Elena Isabelle Tamba

The role of the Institutes of the Romanian Academy in the digitalization process of linguistic research

Abstract

In the last few years, measures have been taken in Romania to create the necessary electronic instruments and resources to support the Romanian language and culture on a transnational level in the general context of the digitalization of basic academic research. In today's digital, multicultural society, this had become an absolutely necessary step to take.

Electronic dictionaries and text corpora structured as databases facilitate knowing, preserving and maintaining cultural identity on a linguistic level and allow the inclusion of a national language in the field of interest of digitalized research into natural languages on a global level.

1. Introduction

One of the objectives of European policies is the preservation and valorization of national linguistic identities, as long as there is a general tendency towards using languages which are privileged by the existence of (electronic) means of promoting them.

Linguistics and lexicography around the world have undergone an extensive process of change, including the modernization of means of writing, consulting, etc., through approaches that involve interconnections between different fields of research.

Romanian linguistics and lexicography have also been marked by this change. In the last few years, measures have been taken in Romania, to create the necessary electronic instruments and resources to support the Romanian language and culture on a transnational level in the general context of the digitalization of basic academic research.

A special stage in the evolution of Romanian linguistics and lexicography at present is the digitalization of research, which involves the digitalization of existing resources on the one hand and digitalization – the creation of dictionaries, new resources, and instruments in an electronic format – on the other. In parallel, linguistic and lexicographic resources continue to be created in classical, printed format.

Basically, digitalization involves converting existing resources in printed format into an electronic format. For example, linguistic and lexicographic corpora, can be created by digitizing printed dictionaries; linguistic corpora can be annotated

morphologically, syntactically, and semantically and lexicographic corpora can include digitalized dictionaries.

Digitalization, in turn, involves the development of lexicographic or linguistic resources such as dictionaries directly in electronic format by creating/using dictionary writing programs and/or sample extraction programs, etc.

Most efforts in the digitalization of lexicographic research have been made under the auspices of the Institutes of the Romanian Academy; more recently, research centers at some universities in the country have become involved.

Today various digital linguistic/lexicographic projects are being carried out in Romania including:

– academic initiatives (most lexicographic digitalization projects are taking place at the Institutes for the Romanian Language and the IT Institutes of the Romanian Academy while a few are being developed at the research centers of some universities in Romania[1] or in some libraries).
– private initiatives (for example, https://dexonline.ro/ – a lexicographic platform initiated by volunteers, projects at some publishing houses, etc.).

In this paper we will highlight the projects of the Institutes of the Romanian Academy.

2. Digitalized linguistics and lexicography in Romanian

2.1 Institutes for Language at the Romanian Academy

In the Romanian Academy there are three institutes where research into the Romanian language is done in the fields of lexicography, lexicology, grammar, history of the language, dialectology, sociolinguistics, and onomastics, etc.:

– Institutul de Filologie Română "A. Philippide", Iaşi/"A. Philippide" Institute of Romanian Philology – https://www.philippide.ro/,

[1] Here we would like to mention some lexicographic digital projects developed in two universities in Romania: *The Lexicon from Buda (1825). Amended and electronically processed edition for online consultation* (http://www.bcucluj.ro/lexiconuldelabuda/site/login.php), a project coordinated by the "Babeş-Bolyai" University of Cluj-Napoca and *Primele dicţionare bilingve româneşti* (*secolul al XVII-lea*). *Corpus digital prelucrat şi aliniat* (eRomLex) [*The first Romanian bilingual dictionaries* (*17th century*). *Digitally annotated and aligned corpus.* eRomLex] – the main objective of this project, developed at the "Alexandru Ioan Cuza" University of Iasi, is the elaboration of a comparative digital edition of the Slavonic Romanian dictionaries from the 17th century (all of them are manuscripts) – http://www.scriptadacoromanica.ro/bin/view/eRomLex/.

– Institutul de Lingvistică "Iorgu Iordan – Alexandru Rosetti", Bucureşti/ "Iorgu Iordan – Alexandru Rosetti" Institute of Linguistics – https://www.lingv.ro/,
– Institutul de Lingvistică şi Istorie Literară "Sextil Puşcariu", Cluj/"Sextil Puşcariu" Institute of Linguistics and Literary History – http://www.inst-puscariu.ro/.

