Romanian Dictionaries in the ENeL Network – European Network of e-Lexicography

MARIUS-RADU CLIM, GABRIELA HAJA,
MĂDĂLIN-IONEL PATRAŞCU, ELENA TAMBA[1]

This paper presents a short description on a European project, named *ENeL – European Network of e-Lexicography* supported by the European Union Governments, through the COST funding line. The main purposes of this project are: establishing common standards for the printed and electronic dictionaries and the creation of a European portal with information concerning the dictionaries from the European languages that would allow the access to these dictionaries. To fulfil these ideas more than 200 researchers are involved in four working groups. Until the middle of 2015 more than 240 titles of dictionaries were proposed and a part of them have already been included in the portal. The Romanian team – formed of 15 members – proposed a list of four dictionaries or collections of dictionaries, which fulfilled the parameters requested to be added into the portal. Through this portal, the stage of the process of digitization of the lexicographic academic resources from the European countries is shown very clearly; this way, it will be possible to make decisions in terms of European politics, that would permit surpassing the cultural differences. Yet, maybe the most obvious gain offered by this portal is the inter-connection of digital resources in order to offer a very clear image of a term from a semantic and etymological point of view.

Key-words: *dictionary*, *Romanian lexicography*, *European network*, *e-lexicography*, *dictionary portal*

Starting with 2013, a team of Romanian researchers and teachers has been involved in a European project named *ENeL –*

[1] "A. Philippide" Institute of Romanian Philology, Romanian Academy, Iasi, Romania.

European Network of e-Lexicography. This project is supported by the European Union Governments, through the COST funding line, within the program *Individuals, Societies, Cultures and Health* – ISCH (http://www.cost.eu/COST_Actions/isch/IS1305) and aims at creating a lexicographic network that would turn to profit the European multiculturalism in the field of dictionaries.

ENeL – European Network of e-Lexicography is being coordinated by Martin Everaert, professor at Utrecht University of Leiden, Netherlands, and director of the Instituut voor Nederlandse Lexicologie, an institution whose main research direction is the Dutch vocabulary from the 5th century until present times and the analysis of the neologisms in the current language. The vice-president is Iztok Kosem from the Institute for Applied Slovene Studies, Ljubljana, Slovenia.

Currently, the network includes 30 participant countries and approximately 230 members, the specific characteristic of this program being the fact that, during the project, at any time, other people or entities from European countries can join. Details about the project, the involved people and the project activities can be found by accessing the web-site: http://www.elexicography.eu/

The Romanian team of the project is formed of 15 members and is being coordinated by Marius-Radu Clim from the "A. Philippide" Institute of Romanian Philology of the Romanian Academy – Iasi Branch and by professor Dan Cristea, correspondent member of the Romanian Academy and director of the Research Department within the Computer Science Faculty of the "Alexandru Ioan Cuza" University, Iasi. The other members are researchers within the "A. Philippide" Institute of Romanian Philology within the Romanian Academy – Iasi Branch (Gabriela Haja, Mădălin-Ionel Patraşcu and Elena Tamba), researchers and professors within the "Alexandru Ioan Cuza" University, Iasi (Ana Catană-Spenchiu, Ana-Maria Minuț, Mihai-Alex Moruz), researchers from the "Iorgu Iordan – Al. Rosetti" Institute of Linguistics of the Romanian Academy, Bucharest (Mihaela-Rodica Marin, Nicoleta Mihai, Monica Vasileanu, Florin Vasilescu), researchers from the "Sextil Puşcariu" Institute of Linguistics and Literary History of the

Romanian Academy – Cluj-Napoca Branch (Mircea Minică, Maria Ştefănescu) and a professor from the Faculty of Romanian language of the Bucharest University (Ruxandra Cosma).

The project *ENeL – European Network of e-Lexicography* aims at accomplishing the following objectives[2]:

a) establishing common standards for the printed and electronic dictionaries;
b) creating a European portal with information concerning the dictionaries from the European languages that would allow the access to these dictionaries;
c) making an exchange of information and technology in the field of on-line dictionaries in order to create common standards and to offer support to other authors of electronic dictionaries. On this project, we aim at describing the standards for elaborating the dictionaries, the consulting interface both for the academic dictionaries and for the usual dictionaries, but also at making connections between the information from dictionaries in order to offer new opportunities of comparative research between the European languages;
d) publishing articles for promoting the European lexicographic inheritance;
e) providing support to the young researchers during the project meetings.

In order to accomplish these objectives, the activity of the network members was organized in four work groups that would cover the diversity of European dictionaries: WG 1 *Integrated interface to European dictionary content*; WG 2 *Retro-digitized dictionaries*; WG 3 *Innovative e-dictionaries*; WG 4 *Lexicography and lexicology from a pan-European perspective*.

