The Dictionary of Romanian Language: Steps toward the Electronic Version

Gabriela Haja[1,2], Corina Forascu[2,3], Bogdan Mihai Aldea[2], Elena Danila[1]
1 Institute of Romanian Philology "A. Philippide", Iasi branch of Romanian Academy
2 Faculty of Computer Science, University Al.I. Cuza of Iasi
3 Research Institute for Artificial Intelligence of the Romanian Academy
16, Gen. Berthelot, 700483-Iasi,
Romania

Abstract
In the context of the globalised Information Society and the variety of solutions for computer-aided acquisition of traditional dictionaries, the paper presents the actual stage of development of the new series of the Romanian Dictionary edited by the Romanian Academy. Through a project financed by the National University Research Council of Romania, some preliminary steps toward a computer-aided acquisition of the dictionary have been made and are outlined in this article.

1 Introduction

The Dictionary of Romanian (DLR), among dictionaries of Romance languages, is comparable with Nicot's *Trésor de la Langue Française*. Till now, DLR has been elaborated the traditional way: the lexicographers have defined for each dictionary entry, through manual work, those senses that can be found in handwritten excerpt cards. In order to improve the work-flow in the near future and to create a basis for new, updated, unified versions of DLR, a method for computer-aided acquisition was developed within a project financed by the National University Research Council of Romania.

The next section presents some computerized dictionaries of Romance languages, which have similar structure/profile with DLR. After a brief description of DLR, the advantages of an electronic version of DLR are discussed. The paper concentrates then on the solutions adopted for the computerized acquisition and use of the Romanian dictionary. A tool specially created for the acquisition of the Romanian dictionary, DLRex, is sketched. The last section outlines some work perspectives in Romanian lexicography by using computerized linguistic resources and tools.

2 Computerized dictionaries

For all communities, compiling dictionaries of their traditional languages is among the most important tasks of cultural preservation since the dictionary is a storehouse of all the special ways of talking about their culture and environment. While at least 126 languages do

have a kind of electronic dictionary,[1] only few of them have a thesaurus dictionary in an electronic format.

Among the Romance languages, French is one of the most "present" on the Internet, the *Trésor de la Langue Française informatisé*[2] (TLFi), including 100.000 words with their histories, 270.000 definitions, and 430.000 examples. The 16 volumes of TLF were computerized by the compilation of a reliable archive containing the text of the printed version of TLF; the retroconversion into a structured text in which the various subjects of the articles (definitions, quotations, synonyms and antonyms, and semantic, etymologic, grammatical, stylistic and historic indicators) are delimited; and the development of the software to interrogate the structured text at 3 different consultation levels.

The historical dictionary of the Italian language, *Tesoro della Lingua Italiana delle origini (TLIO)*, is based on the *Opera del Vocabolario Italiano (OVI)*, a database with 1849 vernacular texts (21.2 million words, 479.000 unique forms) that could be interrogated through GATTO (Iorio-Fili, 1997), a lexicographic software created at CNR. A lot of other research projects have benefited from the OVI database (Dupont, 2001).

The *Diccionario de la Lengua Española*[3] is continuously updated by the Real Academia Española and 21 Hispano-American academies based mainly on the synchronic and diachronic textual databases and the historical file of the Spanish Academy. For Portuguese more than 95.000 entry words, 13.000 verb conjugations, among others, are available in *Língua Portuguesa On-Line.*[4]

The Romanian language has an explanatory dictionary available on-line,[5] collected from the original version of the *Romanian Academy Explanatory Dictionary* (DEX, 1998), dictionaries of synonyms and antonyms and web-volunteers.