The main projects involving basic research concern the following reference works:

– Dictionary of the Romanian Language,
– Grammar of the Romanian Language,
– History of the Romanian Language,
– Linguistic Atlases covering different areas for the Romanian Language, etc.

Researchers from the above-mentioned institutes are also involved in some international projects, like: DERom (*Dictionnaire Étymologique Roman*, http://www.atilf.fr/DERom/), ENeL (*European Network of e-Lexicography*, https://www.elexicography.eu), ALE (*Atlas linguarom Europae* – https://lingv.ro/atlas-linguarum-europae/), etc.

Research into the Romanian language is also carried out at the IT Institutes of the Romanian Academy, namely in the fields of natural language processing or computational linguistics:

– Institutul de Cercetări pentru Inteligență Artificială „Mihai Drăgănescu", Bucureşti/"Mihai Drăgănescu" Research Institute for Artificial Intelligence – http://www.racai.ro,
– Institutul de Informatică Teoretică, Iaşi/Institute of Theoretical Informatics – http://iit.academiaromana-is.ro/.

2.2 Thesaurus Dictionary of the Romanian Language in the digital age

Great European cultures have had thesaurus dictionaries and text corpora in electronic format for many years now. The main Romanian lexicographic project is the *Thesaurus Dictionary of the Romanian Language* (DA/DLR), which is edited by the Romanian Academy and was started 115 years ago. That is why creating an electronic format which is accessible to scientists and everybody who is interested in learning or studying Romanian in our country or abroad became an absolutely necessary step to take in today's digital, multicultural society.

Fig. 1: *Thesaurus Dictionary of the Romanian Language* (DA/DLR)

For a better understanding of the dimensions of the *Thesaurus Dictionary of the Romanian Language*, we present some statistics and compare them to other large European dictionaries:

– The first edition of the *Thesaurus Dictionary of the Romanian Language* was published in two series: DA (1907-1944) and DLR (1965-2010). It includes 14 tomes with 37 volumes, 20,000 lexicon type pages (between 7,000 and 11,000 characters per page), over 175,000 words (with variants) and over 1,300,000 quotes; the electronic form is being elaborated (first attempt 2007-2010; work in progress). The second edition is also work in progress.
– *Diccionario de la lengua espanola de la Real Academia Espagnola* (DRAE, https://dle.rae.es/diccionario): first printed edition – 1780; 23rd edition – 2014; 93,111 lemmas; first electronic format – 1992.
– *Dictionnaire de l'Académie Française* (https://dictionnaire-academie.fr/): first printed edition – 1694; 9 editions; available online, 55,000 words.
– *Deutsches Wörterbuch der Grimm* (DWB, http://germazope.uni-trier.de/Projects/DWB): 1838-1961; 32 volumes; 350,000 words and variants; first electronic format: 1997-2004.
– *Oxford English Dictionary* (OED, http://www.oed.com/): first edition – 1928, 20 volumes; second edition – 1989; 301,100 words, 2,412,400 quotes; first electronic format – 1988.
– *Tresor de la Langue Française* (TLF), XIXth-XXth centuries (http://atilf.atilf.fr/): first printed edition – 1971-1994; 16 volumes; 100,000 words, 270,000 definitions, 430,000 quotes; electronic format: 1990-2004.
– *Tesoro della lingua italiana delle origini* (TLIO, http://tlio.ovi.cnr.it/TLIO/index2.html): 44,000 words (37,864 published online) out of an intended 57,000 words.

Based on the data above, the *Thesaurus Dictionary of the Romanian Language* can be compared, both in terms of its conception and realization, with similar dictionaries of European languages, and its digitalization is, thus, a normal step in the evolution of Romanian lexicography.

We are preparing the digital form of the *Dictionary* in three projects:

– digitalization of the printed form in the **eDTLR** project (scanning, OCR correction, correction, parsing and uploading to a platform which allows complex searches in the body of each lexicographic entry);
– digitalization of the printed form in the **CLRE** project (scanning and processing in the CLRE platform, which allows, for the time being, consultations at headword level and displaying an image of the page from the dictionary);
– digitalization of the **second edition** of the **DLR** (editing done entirely and directly in a dictionary-writing program).

