Dicționariul Limbei Române (LM) by A. T. Laurian and I. C. Massim – the Digital Form of the First Romanian Academic Dictionary

Marius-Radu Clim, Mădălin-Ionel Patrașcu, Elena Isabelle Tamba

"A. Philippide" Institute of Romanian Philology, Romanian Academy, Iasi, Romania
E-mail: marius.clim@gmail.com, madalin.patrascu@gmail.com, isabelle.tamba@gmail.com

Abstract

We aim to present a project called DICTIONE, the digitalization of the first Academic Dictionary written in the Romanian language, the *Dicționariul limbei române* (LM) by A. T. Laurian and I. C. Massim. LM was published in three volumes between 1873-1877, has 3600 pages and includes 70,000 headwords out of which over 20,000 words are the personal creation of the authors. This dictionary is unique in Romanian cultural history, due to the fact that the two lexicographers did not aim to illustrate the language vocabulary in a particular moment of its history, but instead intended to impose a certain direction to the language, beyond its use at that time. A novelty is the fact that the authors proposed new words which they attempted to popularize through this dictionary. The impact of this work on Romanian culture is a significant one: from more than 20,000 newly-created words, based on terms taken from Latin, most stayed in use and became neologisms. This fact led to the enrichment of Romanian terminology in many domains and to the modernization of the Romanian language at the same time as that of other European cultures.

Keywords: digitization; academic; neologism; cultural heritage

1. European context

European cultures were, and still are, preoccupied with the recovery and valuation of their own lexicographical thesauruses, within which the cultural stages of a language are stored. The current digital means permit not only the recovery of these "cultural databases", but their promotion by making this lexical richness available to the public. We mention several examples, such as:

a) *Trésor de la langue française* (1971–1994, first printed edition), http://atilf.atilf.fr;

b) *Dictionnaire de l'Académie française.* La 9 édition en ligne (1694, first printed

edition), http://atilf.atilf.fr/academie9.htm;

c) *Diccionario de la lengua española de la Real Academia Española* (DRAE) (1780, first printed edition), http://buscon.rae.es/draeI/;

d) *Tesoro della lingua italiana delle origini* (TLIO); http://tlio.ovi.cnr.it/TLIO/index2.html;

e) *Deutsches Worterbuch der Grimm* (DWB) (1838-1961, first printed edition), http://germazope.uni-trier.de/Projects/DWB;

f) *Oxford English Dictionary* (1928, first printed edition), http://www.oed.com.

In 2016 we celebrated 150 years since the establishment of the Romanian Academy, whose original purpose was to create a dictionary and a grammar of the Romanian language, to solve the orthographic problem and to write a book of Romanian history. In this context, what we intend to do through the project DICTIONE, namely the digitalization of the first Academic Dictionary written in the Romanian language, is to both recover the Romanian cultural heritage and make profitable the activity, ideas and erudition of the first scholars who realized this academic thesaurus. The project DICTIONE fits perfectly into the European trend and aims to capitalize on the Romanian lexicographical heritage by digitizing the first academic dictionary to have been printed in the Romanian language.

2. Related Romanian research projects

This project had several related and relevant predecessors concerning the digitization of various lexicographic works. We would like to mention the most representative. The project *CLRE* concerned the *Corpus lexicografic românesc esenţial. 100 de dicţionare din bibliografia DLR aliniate la nivel de intrare* [Essential Romanian Lexicographic Corpus. 100 dictionaries from DLR Bibliography aligned by entries][1] received national financing between 2010 and 2013, and was conceived as a linked database between 100 dictionaries from the DLR Bibliography aligned by entry level. Today, this project continues as part of the research plans of the Romanian Academy – Iasi Branch, by the "A. Philippide" Institute of Romanian Philology. This constitutes a great resource for lexicographers, providing fast access to dictionaries and helping to present a better historical perspective of the Romanian lexicography (Clim, 2015).

Another Romanian lexicographic resource available online is *Lexiconul de la Buda* [The Lexicon of Buda], the electronic edition[2] of the first etymological and explanatory dictionary of the Romanian language, a benchmark for modern Romanian

[1] The CLRE project will be accessible at http://clre.philippide.ro at the end of 2017. More about this project in Clim et al. (2016).

