest, pp. 213–224.

Felea, I. M. (2021): The Staicu lexicon in relation to lexicons belonging to the Berynda family: orthography and structure. In: Diacronia 13, A181, pp. 1–13.

Ganeva, G. (2018): Electronic diachronic corpus and dictionaries of Old Bulgarian. In: Studia Ceranea 8, pp. 111–119.

Gînsac, A.-M./Ungureanu, M. (2018): La lexicographie slavonne-roumaine au XVIIe siècle. In: Zeitschrift für romanische Philologie 134 (3), pp. 845–876.

Gînsac, A. M./Moruz, M. A./Ungureanu, M. (2021): Slavonic–Romanian lexicons of the 17th century and their comparative digital edition (the eRomLex project). In: Diacronia 14, A192, pp. 1–11.

Knoll, V. (2021): Gorazd: An Old Church Slavonic digital hub and the Romanian Slavonic studies. In: Diacronia 14, A197, pp. 1–9.

Miklosich, Fr. (1865): Lexicon Palaeoslovenico-Graeco-Latinum. Lipsiae. http://www.monumentaserbica.branatomic.com/mikl2/ (last access: 22-03-2022).

Patraş, V. S./Pavel, G./Haja, G. (2007): Resurse lingvistice în format electronic. Biblia 1688. Regi I, Regi II – probleme, soluţii. In: Pistol, I. D./Cristea, D./Tufiş, D. (eds.): Lucrările atelierului Resurse lingvistice şi instrumente pentru prelucrarea limbii române, Iaşi, 14–15 decembrie 2007. Iaşi, pp. 51–60.

Seche, M. (1966): Schiţă de istorie a lexicografiei române, vol. I: De la origini pînă la 1880. Bucharest.

Tasovac, T. (2020): The historical dictionary as an exploratory tool: a digital edition of Vuk Stefanovic Karadzic's Lexicon Serbico-Germanico-Latinum. Ph. D. thesis. Dublin. http://www.tara.tcd.ie/bitstream/handle/2262/92750/Tasovac.pdf?sequence=1&isAllowed=y (last access: 22-03-2022)

Totomanova, A. M. (2021): Electronic research infrastructure for Bulgarian medieval written heritage: history and perspectives. In: Diacronia 14, A193, pp. 1–9. https://doi.org/10.17684/i14A193en.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.

Contact information

Ana-Maria Gînsac
Institute of Interdisciplinary Research, Department of Social Sciences and Humanities, "Alexandru Ioan Cuza" University, Iași
anamaria_gansac@yahoo.com

Mihai-Alex Moruz
Faculty of Computer Science, "Alexandru Ioan Cuza" University, Iași
mmoruz@info.uaic.ro

Mădălina Ungureanu
Institute of Interdisciplinary Research, Department of Social Sciences and Humanities, "Alexandru Ioan Cuza" University, Iași
madandronic@gmail.com

Acknowledgements

This work was supported by a grant of the Ministry of Research, Innovation and Digitization, CNCS/CCCDI – UEFISCDI, project number PN-III-P1-1.1-TE -2019-0517, within PNCDI III.

This paper is part of the publication: Klosa-Kückelhaus, Annette/Engelberg, Stefan/Möhrs, Christine/Storjohann, Petra (eds.) (2022): Dictionaries and Society. Proceedings of the XX EURALEX International Congress. Mannheim: IDS-Verlag.