Digitalization of the Dictionary started in 2007 (until 2010), in a complex project eDTLR *Dicționarul tezaur al limbii române în format electronic* (*Romanian Thesaurus Dictionary in electronic format*) which had as its main objective the acquisition of the complete form of the *Thesaurus Dictionary of the Romanian Language* into electronic format as a result of retro-digitalization, but the research is continuing. The results of the eDTLR project will make the electronic format of the *Thesaurus Dictionary of the Romanian Language* accessible for everybody who knows or is interested in Romanian. The digital form of this *Dictionary of the Romanian Language* in CLRE will be presented in the next section.

The second edition of the *Thesaurus Dictionary of the Romanian Language* is called DLRi (*Dicționarul Limbii Române informatizat – Digital Dictionary of the Romanian Language*). It was started in 2010 by electronically acquiring the textual resources of the Bibliography. (The Romanian language does not have yet a complete electronic corpus – it is still work in progress.) We are now working with an electronic editing interface, adapted by Oxygen. The DLRi is being developed completely in electronic form. A printed format will also be published in parallel. The first part of the letter A was presented to the public in digital format in May 2021 – https://dlri.ro/.

2.3 CLRE: The Essential Romanian Lexicographic Corpus

Creating an Electronic Romanian Lexicographic Corpus has been a constant concern of Romanian lexicographers in the last fifteen years, a fact justified by the broader context of the digitalization of Romanian research.

CLRE. *Corpus lexicografic românesc electronic* (***CLRE***. *Electronic Romanian Lexicographic Corpus*) is a project carried out by the Romanian Academy which involves an electronic collection of dictionaries of Romanian aligned at the entry level. It includes the most important lexicographic works from the very first one written in Romanian in the 17th century to the latest ones. The corpus includes, as its main lexicographic work, the *Thesaurus Dictionary of the Romanian language* (DA/DLR).

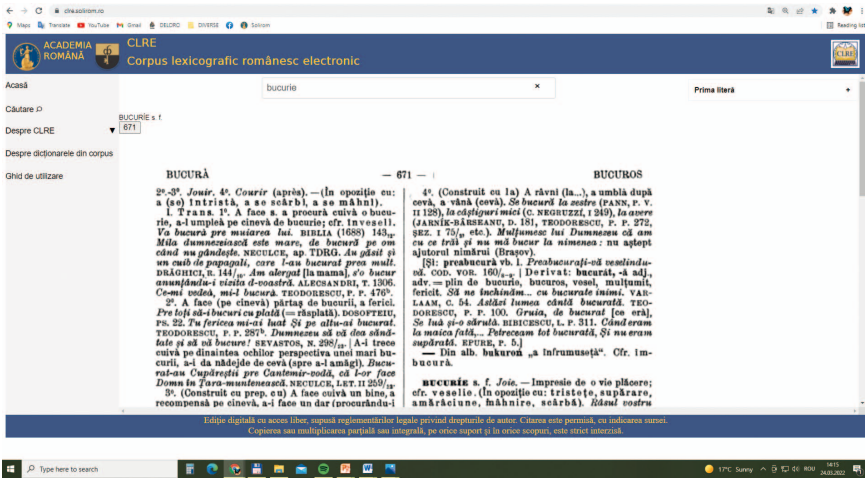

Fig. 2: *Corpus lexicografic românesc electronic* (CLRE)/*Electronic Romanian Lexicographic Corpus* (CLRE)

The main objectives of the CLRE project are:

– to create the largest digital diachronic corpus of dictionaries of Romanian consisting of lexicographic works from the digitized DLR Bibliography (transposed from its classical format, on paper, into digital format) and from digitalized dictionaries (created in an editable electronic format);
– to promote lexicographic works produced under the auspices of the Romanian Academy;
– to provide information from CLRE with free access for the general public.

The first work chosen by the lexicographers from Iasi for publication in CLRE is the *Dictionary of the Romanian Language* produced by the Romanian Academy. The first volume was digitized and published online in September 2021 as Volume I. Part I: A-B and contains 8,517 entries (https://clre.solirom.ro/). This choice was justified by the fact that this is the first volume of the *Dictionary of the Romanian Language* published under the auspices of the Academy and by its parallelism with the publication of the first part of the second edition of DLRi, letter A, written by fellow lexicographers from the Institute of Linguistics "Iorgu Iordan – Al. Rosetti", Romanian Academy, Bucharest (https://dlri.ro/).

CLRE can be compared to two other European lexicographic corpora which are similar in their technical approach:

– *Diccionarios de la lengua española* – a database containing dictionaries edited and published by the Real Academia Espagnola (https://www.rae.es/obras-academicas/diccionarios).