[2] http://www.bcucluj.ro/lexiconuldelabuda/site/login.php

lexicography. The current electronic edition restores the volume that was printed at the Buda press, in 1825, under its full title, *Lesicon romanescu-latinescu-ungurescu-nemtescu quare de mai mulţi autori, in cursul a trideci, si mai multoru ani s'au lucrat. Seu Lexicon valachico-latino-hungarico-germanicum quod a pluribus auctoribus decursu triginta et amplius annorum elaboratum est.* This is a multilingual dictionary, targeted more towards the equivalence of terms in four languages (Romanian, Latin, Hungarian and German) and less towards the semantic description of lemmas (Patrașcu et al., 2016).

3. DICTIONE project

3.1 DICTIONE – highlights of the first Romanian academic dictionary

In contrast to this dictionary, the one implicated in the DICTIONE project is a dictionary exclusively dedicated to Romanian vocabulary, etymology and semantics. *Dicționariul limbei române* (LM) [Romanian Language Dictionary] by A. T. Laurian and I. C. Massim was published in three volumes between 1873-1877, has 3600 pages and includes 70,000 headwords, of which over 20,000 words are the personal creation of the authors, based on latinized terms. In this dictionary, the authors were interested in achieving two main outcomes: creating Romanian terminology, capable of expressing the concepts of modern culture; and proving by lexical means the latinity of the Romanian language. However, in order to facilitate the process of consulting this lexicographical work by the foreign specialists, the Romanian lemmas are accompanied by their Latin equivalent.

The preface contains three chapters describing the principles used in elaborating the dictionary, the conception of the Romanian language and how it can be modernized. The authors, who are representatives of the latinizing direction in Romanian linguistics, proclaim the purity of the language as a basic principle of the entire work and argue for maintaining the latinity of the Romanian language and eliminating foreign words, using the French Academy as a model. The last chapter promotes the orthography according to the etymological principle, the authors aim to admitting only the letters that correspond to the primitive sounds preserved by the Romanian language from Latin. Laurian was a supporter of etymological writing, to the detriment of phonetic writing, and attempted to demonstrate as clearly as possible the Latin form and origin of Romanian words (Clim, 2012). Thus, access to understanding this writing was possible only for those who knew Latin. Also, the authors tried to assimilate the current forms of Romanian words – forms that resulted from a long historical evolution – with the appropriate forms from Latin, and the difference between the words inherited from Latin and the Latin-Romance neologisms was ignored.

Laurian and Massim divide the words in this dictionary in two categories, considering this historical criterion:

a) words in use before 1830;

b) words in use after 1830.

The words of the first category are numerous in regions inhabited by Romanian people, while those from the second category had not yet spread until the dictionary was published and were thus marked in this work with an asterisk.

The two authors declared themselves against defining the terms through synonyms, as, in their opinion, this creates more confusion regarding the meaning of the headwords. Therefore, they adopted definition by periphrasis. Broad definitions are followed by illustrative phrases. Another lexicographical novelty is the fact that LM includes detailed orthoepic explanation at the beginning of each letter, taken from the work of Laurian, *Tentamen criticum in originem, derivationem et formam linguae romanae in utraque Dacia Vigentis vulgo valachicae*, Vienna, 1840. In general, the dictionary article is structured as follows: the entry word, bearing no accent indication, followed sometimes by the explicit indication of pronunciation (e.g. *coctoriu*, *pronuntiatu coptoriu*), and then details about the word flexion and about the grammatical category. The equivalence of the Romanian term with another from a well-known foreign language like Latin should be appreciated. Also, the fact that the etymology of the entry word is indicated before its explanation is a novelty in the lexicographical technique. It also has a solid scientific argumentation in the preface of the work, many of the etymological solutions proposed by the authors are used even today. Furthermore, the adoption of multiple etymologies as a way of explaining the origin of Romanian terms is notable.