3 The dictionary of Romanian Language

3.1 Current status

Three institutes of the Romanian Academy (Institute of Linguistics „Iorgu Iordan – Al. Rosetti" – Bucharest, Institute of Romanian Philology "A. Philippide" – Iasi and Institute of Linguistics „S. Puşcariu" – Cluj-Napoca) are finishing nowadays this monumental work, the 'Dictionary of Romanian Language. New series' (*DLR*). Some statistics performed on DLR are illustrated in Figure 1. Started in 1959 (Seche, 1969), the DLR includes now 22 printed volumes, with more than 10.000 pages containing 18 letters (out of 28) from the Romanian alphabet. The 26 volumes intended to be created till the end of the project will have the letters *D, E* and *M-Z.* One previous edition of the Romanian Academic Dictionary, started by Al. Philippide at the beginning of the previous century, was continued in an academic collec-

[1] http://www.lexilogos.com/

[2] http://atilf.atilf.fr/

[3] http://buscon.rae.es/diccionario/drae.htm

[4] http://www.priberam.pt/dlpo/dlpo.aspx

[5] http://dexonline.ro/

tive guided by S. Puscariu. Between 1906 and 1944 almost half of the whole dictionary was edited: 2.956 pages with the letters *A-C*, *F-K* and a part of *L* (DA, 1913; Seche, 1969). The dictionary was created using the traditional lexicographic technology till the nineties, when the information and publication had started being introduced by the aid of computers. Using computers, not only the efficiency has grown, but also the possibility to use the electronic versions of the texts for computer-aided creation of new versions of the dictionary.

Publication year	Tom	Part	Letter	No. of pages	No. of word entries
1965 – 1968	VI		*M*	1076	9653
1971	VII	1	*N*	548	5493
1969	VII	2	*O*	400	3622
1972	VIII	1	*P*	357	4006
1974	VIII	2	*P*	336	3784
1977	VIII	3	*P*	255	2738
1980	VIII	4	*P*	393	4558
1984	VIII	5	*P*	525	4692
1975	IX		*R*	641	7255
1986	X	1	*S*	388	3540
1987	X	2	*S*	300	2212
1990	X	3	*S*	349	2692
1992	X	4	*S*	371	2757
1994	X	5	*S*	721	5742
1978	XI	1	*Ş*	271	4528
1982	XI	2	*T*	376	5027
1983	XI	3	*T*	387	4217
1994	XII	1	*Ţ*	240	3856
2002	XII	2	*U*	468	2347
1997	XIII	1	*V*	325	1747
2002	XIII	2	*V*	426	2396
2000	XIV		*Z*	409	4088
				9562	**90950**

Figure 1. The Dictionary of Romanian Language – statistic

3.2 DLR – general and explanatory dictionary of Romanian

As a general dictionary, DLR explains all the words which are attested in popular, literary and artistic speech (DLR, 1965). Special terminologies are present only if they are attested in at least two linguistic styles. DLR has a historic character as it contains all possible regionalisms, archaisms and popular technical terms (Sala, Mihaila, 2000). Only the argotic terms used in familiar or artistic speech were included; the same rule applies to the personal creations of Romanian authors or if they are used in the general literary terms. The derivatives and compounds have separate entries. The homonyms are grouped under their etymon. The words used in specialized languages (children games, riddles, magic spells), phrases, expressions, proverbs and sayings are listed only if they are explained in their original sources.

DLR is an explanatory dictionary as the words are defined and explained for all their senses, no matter their frequency or geographical area of provenance or use (DLR, 1965).

The linguistic facts are defined through *genus proximus* and *differentia specifica*, by delimiting the semantic values through synonyms, by indicating the syntactic functions for the words being grammatical instruments. The variety and complexity of the factors determining the semantic evolution of the words is reflected in the lexicographical techniques used to group word senses (DLR, 1965). The principle of the etymon is used when establishing the order of senses. Even though the first written attestations of a word have semantic values developed in Romanian, not representing the initial meanings of the etymon, still the semantic evolution of the word starts from the sense of the etymon. A different perspective appears with the neologisms, for which the sense used in the first Romanian attestations is listed as the first one.