– *Das Wörterbuchnetz* – a collection of 37 electronic dictionaries created at the University of Trier in Germany (https://www.woerterbuchnetz.de/).

The development of CLRE, mirroring other directions for the development of electronic resources, represents a starting point for future research, which may be part of a medium- and long-term research strategy, such as:

– aligning the *Romanian Thesaurus Dictionary* in electronic format (eDTLR) with CLRE DA/DLR and other dictionaries from the corpus;
– using CLRE to elaborate the DLRi (the second edition) and for other lexicographic projects;
– developing large-scale applications on the semantic disambiguation of words;
– selecting entry types to produce new, specialized dictionaries (thematic, etymological, etc.);
– highlighting dictionaries from the database by publishing them online or republishing a dictionary in a mixed format (classical and online);
– turning CLRE into an open corpus (in the sense of the possibility of adding new lexicographic works) for all researchers from the Romanian Academy;
– associating it with other linguistic or multimedia resources, which would bring Romanian lexicography to a level comparable with European lexicography (for example, with the *Dictionnaire Étymologique Roman* (DÉRom) (http://www.atilf.fr/DERom/) or ENeL: European Network of e-Lexicography (http://www.elexicography.eu).

2.4 TDRG

Another lexicographic project published online by the Romanian Academy is the electronic version of the **TDRG** – H. Tiktin, *Rumänisch-Deutsches Wörterbuch* (first edition 1896-1926). The third edition of this dictionary (published in 2003-2005) was digitalized, following the model of eDTLR, in a project involving the Albert-Ludwigs-Universitat in Freiburg, Germany, and the Romanian Academy, and it has now been published online (https://tdrg.solirom.ro/).

2.5 SOLIROM

All of the results of the digitalization process of linguistic/lexicographic research in the Institutes for Language of the Romanian Academy are planned to be published together online on an academic platform called **SOLIROM** (https://solirom.ro/). It will include all electronic resources (DLRi, CLRE, TDRG, eDTLR, etc.) either directly or via a link to the homepage of the project.

Until last year, every academic project mentioned above was published online on a separate web page, but now the results of these projects (digitalized or digital

dictionaries) have been published (or will be published) online on this single platform of the Romanian Academy. SOLIROM promotes a unitary way of working, at the level of specialized institutes, regarding the creation of digital linguistic resources and tools dedicated to the Romanian language and literature. Important projects of the Romanian Academy, such as the new digital edition of the *Dictionary of the Romanian Language* involves permanent collaboration between teams of researchers from several institutions in the country, the use of the same documentation sources and writing tools, so the approach offered by the SOLIROM platform is welcome. This allows, among other things, the alignment of developed language resources, the simplification and streamlining of the publishing process using website templates, as well as the management of published digital resources with minimal resources, which is an important element in the management of research activity.

The platform consists of two sections, a public one which provides digital language resources for public access and a private one with the digital tools needed to manage the platform's digital language resources for the researchers developing the platform.

Now the Romanian Academy is developing a new site with a special area dedicated to Romanian language resources.

2.6 CoRoLa

Another very important project concerning digital resources for Romanian is ***Corpus computațional de referință pentru limba română contemporană*** [Reference computational corpus for contemporary Romanian language] – **CoRoLa** (http://corola.racai.ro/).[2]

The purpose of CoRoLa is to be an online resource for the study and learning of Romanian and so it is a very important resource for lexicographic research as well.

Starting in 2014, this corpus was developed as a priority program of the Romanian Academy. It contains various texts, dating from 1989 to the present day, the purpose of its creation being to provide an objective image of current written and spoken Romanian. The corpus is publicly accessible via two interfaces, one for searching for text data and one for searching for audio data. The main fields of use of the CoRoLa corpus are: linguistic studies; language modeling for the automatic processing of Romanian; developing translation models; language learning; intelligent and multi-criteria indexing and retrieval of textual and oral

[2] Another online resource related to the Romanian Academy is **DIGIBUC** (http://www.digibuc.ro/), the most important Romanian digital library, a project run by the Bucharest Metropolitan Library and the Library of the Romanian Academy. It is the official partner of the European Digital Library EUROPEANA (http://www.europeana.eu/portal/).

information; semantic classification of large volumes of data (text and audio); extracting knowledge from data (text and audio); automatic document summaries; question-answer systems; automatic speech recognition and synthesis; and so on.

