Towards a Romanian Lexicographic Corpus

Elena DĂNILĂ, Marius-Radu CLIM,
Ana CATANĂ-SPENCHIU*

Key-words: *lexicography, computerized lexicography, linguistic resources, computerized lexicographic instruments*

Great European cultures have had, for many years, thesaurus dictionaries and texts corpora in electronic format.

In this context, our project, *CLRE. Corpus lexicografic românesc esențial. 100 de dicționare din Bibliografia DLR aliniate la nivel de intrare şi la nivel de sens* (*CLRE. Essential Romanian Lexicographic Corpus. 100 Dictionaries from DLR Bibliography Aligned by Entries and Meaning*) is a natural continuation of the projects which dealt with the digitizing of the *Romanian Language Dictionary*, also proving the valorisation capacity of some of the results complex project e DTLR, project that has initiated a series of techniques and methodologies for the electronic acquisition and use of data of the great *Dictionary*.

This project's aims: 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.

In order to achieve the objectives of the project it was necessary to use advanced equipment to facilitate the acquisition of electronic dictionaries and also software used in processing scans, character recognition so it could be allowed the smooth implementation of the database. Further on we will detail some of the equipment and software used in the project.

First it was purchased a special scanner for books, Atiz Book DIY[1]. This proved to be the best solution for the digitization of books, in terms of costs and efficiency.

* "A. Philippide" Institute of Romanian Philology, Iași, Romania.

This paper was realized in the project *CLRE. Corpus lexicografic românesc esențial. 100 de dicționare din Bibliografia DLR aliniate la nivel de intrare şi la nivel de sens*, CNCSIS-UEFISCSU, code TE_246/2010, 2010–2013.

The efficiency of this product is due to the performance of digital cameras SLR ("single-lens reflex") and the ingenuity of a unique v-shaped, auto-adjusting book cradle and platen to capture sharp images at up to 700 pages an hour. The Atiz scanner has two cameras Canon EOS 450D with 35 mm lens. The EF 35 mm lens allow a better focus and they are specially used for A3 or A2 book format.

The conventional flatbed book scanners have many disadvantages. First of all, they cause damage to books as a result of applying force in an attempt to flatten pages. Secondly, the manipulation of books is more difficult and requires a human effort at the same time. Another disadvantage is the curvature of the page which leads to numerous errors in character recognition program.

BookDrive DIY has a great advantage to produce sharp images with no page curvature and it prevents books from being damaged. The scanner has two powerful digital cameras, mounted either side of the book cradle in order to produce sharp, flat-looking images of the pages. The book is placed facing up at 120°on the v-shaped cradle. So, the bending of the page is removed and the spine of the book is no longer forced and damaged and the productivity is much higher in the end. By engaging the transparent platen there will be spared time and energy used in manipulation of the book and every single page is preserved. BookDrive DIY can be used to scan all kinds of books and can facilitate many different sizes, thicknesses and types of bindings. It should be added that this scanner is best use in processing old books which require attention.

Atiz scanner is available with two programs used in capturing and processing images. Every BookDrive comes with BookDrive Capture and BookDrive Editor Pro software. So BookDrive Capture is the application that controls the cameras. It supports a wide range of Canon EOS SLR cameras and allows you to change camera settings from directly within the software (e.g. shutter speed, aperture and ISO values). With a simple click the cameras capture both left and right pages at the same time. Each shot takes less than a second to photograph. The program allows the user to insert the missing pages which have been skipped or if some shots look bad to replace them with new ones.

After the scanning of a book, scanned images are converted using another program, BookDrive Editor Pro. With this program the scanned pages are processed and transformed in PDF (single-page or multi-page files), TIFF (LZW and CCITT Group 4, single-page or multi-page file), and JPEG formats fit for distribution or archiving and ideal for OCR text conversion. This program replaces an unwanted tinted page background common in old books with a bright and clear background free from speckles and ink stains. Other features include rotation, de-skew, crop, auto level, brightness and contrast adjustment, sharpen, black border removal, image resize and DPI adjustment and also the saving is made for different formats for each page or in folders for more pages.

For recognition accuracy and text conversion capabilities it was purchased an ABBYY FineReader Engine, which includes an optical character recognition program (OCR), an intelligent character recognition, an optical mark recognition (OMR), a barcode recognition (OBR), a document imaging, and PDF conversion.

