The evolution of the Romanian digitalized lexicography. The essential Romanian lexicographic corpus

Elena Tamba Dănilă, Marius-Radu Clim, Mădălin Pătraşcu & Ana Catană-Spenchiu

Keywords: *Romanian lexicography*, *computerized lexicography*, *linguistic resources*, *computerized lexicographic instruments*.

Abstract

The aim of this paper is to highlight the present stage of the digitalized lexicographic research from Romania and the importance of creating a Romanian Essential Lexicography Corpus. In the last years there have been taken measures for creating electronic instruments and resources that are necessary for supporting the Romanian language and culture on a transnational level, in the general context of the computerization of the fundamental academic research.

The Romanian academic specialists in linguistics and applied informatics, as well as in computational linguistics fields, have initiated research projects by which they want to valorise the non-digitized resources by acquiring them in electronic formats and to create new resources and instruments for the automatic processing of the language.

The project presented in this paper has as purpose the valorization of certain results from the complex project eDTLR, by using, as reference text for the alignment, the Thesaurus Dictionary in electronic format and creating a Romanian lexicographic corpus. This project's aims are: the realization of a scanned corpus, with the reference dictionaries of DLR (taking into account the present legislation regarding copyright); scanning and processing of these dictionaries (by OCR – optical character recognition – the conversion from image to text; parsing the text at entry); realizing an on-line interface for validating/correcting of the parsing (= automatic identification of the entries from previously scanned and converted dictionaries), as well as validating the alignment between the text of the Romanian Language Thesaurus Dictionary (in electronic format, from eDTLR project) and the reference dictionaries from DLR Bibliography. The final database will include an important number of essential Romanian language dictionaries (100 dictionaries from the 16th century to present day) aligned at entry level, fact that will offer Romanian specialists an excellent working instrument and will set basis for future research.

1. Introduction

One of the objectives of the European policy is the preservation and the exploitation of the national linguistic identity. In Romania, measures have been taken also for creating electronic tools and resources that are necessary for supporting the Romanian language and culture on a cross border level, in the general context of the computerization of the fundamental academic research.

2. The Romanian digitalized lexicography

The Romanian academic specialists in linguistics and applied informatics, as well as in the computational linguistics field, have initiated research projects by which they want to turn to profit the non-digitized resources by transposing them in electronic format and to create new resources and tools for automatically processing the language. Thus, on some previous projects, the tendency was to obtain the *Thesaurus Dictionary of the Romanian Language* in electronic format (14 tomes, 36 volumes, over 17.000 lexicon type pages, having between 7000 and 11000 characters / page), which constitutes, in a synthetic form, the Romanian spirituality manifested in language, under all its aspects, from the first writings to the present

day. That is why the elaboration of an electronic version, which would be accessible to the scientists and to all those who are interested in learning or studying the Romanian language became an absolutely necessary step to take in the digitalized, multicultural society.

A special place is held by the digitalization of the *Romanian Language Thesaurus Dictionary* that was obtained on the eDTLR project, which had been announced and prepared through a series of other grants that proved the feasibility of this idea and created new tools necessary for achieving the results. These grants had as purpose finding solutions for:

– obtaining the electronic version of the printed edition of DLR (*Dicţionarul limbii române în format electronic. Studii privind achiziţionarea* – 2003-2005). This project made possible the process of testing and providing evidence to prove the possibility of transforming the *Romanian Language Dictionary* from a printed text into an annotated electronic text.

– researching the lexical material from DA[1] and DLR[2] according to certain criteria by using computer means (*DLRI. Baza lexicală informatizată. Derivate* – 2007–2008). The current project contributed to the elaboration of a lexicographic sample composed of the Romanian language derivatives formed with the Latin suffix *–ime* and the old Slavic suffix *–işte*, of the old series of the DA and of the new series of the DLR, together with the technical-lexicographic unification of the articles from the DA and DLR through digitalized means.

– elaborating the electronic version of some bibliographic resources of the *Romanian Thesaurus* (*Resurse lingvistice în format electronic: Monumenta linguae dacoromanorum. Biblia 1688. Regum I, Regum II – Ediţie critică şi corpus adnotat* – 2006–2007). The current project contributed to finding a method for obtaining the electronic version of old books from the DLR Bibliography; this method is applicable to two books of the *Bible* that were printed in Bucharest in 1688, i.e. *A împărăţiilor cea dentâiu* (Kings 1)*, A împărăţiilor a doua* (Kings 2), together with the elaboration of a series of tools for indexing and automatically annotating the old Romanian texts, for each word.

These steps have prepared the complex project which is *eDTLR Dicţionarul tezaur al limbii române în format electronic* (*The Romanian Thesaurus Dictionary in electronic format*) (2007-2010), with the main objectives of achieving the complete version of *The Romanian Thesaurus Dictionary* in electronic format and elaborating a corpus which integrates all the texts of the *Dictionary*'s Bibliography (scanned and text versions), fact that will allow both a complex consultation of the *Dictionary*, and the editing and up-dating process.

For many years, the great European cultures had thesaurus dictionaries and text corpora in electronic format. For better understanding the dimensions of the *Romanian Thesaurus Dictionary*, we present several statistics, in comparison to other large European dictionaries: *The Thesaurus Romanian Dictionary* (two series: DA – 1907-1944, DLR – 1965-2009), 33 volumes, 160.000 words and their versions, over 1.300.000 quotes; first electronic version: 2007-2010; *Oxford English Dictionary* (*OED*, http://www.oed.com/) – first edition 1928, 20 volumes (second edition – 1989), 301.100 words, 2.412.400 examples; first electronic version: 1988; *Deutsches Wörterbuch "der Grimm"* (DWB, http://germazope.uni-trier.de/Projects/DWB: 1838-1961), 32 volumes, 350.000 words and their versions; first electronic version: 1997-2004; *Tresor de la Langue Francaise* (TLF), XIXth –XXth Century (http://atilf.atilf.fr/: 1971-1994 – first printed edition), 16 volumes, 100.000 words, 270.000 definitions, 430.000 examples; first electronic version: 1990-2004; *Diccionario de la lengua espanola de la Real Academia Espagnola* (DRAE, http://buscon.rae.es/draeI/): 1780 – first printed edition; the 23rd edition is still in work; 88.500 words, 161.962 examples; first electronic version: 1992.