WG 1 implies creating a European portal with information on the dictionaries from the European languages (type of dictionary, the used language etc.) that would permit the access to these

[2] Further details concerning the objectives can be obtained on the website of the project (http://www.elexicography.eu/action/ objectives/) and also on the pages of every working group.

dictionaries; this action will comply with the author rights, according to each case. The purpose of WG2 is to digitize the printed dictionaries, ensuring the application of various standards for accomplishing this objective. This process implies the elaboration of standards for coding and establishing the types of information from the dictionaries, the possibility of introducing new categories of information in order to make the dictionaries more accessible and inter-operable, developing a plan of digitization that would include the necessary parameters and the estimated costs, analysing the possibility of using the information from the dictionaries for computational linguistic applications.

WG 3 aims at developing the dictionaries converted in electronic format by analysing the last evolutions in the field of electronic lexicography and the connections between lexicography and computational linguistics. In this work group, the existing programs are presented in order to elaborate the electronic dictionaries, we analyse the possibility of automatically purchasing lexicographic information and the inter-connection between these dictionaries and other databases morphologically, syntactically and semantically annotated. In the WG4 work group the focus is on the possibility of correlating the information from the European dictionaries through the study of loans, migration of words and meanings between the European languages.

In order to create the European portal, the members of the project gathered the necessary information about the academic dictionaries from all the involved countries. Until de middle of 2015, approximately 240 titles were proposed. Through this action, the objective was to have as much information as possible about the dictionaries of high reputation from the European languages, and to make these normative lexicographic works accessible to the large public, and also to the people who elaborate dictionaries, through this portal. The special focus was on the monolingual dictionaries, which offer a lot of information about a certain language and which is already on-line or in process of digitization. For the selection and sorting of these dictionaries 22 parameters were used, that is: Language [language used for editing the dictionary], Country, Original Title (and subtitle), Translated Title (in English), Abbreviation,

URL, Organization/Publisher, Editor, Accessibility (not on-line, partly on-line, on-line, paywall partly accessible, paywall not accessible), Digitization and mark-up language (HTML full text digitization, page image scan with OCR, page image scan), Search on (lemma, lemma and meaning, not possible to search because of no OCR), Form (audio, no audio), Grammatical Characterization (Y/N), Meaning/definition (in what way does the dictionary provide meaning and definition: no definition or meaning given, synonyms, paraphrases, translation into other languages, images –depicting the meaning–), Etymology (does the dictionary provide etymological information: Y/N), Examples (no, given with source, given without source), Usage information (does the dictionary provide usage information: Y/N), Cross-references (no, other entries, other dictionaries), Dictionary Portal (Y/N), Metalanguage [language used for describing the analysed language], Etymological (if it is an etymological dictionary: Y/N), Dialectal information (Y/N).

All these dictionaries and the information about them will be available on the portal dictionaryportal.eu until the end of the project. Through this portal, the addresses of these dictionaries will be disseminated and, this way, it will be easy to see the stage of the digitizing process of the academic lexicographic works from each European country. Moreover, the portal permits the permanent addition of new dictionaries, under the condition that they would respect the requested conditions. We present here all the nine criteria, as they are listed on the website, including the explanations (http://www. dictionaryportal. eu/en/about/):

a) *trustworthy*: the dictionary does not contain errors, inaccuracies or misunderstandings;
b) *authoritative*: the dictionary has been written by somebody whose judgment on questions of language is respected by the language community;
c) *large*: the dictionary is considerable in size and coverage. Dictionaries of general (non-specialized) vocabulary should normally count their entries upwards of 10,000;

d) *detailed*: the dictionary does not skim over important details. Dictionaries and glossaries which merely list things off, as opposed to explaining things, are excluded;
e) *original*: the dictionary is an original artefact. This definition does not exclude newer or reworked editions of older dictionaries, but does exclude dictionary aggregator websites and glossaries based on harvested data from the Internet;
f) *intended for humans*: in other words, it is not primarily a machine-readable lexical resource for computational linguistics. This means that various lexical databases such as WordNet and FrameNet are excluded, and so are lexical databases extracted automatically from a corpus;
g) *focused on a single language*: the dictionary must be intended to describe one particular language. Monolingual dictionaries meet this criterion by default, but bilingual or plurilingual dictionaries must make a clear distinction between the language they are describing – the *object language* – and the language(s) they are using to describe it – the *metalanguage(s)*;
h) *identifiable*: the dictionary has a name (as opposed to merely a description) and has its own identity which makes it distinct from other dictionaries. In practical terms this means that the dictionary has its own website, or its own section in a larger website, where it can be found and consulted in isolation from others;
i) *usable*: the dictionary can be searched and consulted online and offers a reasonably pleasant user experience. Superficially digitized dictionaries which only have static facsimiles and no significant additional features are excluded by this definition. Dictionaries which require login or are behind paywalls are not excluded but are labelled as such.