[1] More information about this product is available on the website http://diy.atiz.com/.

This program turns scans, PDFs and digital photographs into different searchable and editable documents. We can add that this program has a conversion utility that instantly turns the different elements of formatting like content, titles, footnotes, page number, headings are saved into various electronic formats, including Microsoft Word, Excel and PDF, because of the ADRT (Adaptive Document Recognition Technology) for intelligent reconstruction of the logical structure and format documents.

All this equipment and computer software facilitates the accurate processing of lexicographical material in question.

CLRE - Information Indexing

For the first phase of the project mentioned above, which involves aligning entry-level dictionaries it is necessary to store in electronic format the lexicographic resources in order to establish connections. This operation was performed in three steps:

1. Dictionary scanning with a vertical edge scanner, which gives extra quality results;

2. Applying OCR (Optical Character Recognition) program on the obtained images. The process involves recognition of graphic signs and their electronic storing. For a more efficient version it has been used ENGINE Abby Fine Reader program;

3. Storing all this information in a database, both in image format, as well as the structured alphanumeric.

CLRE – Entry identifying

Due to the graphical format diversity of dictionaries in work, it was engaged a less traditional idea involving giving up writing parser versions.

To solve this problem we used machine learning concepts to delineate definitions. This is possible by using clustering algorithms (cluster = group of objects with similar features), applied according to a graphical specified criteria for groups of dictionaries with similar graphic format. There are built this way geometrical shapes that include constituent words of a definition.

The present project continues and refines the new work/study/research methods in Romanian lexicography, including its digitized side, offering, in completion of the results of eDTLR, a modern way for the completion and bringing up-to-date of the great dictionary, the possibility for interactively consulting the dictionaries from DLR Bibliography by any Romanian or foreigner philologist/ linguist/ lexicographer and, why not, by any Romanian language user within its area or not.

The specific elements of the project are also obvious in using new methods of work and Romanian lexicographic research, including its computational side, thus a modern way for the completion and up-to-date bringing of the great dictionary being offered, also giving the possibility for interactively consulting the CLRE corpus by any Romanian language speaker from the area or outside it.

Thus, this project will have both classic/ traditional linguistic methods (for example, transliterating the entries in the Cyrillic or transition alphabet or the comparative, semantic level, study of the dictionaries), as well as new, lexicographic-computational methods.

The results of the project 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 words' semantic unambiguousness, types and entries selections with the purpose of elaborating new specialized dictionaries (etymologic, semantic etc.), the correlation with other linguistic or media resources, fact that would take Romanian lexicography at a level which would be close to European lexicography (see *Le rayon des dictionnaires*, http://www.atilf.fr/ – a collection of digitized French dictionaries, from the 16th to the 20th 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). The results of our project will be available on-line too.

In order to find out about the study of the lexicography computing, two of the members of the team participated in the Lexicom 2011 *Workshop in Lexicography and Lexical Computing* – Sankt Petersburg, Rusia, 14–18 Juin 2011.

Lexicom is a five-day intensive workshop in lexicography and lexical computing created by the Lexicography MasterClass, where you can learn how to create dictionaries and other lexical resources, from the preparation of corpora to the writing of entries. Seminars on theoretical issues alternate with practical sessions at the computer, working in small groups and individually.

Tutorials were led by: Adam Kilgarriff, Simon Krek, Jan Pomikalek, Anna Rylova and Liz Walter.

The topics of the LEXICOM were:

– designing and building text corpora

– installing corpora into the Sketch Engine and building word sketches

– dictionary databases

– corpus analysis: discovering word senses, recording contextual information

– writing entries

At Lexicom 2011 there were presented some new lexicographic resources and methods designed to facilitate the lexicographers' work.