However, the greatest value of this dictionary is that it tried to impose a large number of neologisms into the Romanian language, over 20,000 of them. Although unnatural, this effort to fill the gaps of Romanian vocabulary remains quite impressive. Here are some of the neologisms proposed by the authors of this dictionary: *accelerator, adjunct, admirativ, adversitate, aerofagie, austeritate, anxietate, benign, bibliologie, biochimie, biotic, calvar, fabricabil, fabulație, factură, fastuozitate, felin, feroce, ferocitate, figurativ, fluență, formativ, fotogenic, fracționa, frazeologic, genetic, genuin, germina, ginecologic, giratoriu, gnoseologie, gnostic, grandilocvență, imersiune, imixtiune, imobiliar, imobiliza, imortaliza, matador, mercantil, meteoric, metronom, micrometru, miligram*, etc. These terms exist in the current Romanian language, many of them being (re)borrowed from Romance languages. In order to illustrate the peculiarity of this dictionary we present the definition of the term *factură* [invoice]:

> factură: *factura*, s. f., **factura**, resultatu allu facerei, opera; 2. in commerciu, (it. **fattura**, fr, **facture**), statulu care arréta in detaliu speci'a, cantitatea, calitatea si pretiulu merciloru ce tramitte unu fabricante sau unu negotiatoriu la veri-unulu dintre confratii sau associatii sei, la veri-unu commissionariu, etc. (dar și ceilalți termeni din familia de cuvinte: *factura*, *facturariu*, *facturat*).

Laurian and Massim did not aim to highlight the state of the language in their time, but instead intended to set a certain direction, outside the use of the period, and thus proposed new words (an impressive novelty in the lexicographical works) that they tried to popularize through this dictionary. In other words, they pointed out to the maximum the regulatory role of this Academic dictionary. It is recognized that the two lexicographers have sought to provide modern definitions, both for the new terminology, and for old words. This is the reason for which the work has been appreciated by subsequent lexicographers who have treated it as a valuable source of information and documentation. Through the efforts of the authors to enrich the vocabulary of the Romanian language, many neologisms that have been preserved in the language were put into circulation. Taking into consideration the large number of registered neologisms, this dictionary is unprecedented in Romanian culture. This dictionary is mentioned in prestigious lexicographical works and represents a documentary base exploited by Romanian language researchers. Regarding the problem of etymology, and especially the primary, not only direct, etymology problem studied by European researchers interested in the migration of words, this dictionary is a valuable bibliographic resource. The conversion of this first dictionary of the Romanian Academy in a digital format, easy to read and to use in the documentation process and in linguistic research, would facilitate access to the information gathered by Laurian and Massim to Romanist specialists.

In the current European context, there is interest in collating linguistic resources for determining a correct etymology of a new term for any given language, of finding the primary etymon and also the transition of the term from the source language to the host language. Thus, the digitization of this work will help Romanian researchers solve etymological problems, while also assisting foreign researchers who want to study the filiation of the terms or meanings. Therefore, this dictionary is proposed for inclusion on the list of prestigious European dictionaries to be used by all the people interested in the study of the Romanian language.[3]

3.2 DICTIONE goals

Through its digital version, this dictionary will have a significant impact both on lexicographic Romanian works, and on research of Romanian language history.

The objectives of this project are not new and they reflect the usual objectives of digitizing an old dictionary. The project DICTIONE aims to digitize the first academic dictionary printed in the Romanian language. The proposed objectives are the following:

[3] This dictionary will be included in the European Dictionary portal www.dictionaryportal.com created in the COST project *ENeL European Network of e-Lexicography.*

1. transforming the digital form into an editable format and correcting the text of the dictionary;

2. recognizing the entries of the dictionary;

3. creating a site for the project that would contain, among other information, the scanned and the editable versions of the dictionary. The written version will be annotated at the levels of morphology, etymology, and so on. This will permit various types of search.

4. aligning the dictionary to other digital versions of Romanian dictionaries. In the primary stage, this will be done at headword level. After this step, annotated information will be considered to be linked with similar data from other dictionaries.

5. aligning the dictionary to other Romanian dictionaries and integrating it into the list of representative dictionaries of European languages;

6. creating an exhaustive list of neologisms defined into the dictionary and developing a study regarding their circulation during the period 1862-1927.