3.3 A computerised DLR – a necessity and a goal

Acquiring DLR in electronic form constitutes a solution to all the problems of re-editing the whole Romanian language dictionary, which implies the unification and updating of the material from DLR – new and old series. The computer-aided acquisition of DLR has many advantages:

- it will be possible to have a broad database of texts/attestations which can be used to update the articles in the old series of the dictionary (published between 1905 and 1949) and those of the first volumes of DLR, as DLR contains about 3.200.000 examples, representing about 88% from the whole text (Vintila-Radulescu, 2002);
- additions and corrections can be performed directly in the electronic text of the DLR, without retyping the text, as has been the practice with the tomes published after 1990;
- it will be easier to notice and correct the inconsistencies/differences between different tomes of DLR, as they came up as a normal consequence of the fact that three different generations of editors/lexicographers have worked during time in three different centres;
- it will be easier to extract defining models for standard entries, such as certain geological terms, months and days of the year, the name of plants and animals or for certain parts of speech, such as numerals or pronouns; these models can be used to create a software interface that can assist the lexicographer when completing the fields of such an entry;
- it will be possible to align word senses from other dictionaries, which is very useful in Word Sense Disambiguation and Machine Translation; such an alignment was performed between the DLR and the Romanian WordNet (Tufis et al., 2004), for a part of the letter "V": for 73 common word entries, DLR contains 2.300 senses while the Romanian WordNet contains 100 corresponding synsets (Tanasescu, 2004).

4 DLRex – a tool for computerised creation and use of DLR

In the framework of the research project (Haja et al., 2005), a dedicated tool was developed, DLRex, as an instrument for acquiring, processing and browsing the electronic version of the Romanian Language Dictionary. The main functionalities of DLRex are the followings: converts the text from Word (RTF) format into XML format; creates the XML files for the entire DLR; allows browsing and interrogating the XML files of the DLR; permits to update and unify the DLR.

After scanning the printed files of the dictionary, an OCR is used to obtain the texts which are further corrected by lexicographers and/or linguists (which are familiar with the

standard format of DLR). A Word document in .doc format is obtained this way. Since the title abbreviations used for the references in DLR are very important and recognizing them is very helpful when parsing the files of DLR, the Java programming language was used, as it handles very well the strings of characters. In order to improve the conversion to XML, the Word files were saved in RTF format.

When parsing the DLR the formatting of the text and the presence of special symbols are very important: the ♦ symbol introduces a closely related sense, whereas the ◊ symbol introduces a more distant one. Both symbols can appear randomly in a word entry, as reproduced in Figure 2.

After reading an entry, a vector is built in order to keep each fragment with different format with respect to the fragments immediately before and after it. Taking into consideration one entry, one parsing of the vector would give the desired XML corresponding file. But some inconsistencies came up: fragments with the same format in the dictionary, hence with the same "meaning" in parsing, were saved in the RTF file with different formatting. Moreover, some characters (white space, quotation marks, periods and commas) do not maintain their formats mainly when used in fragments written in Boldface or Italics, due to the scanning process.

When first trying to parse the vector, another problem aroused: at the end of each word entry there is a list with the etymology, pronunciation of the word, in which the formatting is, again, different from one entry to another. In order to avoid all the problems mentioned above, that can cause errors in parsing, the vector was pre-processed before the parsing so to have a more restricted form. The vector is then parsed, based on the succession of the formatting/styles used for each fragment and it takes into consideration the special situations that have been recorded while testing and evaluating the DLRex on a very broad sample of DLR.

WORD part of speech, sense definition [*example illustrating the use* TITLE ABBREVIATION]
A sense definition [*example illustrating the use* TITLE ABBREVIATION]
I sense definition [*example illustrating the use* TITLE ABBREVIATION]
1 sense definition [*example illustrating the use* TITLE ABBREVIATION] ◊...
a) sense definition [*example illustrating the use* TITLE ABBREVIATION]
b) sense definition [*example illustrating the use* TITLE ABBREVIATION]
.....
2) sense definition [*example illustrating the use* TITLE ABBREVIATION]
.....
II sense definition [*example illustrating the use* TITLE ABBREVIATION] ◊... ♦
.....
III sense definition [*example illustrating the use* TITLE ABBREVIATION]
.....
B sense definition [*example illustrating the use* TITLE ABBREVIATION]
.....
C sense definition [*example illustrating the use* TITLE ABBREVIATION] ♦...
.....
- Grammatical, orthoepical information ...
- Etimological information

Figure 2. A word entry in DLR – general scheme

As the DLR volumes were written in almost 50 years by different authors, the DLRex will be continuously improved so that parsing the whole DLR can have a bigger success rate. This way there will be no need for many interventions from the specialists to treat manually specific cases.