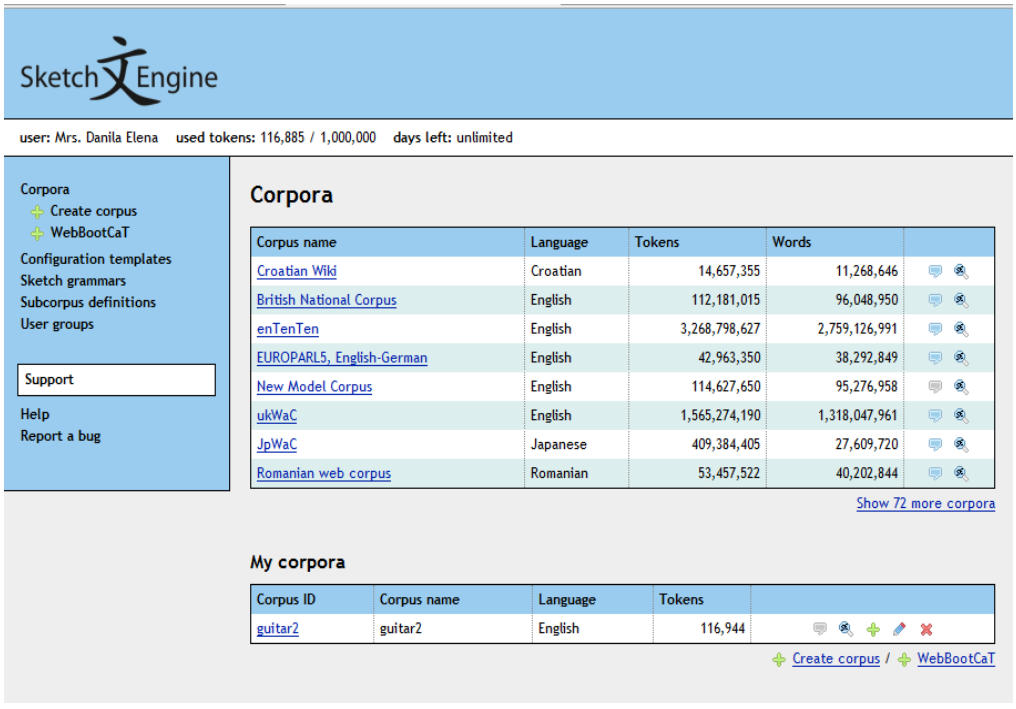
Sketch Engine (http://the.sketchengine.co.uk/)

It is a Corpus Query System incorporating word sketches, one-page, automatic, corpus-derived summary of a word's grammatical and collocational behavior.

The Sketch Engine is a web-based program which takes as its input a corpus of any language with an appropriate level of linguistic mark-up. The Sketch Engine has a number of language-analysis functions, the core ones being:

– the Concordancer – a program which displays all occurrences from the corpus for a given query. The program is very powerful with a wide variety of query types and many different ways of displaying and organizing the results.

– the Word Sketch program – this program provides a corpus-based summary of a word's grammatical and collocational behavior.



The program aims to generate Word Sketches (a brief analysis of a word, consisting of grammar and collocation behavior description, performed on a corpus).

With Word Sketches and concordances[2], Sketch Engine provides support for identifying the correct meaning of a word in a certain context.

At this moment, Sketch Engine offers other facilities, such as Word List or Thesaurus (described in Chapter 1.1.4.2 and 1.1.4.4). In order to be used it is necessary to have a formatted corpus as specified by Sketch Engine and a corresponding set of rules. Software package is available at http://www.sketchengine.co.uk/ and can be used by anyone who creates a user account (with limited functionality and limited term, for the free version).

There were also mentioned several dictionary writing systems that represent software which allow the lexicographer to compile and edit the dictionary entries, to organize teamwork with other lexicographers, but also to print the dictionary in a certain format.

These programs were created after 1990 and were heavily supported by the publishing houses developed specially for the purpose of dictionary writing.

Few of these programs are freely available (e.g. Lexique Pro http://www.lexiquepro.com/ and MyWords http://flex-aspnet.blogspot.com/2008/01/ mywords-dictionary-writing-system-flex. html), most of them were created by academic research institutions or by specialized companies. Some of these dictionary writing systems are: DPS (IDM developed in France), Lingvo Content

[2] The concordance means the organization as a dictionary of all words used by a particular author, with the indication of all the word occurrences and the contexts of these occurrences.