3.3 DICTIONE project steps

This project is meant to last two years and comprises a number of steps that we will present briefly here. The first step is administrative and presumes the preparation of the lexicographic material and the establishment of the working stages for the entire team. That means, first of all, to check the current status. There are already scans of this dictionary (some made in Romania: for example one made by the Bucharest Metropolitan Library, available on the site www.digibuc.ro, or in CLRE project and another made by Google Books) and it is necessary to verify and compare the existing scans and to select the most appropriate scan for the project DICTIONE. Before using the OCR program, it is necessary to process the 3600 scanned images to eliminate page noise and any typographical points and to optimize them for the process of recognizing characters. This process will be followed by a OCR testing phase. It is possible for the program to recognize most letters, but it will not be able to recognize many words because of the latinizing orthography: vowels, doubled consonants, the lack of diacritical marks and others. After recognizing the characters, a program will be created in order to allow the validity of each term separately. Then the text will be corrected for keeping the exact orthographic version of the dictionary, as it was drafted. This step, to the text correcting arising from the character recognition process, will be assisted by software specifically developed for this purpose. This computer program should identify types of corrections made to the text by the specialist and propagate similar changes throughout the text. We rely on the fact that we can achieve this desideratum, because from our experience with OCR-ized texts we

have noticed that these software tools introduce inaccuracies towards printed text which can be resolved systematically. In the first stage, the program will identify possible errors by statistical analysis of the text (for example: words with one or two occurrences). Secondly, the program will assist a human specialist and propose words which are similar to the modified words operated on the text. The similar terms will be generated by considering the OCR score of each recognized sign and an analysis based on n-grams. This is the approach by which we hope to reduce the length of this stage and at the same time to develop a software instrument that will later be used for similar texts. Another goal of the project would be that the electronic text obtained from the correction of the OCR stage be similar to the printed one in both content and graphic form.

After the text is revised the next step is to recognize the dictionary entries. In order to validate the delimitations of the dictionary entries, a program created through the CLRE project will be used. This program will be adapted to insert the modern orthographic form for each headword, to align this dictionary to the other Romanian works. Subsequently, a parser will be created to delimit the information (the fields) of a drafted article: entry word, morphological information, etymology, the translation of the term, examples etc. Once the text from the dictionary is parsed, it is necessary to make the correlations between the terms from the dictionary and those mentioned by the authors in the preface in order to put at the disposal of the users not only the definitions, but also the commentaries and analyses made by Laurian and Massim [for example, a term as *federatione* [federation] will be found according to the modern orthography „federație/federațiune" both in the text dictionary and also in the preface. In addition, parsing the text will enable searching the Latin etymon of the terms if, for example, a term borrowed from Latin is searched.

The last step involves the creation of a website for the dictionary, aligning the project DICTIONE to other digital Romanian dictionaries, such as CLRE, and including this dictionary on the list of representative dictionaries at a European level (www.dictionaryportal.eu). After the dictionary is fully digitized we intend to extract an exhaustive list of new terms included in Laurian and Massim's work, terms that did not exist in the language at the time, and to conduct research into their adaptation in the Romanian language.

The digitization of this dictionary comes with some risks, due to its uniqueness in Romanian culture. The greatest risk is the potential incorrect automatic recognition of words, because the latinizing orthography requires that the user should mentally transpose the terms into a modern orthographical version. Also, this dictionary can cause problems to public users because of the Latinist spelling. However, transposing the scanned text into its latinizing version reflects the novelty of this dictionary and for this reason researchers will employ automatic or semi-automatic electronic means, but also using manual verification to maintain the accuracy of the dictionary text. In addition, all headwords will appear in their modern orthographic form in order to

allow easy access for all users familiar with the Romanian language. If the headwords are commented in the Preface, the user will be able to search according to the current form of the terms. Because the entire text will be corrected, researchers can search, for example, for a Latin term and verify if it has been borrowed into the Romanian language and determine under which form it is mentioned in LM. This will allow correlations to other languages (Romanic or not) which have also borrowed that Latin term.

This dictionary remains a valuable source for the lexicographers involved in drafting *The Dictionary of the Romanian Language (DLRi),* the current academic thesaurus, but also for the dialectologists and philologists interested in the etymology of old words or their semantic evolution.

4. Acknowledgements

This work was supported by a grant of the Romanian National Authority for Scientific Research and Innovation, CNCS – UEFISCDI, project number PN-II-RU-TE-2014-4-0195.

5. References

Catană-Spenchiu, A.-V. & Clim, M.-R. & Patrașcu, M. & Tamba, E. (2015). CLRE. Corpus lexicographique roumain essentiel. Résultats et perspectives in *Integrare europeană/identitate națională; plurilingvism/multiculturalitate – limba și cultura română: evaluări, perspective (European Integration/ National Identity; Plurilingualism/Multiculturality – Romanian Language and Culture: Evaluation, Perspectives)*, Luminița Botoșineanu, Ofelia Ichim (eds). Roma, Italia, ARACNE Editrice, Colecția „Danubiana", p. 323-332.