3. Conclusions

The aim of this paper is to highlight, in general, the current status of linguistic and lexicographic research in the Institutes of The Romanian Academy in the digital age.

Trends in Romanian linguistics and lexicography include:

- Writing online dictionaries based on continuously increasing text corpora and on various tools (programs for extracting the quotations, for example);
- Developing a Romanian Language Text Corpora (for Contemporary Romanian we have the CoRoLa corpus; a diachronic corpus – work in progress), and linking it to the Thesaurus Dictionary;
- Developing lexicographic corpora (CLRE – work in progress);
- Using dictionary writing systems (DLRi – work in progress);
- Further editing of the printed edition of the *Thesaurus Dictionary of the Romanian Language*;
- Aligning various lexicographic works and creating collaborative programs between academics with lexicographically-oriented publishers etc., as an important subsequent goal;
- Matching electronic lexicographic resources for Romanian – DLRi – CLRE – eDTLR etc. – and all of them with other linguistic resources (possibly multimedia) from Romania and abroad.

Electronic dictionaries and text corpora structured as databases facilitate knowing, preserving, and maintaining cultural identity on a linguistic level and allow the inclusion of a national language in the field of interest of digitalized research into natural languages on at a global level.

References

Ernst, G. (2013): "Romanian". In: Heid, U./Gouws, R.H./Schweickard, W./Wiegand, H.E. (eds.): *Dictionaries. An international encyclopedia of lexicography. Supplementary volume: Recent developments with special focus on computational lexicography*, Berlin/Boston, 687-701.

Hartmann, R.R.K./James, G. (1998): *Dictionary of lexicography*, London.

Kirchmeier, S. (2020): Trends in European language policies with a view to language technology. In: *Bendrin*ė Kalba 93. http://journals.lki.lt/bendrinekalba.

Tamba, E. (2014): La lexicografía Rumana. Historia y Actualidad. In: Córdoba Rodríguez, F./González Seoane, E./Sánchez Palomino, M.D. (eds.): *Lexicografía de las lenguas románicas. Perspectiva histórica. Vol. I*. Berlin/Munich/Boston, 265-282.

Tamba Dănilă, E./Clim, M.-R./Catană-Spenchiu, A./Patraşcu, M. (2012): The evolution of the Romanian digitalized lexicography. The Essential Romanian Lexicographic Corpus. In: Vatvedt Fjeld, R./Torjusen, J.M. (eds.): *Proceedings of the 15th EURALEX International Congress, 7-11 August 2012*. Oslo, 1014-1017. http://www.euralex.org/proceedings-toc/euralex_2012/.

Tamba, E.I. (2017): CLRE. Corpus lexicografic românesc esențial. 100 de dicționare din Bibliografia DLR aliniate la nivel de intrare şi la nivel de sens. In: Haja, G. (ed.): *Lexicografia academică românească. Studii. Proiecte*. Iaşi, 221-234.

Dictionaries

DA = *Dicționarul limbii române*, tom I-II, Tipografia ziarului "Universul". Bucharest, 1907-1944.

DAF = *Dictionnaire de l'Académie Française*. https://dictionnaire-academie.fr/.

DLR = *Dicționarul limbii române (DLR)*, Serie nouă, tom. VI-XIV, Bucharest, 1965-2010.

DRAE = *Diccionario de la lengua espanola de la Real Academia Española.* http://buscon.rae.es/draeI/.

DWB = *Deutsches Wörterbuch "der Grimm"*. http://germazope.uni-trier.de/Projects/DWB.

OED = *Oxford English Dictionary*. http://www.oed.com/.

TLFi = *Le Trésor de la Langue Française Informatisé*. http://atilf.atilf.fr/.

TLIO = *Tesoro della lingua italiana delle origini*. http://tlio.ovi.cnr.it/TLIO/index2.html.

Bibliographical information

This text is part of the book:

Željko Jozić/Sabine Kirchmeier (eds.) (2022): The role of national language institutions in the digital age. Contributions to the EFNIL Conference 2021 in Cavtat. Budapest: Nyelvtudományi Kutatóközpont/Hungarian Research Centre for Linguistics.

This electronic PDF version of the text is accessible through the EFNIL website at:
http://www.efnil.org