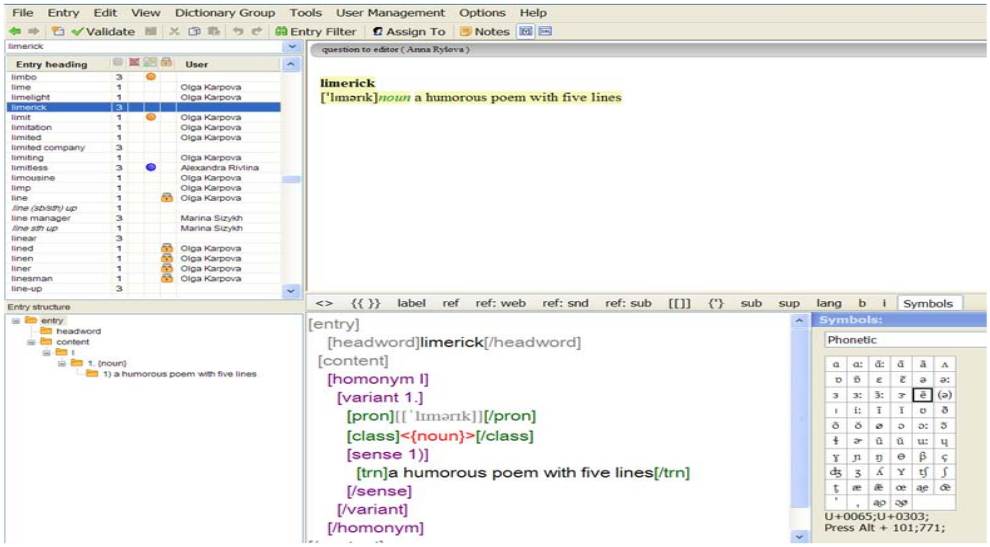
(by ABBYY, the Russian Federation) and Ilex (designed by Erlandsen Media Publishing in Denmark).

Generally, such systems have three components: a database which stores the dictionary itself, the bibliographic resources, a working interface for editing dictionary entries and dictionary tools for viewing in different formats.

The database includes all dictionary entries, whatever form they have, as well as dictionaries or other resources needed to achieve the desired dictionary. This database resides on a server which stores all the work done by lexicographers and offers also a history of each entry: how many interventions were made on the text, who is the author, the reviewer and how much time was required for editing.

The editing interface facilitates the structuring of the entries in the same way, allowing automatic checking of the structure of a definition, making automatic correlations, but also various dictionary text statistics. With the interface it can be selected for each editor separately the entries that have to be done and it can be checked at any time the status of the whole dictionary. The interface, as shown in picture, may contain four types of views. It can display the full list of words that can be generated separately and can be used for other linguistic researches.

The editing window allows the lexicographer to write each entry in part using the given structure. Writing is automatically converted into XML and the lexicographer can see the tree-structure of each entry. By this view the structure of the definition comes out more clearly and an editor can move easily from one place to another a sub-meaning and all the links are done automatically again. Also, the lexicographer can see how it will look in the dictionary the entry edited by him. The dictionary format visualization tools can determine the design for each type of dictionary is required. Therefore, such a database can create smaller dictionaries, reversed dictionaries or different dictionaries can be compared and can be watched both in printed format and also with other electronic means.



The ABBYY Lingvo Content interface

Based on information received and the practical demonstration that the two team members attended there were gathered a number of data needed to achieve the objectives of the project CLRE. Also there was presented particularly the CLRE project, and the computerized activities of the Romanian lexicography generally, establishing links with specialists from different countries.

The final result of this project is an essential Romanian lexicographic Corpus, which will include an important number of essential Romanian language dictionaries, formally and semantically aligned, fact that will offer Romanian specialists an excellent working instrument and will set basis for future research.

Bibliography

DA = *Dicționarul limbii române*, tome I-II, Tipografia ziarului „Universul", Bucureşti, Imprimeria Națională, 1907–1944.

DLR = *Dicționarul limbii române*, Serie nouă, tome VI–XIV, Bucureşti, Editura Academiei, 1965–2010.

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

Abstract

This paper presents the study of the research in the project CLRE. In order to achieve the objectives of the project it was necessary to use advanced equipment to facilitate the acquisition of electronic dictionaries and also software used in processing scans, character recognition so it could be allowed the smooth implementation of the database. Therefore we present these hardware and software. We also present some lexicographic resources and methods that some members of the research team studied in LEXICOM 2011.