These conditions do not only aim the lexicographic criteria, but also those through which these dictionaries can be accessible for free in order to be consulted by the interested users. The purpose is that each European language would have at least one dictionary on this portal.

The list with Romanian dictionaries[3] proposed on this portal includes the following titles:

- *Corpus lexicografic românesc electronic* [Romanian lexicographic electronic Corpus] (CLRE, http://clre.philippide.ro),
- a collection of *Dictionaries of Romanian Language* (http://dexonline.ro/),
- *Dicţionarul limbii române* [The Romanian Language Dictionary] (DLR),
- *Dicţionarul limbii române în format electronic* [The Romanian Language Dictionary in Electronic Format] (eDTLR, http://edtlr.philippide.ro/)[4],
- *Lesicon românescu-latinescu-ungurescu-nemţescu* [The Romanian-Latin-Hungarian-German Lexicon] (LB, http://www.bcucluj.ro/lexiconuldelabuda/site/login.php).

Through this list, the objective is to promote the academic dictionaries, but also those which are already accessible on-line. In addition to the famous dexonline.ro, – currently the most used site of dictionaries of the Romanian language – two other sites were proposed, which appeared on the UEFSCDI projects, during the period 2010–2013, namely: http://www.bcucluj.ro/lexiconuldelabuda/site/login.php – which is the on-line form of the lexicon in 4 languages, published in 1825[5], very important for the evolution of the Romanian language and http://clre.philippide.ro – which is a corpus of dictionaries from the DLR bibliography, scanned, OCR processed, validated and aligned according to entry-word[6].

On this portal, the stage of the process of digitization of the lexicographic academic resources from the European countries is shown very clearly; this way, it will be possible to make decisions in terms of European politics, that would permit

[3] A more detailed presentation of the Romanian lexicography, concerning its history and the stage of digitization can be found in Clim 2015: 95–110 and Tamba 2014: 265–282.

[4] A short presentation of the process of digitization of this academic dictionary can be accessed in the article of Cristea, Haja 2011: 10–11.

[5] Cf. Patraşcu et al. 2016: 115.

[6] For a detailed presentation of this project, see Clim et al. 2016: 83–94.

surpassing the cultural differences. Yet, maybe the most obvious gain offered by this portal is the inter-connection of digital resources in order to offer a very clear image of a term from a lexical point of view. Such a portal will offer ideas for future lexicographic projects and will favour the contact between European lexicographers. By aligning various European dictionaries, answers will be offered to many etymological problems and semantic evolution.

Bibliografie

CLIM M.-R., *La lexicografía rumana informatizada: tendencias, obstáculos y logros*, in María Dolores Sánchez Palomino y María José Domínguez Vázquez (coords.), María José Domínguez Vázquez, Xavier Gómez Guinovart y Carlos Valcárcel Riveiro (eds.), *Lexicografía de las lenguas románicas. Aproximaciones a la lexicografía moderna y contrastiva*, vol. II, de Gruyter Verlag, p. 95–110, 2015.

CLIM M.-R., TAMBA E., CATANĂ-SPENCHIU A.-V., PATRAŞCU M., *CLRE. Corpus lexicographique roumain essentiel. 100 dictionnaires de la langue roumaine alignés au niveau de l'entrée et, partiellement, au niveau du sens*, in David Trotter, Andrea Bozzi, Cédric Fairon (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, p. 83–94, http://www.atilf.fr/cilpr2013/actes/ section-16.html, 2016.

CRISTEA D., HAJA G., *The Thesaurus Dictionary of Romanian Language in Electronic Form*, in "Clarin", Newsletter of Clarin Project, no. 13, p. 10–11, January–June, http://www.clarin.eu.newsletter, 2011.

PATRAŞCU M., HAJA G., CLIM M.-R., TAMBA E., *Romanian Dictionaries. Projects of Digitization and Linked Data*, in D. Trandabăţ and D. Gîfu (eds.), *Linguistic Linked Open Data. EUROLAN 2015*, Springer International Publishing Switzerland, p. 110–123, 2016.

TAMBA E., *La lexicografía Rumana. Historia y Actualidad*, in Rodríguez, F.C., Seoane, E.G., Palomino, M.D.S. (eds.) *Lexicografía de las lenguas románicas. Perspectiva histórica*, vol. I, de Gruyter Verlag, p. 265–282, 2014.