Clim, M.-R. & Tamba, E. & Catană-Spenchiu A.-V. & Patrașcu, M. (2016). CLRE. Corpus lexicographique roumain essentiel. 100 dictionnaires de la langue roumaine alignés au niveau de l'entrée et, partiellement, au niveau du sens, in vol. Buchi, Éva, Chauveau, Jean-Paul & Pierrel, Jean-Marie (éd.) : *Actes du XXVII^e Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013)*, Strasbourg, ÉLiPhi (Editions de linguistique et de philologie), vol. 2, Section 16, p.1611-1622, (and on-line in vol. Trotter, David/Bozzi, Andrea/Fairon, Cédric (éd.): *Actes du XXVII^e Congrès international de linguistique et de philologie romanes (Nancy, 15-20 juillet 2013). Section 16 : Projets en cours; ressources et outils nouveaux.* Nancy, ATILF: http://www.atilf.fr/cilpr2013/actes/section-16.html).

Clim, M.-R. (2012). *Neologismul în lexicografia românească*, Iaşi, Editura Universităţii „Alexandru Ioan Cuza", 355 p.

Clim, M.-R. (2015). La lexicografía rumana informatizada: tendencias, obstáculos y logros in vol. *Lexicografia de las lenguas románicas. Aproximaciones a la lexicografía moderna y contrastiva,* vol. II, coords.: María Dolores Sánchez

Palomino y María José Domínguez Vázquez, eds.: María José Domínguez Vázquez, Xavier Gómez Guinovart y Carlos Valcárcel Riveiro, Editura De Gruyter, pp. 95-110.

Dănilă E. & Haja, G. (2005) Neologismul din perspectivă lexicografică, în „Studii şi cercetări lingvistice", LVI; nr. 1–2, ianuarie–decembrie, Bucureşti, p. 71–78.

De Schryver, G.-M. & Joffe, D. (2004). On How Electronic Dictionaries are Really Used. In G. Williams & S. Vessier (eds.) *Proceedings of the Eleventh EURALEX International Congress, EURALEX 2004. Lorient: Faculté des Lettres et des Sciences Humaines, Université de Bretagne Sud*, pp. 187–196.

Krek, S. & Kilgarriff, A. (2006). Slovene Word Sketches. In T. Erjavec & J. Žganec Gros (eds.) *Proceedings of the 5th Slovenian and 1st International Language Technologies Conference.* Ljubljana, Slovenia. Available at: http://nl.ijs.si/is-ltc06/proc/12_Krek.pdf.

Patrașcu, M.-I. & Clim, M.-R. & Haja, G. & Tamba, E. (2016). Romanian Dictionaries. Projects of Digitization and Linked Data in Diana Trandabăț, Daniela Gîfu (eds.) *Linguistic Linked Open Data.* 12th EUROLAN 2015 Summer School and RUMOUR 2015 Workshop Sibiu, Romania, July 13–25, 2015. Revised Selected Papers, Springer, p. 110-123.

Pruvost, J. & Sablayrolles, J.-F. (2003). *Les Neologismes*, Paris, Presses Universitaires de France, 128 p.

Seche, M. (1966). *Schiţă de istorie a lexicografiei române*, vol. I: *De la origini până la 1880*, Bucureşti, Editura Ştiinţifică, 192 p.

Tamba, E. & Clim, M.-R. & Catană-Spenchiu, A.-V. & Patrașcu, M. (2012). Situația lexicografiei românești în context european, in „Philologica Jassyensia", An VIII, Nr. 2 (16), 2012, p. 259-268 and on-line, http://www.philologica-jassyensia.ro/upload/VIII_2_Tamba_Clim.pdf

Tamba, E. & Clim, M.-R. & Catană-Spenchiu, A.-V. (2012). The Evolution of the Romanian Digitalized Lexicography. The Essential Romanian Lexicographic Corpus in *Proceedings of the 15 th EURALEX International Congress,* 7–11 august 2012, Oslo, eds. Ruth Vatvedt Fjeld, Julie Matilde Torjusen, Press Reprosentrales, UiO, ISBN: 978-82-303-2095-2, p. 225; the text can be accessed on-line at: http://www.euralex.org/proceedings-toc/euralex_2012/.