By means of DLRex, the dictionary can be interrogated in different ways: it is possible to search among word entries, among the senses or for specific examples.

5 Perspectives in the Romanian lexicography

The initial objective, to define a lexicographical grammar for DLR, has proved to be almost impossible to be carried out under the given circumstances (restrictions of time, technical means) and it was postponed. After the complete acquisition of the XML format of DLR, by using improved versions of the DLRex, all the problems due to the editing norms of DLR will be discussed among specialists so that the best formalization of the entries in DLR to be adopted.

The instrument created in the framework of the project, DLRex, raises encouraging possibilities. It will be possible to obtain the full computer-aided acquisition of the DLR and, after that, the acquisition, based on an analogous heuristics, of the old series of the same dictionary, edited under the coordination of Sextil Puscariu in the first half of 20th century. Finally, it will be possible to fully update the DLR, from *A* to *Z*, and to create future editions with the regularity and efficiency that current possibilities allow, and in accordance with the advanced researches in the field of computational lexicography. This goal has to be achieved simultaneously with the creation of an electronic corpus of Romanian texts, updated continuously as the language advances, large enough and efficient for compliance with the scientific standards of this fundamental work of the Romanian culture, the Dictionary of Romanian Language.

References

A. Dictionaries

Dictionary of Romanian Language. New Series. (in Romanian: *Dictionarul limbii romane. Serie noua*) (DLR), Romanian Academy Publishing House, Bucharest, Romania, 1965-2002, 22 vol.

Romanian Explanatory Dictionary (in Romanian: *Dictionarul explicativ al limbii romane*), Romanian Academy, Univers Enciclopedic Publishing House, Bucharest, Romania (first edition 1984, second edition 1996).

Romanian Dictionary (DA, 1913) (in Romanian: *Dictionarul limbii romane*). With the recommendation and financial help of HM King Carol I of Romania, Socec&Comp. si C. Sfetea Libraries, Bucharest, Romania.

Trésor de la Langue Française (T.L.F.), CNRS, Gallimard, 1971-1994. 16 vol.

B. Other Literature

Dictionary of Romanian Language. New Series. (DLR, 1965), 'Introductory Notes'. *Dictionary of Romanian Language. New Series. Tome VI, Letter M*, Scientific and Encyclopedic Publishing House, Bucharest, Romania.

Dupont, C. (2001), 'The Opera del Vocabolario Italiano Database: Full-Text Searching Early Italian Vernacular Sources on the Web', *Italica* 78:4, pp. 526-39.

Haja, G., Danila, E., Forascu, C., Aldea, B.M. (2005), *The Dictionary of Romanian Language in electronic format. Acquisition studies*, Alfa Publishing House, Iasi, Romania.

Iorio-Fili, D. (1997), 'Un nuovo software lessicografico: GATTO', *Bollettino dell'Opera del Vocabolario Italiano* 2, pp. 259-70.

Sala, M., Mihaila, G. (2000), 'Foreword' in *Dictionary of Romanian Language. New Series. Tome XIV, Letter Z*, Scientific and Encyclopedic Publishing House, Bucharest, Romania.

Seche, M. (1969), 'History of the Romanian lexicography – Outline', in *Romanian: Schita de istorie a lexicografiei romanesti*, vol. II, Scientific and Encyclopedic Publishing House, Bucharest, Romania.

Tanasescu, V.I. (2004), 'Aligning electronical lexical resources. Applications on DLR and Romanian WordNet'. MSc Thesis, Faculty of Computer Science, University Al.I.Cuza, Iasi, Romania.

Tufis, D., Cristea, D., Stamou, S. (2004), 'BalkaNet: Aims, Methods, Results and Perspectives. A General Overview', *Romanian Journal on Information Science and Technology*, Tufis, D. (ed.) *Special Issue on BalkaNet*, Romanian Academy, 7(1-2), pp. 9-34.

Vintila-Radulescu, I. (2002), 'Romanian linguistic resources developed at the Institute of Linguistics „Iorgu Iordan"', in Tufis, D., Filip, F.G. (eds.), *Romanian Language in the Information Society – Knowledge Society*, Research Institute for Artificial Intelligence of the Romanian Academy, Bucharest, Romania.