Starting from the statistics mentioned above, we notice that the *Romanian Language Dictionary* can be integrated among other similar European dictionaries and its computerization became a normal step in the evolution of the Romanian lexicography.

3. CLRE. Essential Romanian Lexicographic Corpus

In this context, the project *CLRE. Corpus lexicografic românesc esenţial. 100 de dicţionare din bibliografia DLR aliniate la nivel de intrare* (*ERLC. Essential Romanian Lexicographic Corpus. 100 dictionaries from DLR Bibliography aligned by entries*) is a natural continuation of the projects related to the computerization of the *Romanian Language Dictionary*, also proving the capacity to exploit some of the results of the complex project eDTLR, project that has initiated a series of techniques and methodologies for the electronic acquisition and use of the great *Romanian Thesaurus*.

The current project continues and develops new work/study/research methods in the Romanian lexicography field, including its digitized side and it offers, beside the results of eDTLR, a modern way for finalizing and up-dating the great dictionary, the possibility of interactively consulting the dictionaries of the DLR Bibliography by any Romanian or foreign philologist/linguist/lexicographer and, why not, by any user of the Romanian language within or across its region.

During the elaboration of the CLRE, methods for on-line and local editing will be created / used, in order to extract new entry fields from the dictionary, with the purpose of including them into a database, for searching dictionary entries within biographical sources, indexing methods in scanned documents (picture of the original page).

The project results and especially the elaboration of a corpus in which the alignment is to be done at an entry level will allow the development of vast applications regarding the semantic of words and entry selections in order to elaborate new specialized dictionaries (etymologic, semantic etc.), the correlation with other linguistic or media resources, fact that would take Romanian lexicography at a level close to the European lexicography (see *Le rayon des dictionnaires*, http://www.atilf.fr/ – a collection of digitized French dictionaries, from the XVIth to the XXth century or *Nuevo tesoro lexicográfico de la lengua española*, http://buscon.rae.es/ntlle/SrvltGUILoginNtlle – the database containing the facsimiled versions of all dictionaries edited and published by Real Academia Espagnola).

This project has the following purposes: achieving a scanned corpus, with the reference dictionaries of DLR (taking into account the current copyright legislation); scanning and processing these dictionaries (by OCR – optical character recognition – the conversion from image to text; parsing the text at entry); achieving an on-line interface for validating/correcting the text after parsing it (= automatic identification of the entries from previously scanned and converted dictionaries), as well as validating the alignment between the text of the *Romanian Language Thesaurus Dictionary* (in electronic format, from the eDTLR project) and the reference dictionaries from DLR Bibliography.

The CLRE project will include three types of specific activities: 1. elaborating techniques for digitizing the dictionaries of the DLR Bibliography, a software for identifying the fields of a dictionary entry, aligning and organizing them into a database, an interface which would allow the correction and searching through this aligned corpus – activities which would be carried out by the IT specialist; 2. lexicographic activities comprising of the transliteration of the title-words from the dictionaries written in Cyrillic and transition alphabets and also the activity of validation for the final alignment; 3. disseminating the final product – activity carried out by all the researchers in the project

The project result will be published on-line.

Also, this project will have both classic/traditional linguistic methods (for example, transliterating the entries in Cyrillic or in the transition alphabet or comparatively studying the dictionaries at a semantic level), as well as new, lexicographic-computational methods.

4. Conclusions

The aim of this paper is to point out the current stage of the digitalized lexicographic research in Romania, together with the validation / correction programme through which we can check each entry for the different types of dictionaries (especially for the ones that are written in Cyrillic or that are using the transition alphabet), at a particular level.

The final result of this project is an essential Romanian lexicographic Corpus, which will include an important number of essential Romanian language dictionaries, aligned according to their type, fact that will offer the Romanian specialists an excellent working tool and will set the basis for future research.

Notes

[1] DA is the acronym for *Dictionarul Academiei (Academia Dictionary)* the first part of the *Romanian Academic Thesaurus*, published between 1907-1944.

[2] DLR is the acronym for *Dictionarul Limbii Romane (Romanian Language Dictionary)*, the second part of the *Romanian Academic Thesaurus*, published between 1965-2009.

References

A. Dictionaries

DA. ***Dicţionarul limbii române*** **1907-1944.** Tom I-II, Tipografia ziarului „Universul”, Bucureşti, Imprimeria Naţională.

DLR. ***Dicţionarul limbii române*** **1965-2010.** Serie nouă, tom VI-XIV, Bucureşti, Editura Academiei.

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

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

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

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

B. Other literature

Haja, G., E. Dănilă, C. Forăscu and B.-M. Aldea 2005. *Dicţionarul limbii române (DLR) în format electronic. Studii privind achiziţionarea*. Iaşi: Editura Alfa.

Haja, G., C. Forăscu, B.-M. Aldea and E. Dănilă 2006. ‘The dictionary of Romanian Language: steps toward the electronic version.’ In E. Corino, C. Marello and C. Onesti (eds.), *Atti del XII Congresso Internazionale di Lessicografia : Torino, 6-9 settembre 2006*. Alessandria: Edizioni dell'Orso.