# THE DIGITAL FORM OF THE THESAURUS DICTIONARY OF THE ROMANIAN LANGUAGE

Dan CRISTEA[1, 2], Marius RĂSCHIP[1], Corina FORĂSCU[1, 3], Gabriela HAJA[4], Cristina FLORESCU[4], Bogdan ALDEA[1, 4], Elena DĂNILĂ[4]

[1] Faculty of Computer Science, "Al. I. Cuza" University of Iaşi, Romania
[2] Institute for Computer Science, Romanian Academy, Iaşi, Romania
[3] Institute for Artificial Intelligence, Romanian Academy, Bucharest, Romania
[4] Institute of Romanian Philology "A. Philippide", Iaşi branch of Romanian Academy, Romania

The paper argues in favor of an electronic form of the thesaurus dictionary of the Romanian language, the dictionary edited by the Romanian Academy in two editions since 1913. Preliminary steps like scanning, optical character recognition, and pre-processing operations have already been done. The paper presents a prototype for the correction of the digital form of the dictionary. The numerous advantages of the digital thesaurus dictionary are discussed, as a basis for future work in Romanian lexicography and, more generally, in language processing.

*Key words:* Computational lexicography, Romanian language, dictionary, thesaurus, linguistic resources

## 1. INTRODUCTION

The modern society seems to be more and more attentive to its treasures that come from the past and that include all kind of cultural heritage, arts and language. Preserving it has become synonymous with preserving the cultural and linguistic identity of the nations. Even if at first sight the impression could be opposite, this concern is not at all weaker in communities that tend to form large political and economical organizations, like the European Union. Here, keeping alive individual languages has become extremely important in order to maintain the multitude of differences. Unification has nothing to do with flattening the cultural and linguistic differences. On the contrary, the more individual languages are preserved, the richer, more powerful and interesting a large union, as the Europe of nowadays, becomes. The linguistic identity of a nation relies heavily on the written literature and press of its language, and in the greatest degree, on its linguistic thesauri. These should be the recorders of the lexical and semantic stock in all phases of the evolution of the language.

There are no more than few tenths of years since the humans use the computer in printing processes. The intervention of the computer in printing has brought, beyond incredible gain in speed and accuracy, the side effect of producing the electronic form of the original paper document. A document in electronic format, be it a piece of news, literature, real science or anything else, is more valuable than its (only) printed format, as it is more easy to work with, to improve, to share, or to analyse. However, the greatest part of the human knowledge is still unavailable in electronic form, for commercial reasons, for restrictions encumbered by intellectual property rights (IPR), but also because of ignorance and lack of awareness of the tremendous benefits the electronic form of documents might bring. But more and more voices raise the problem of acquiring the writings of the humanity in electronic form (see, for instance, the Gutenberg project[1], the Google book search[2], or the DARPA project GALE[3]). Although, still not yet feasible with the present

[1] http://www.gutenberg.org/wiki/Main_Page
[2] http://books.google.com/
[3] http://projects.ldc.upenn.edu/gale/overview/

technology, when this will become a reality and means to interpret the semantic content of texts will be found, then the intelligent, content-based search will be possible and a tremendous part of the thoughts of the humanity, that actually reside only in the shelves of the libraries will become part of any project in which man and computers co-operate.

The word stock of a language, usually described in dictionaries and thesauri, should be kept updated with the language evolution. In some countries, preoccupations for this enterprise are usually being made in major linguistic institutions, under the prestigious patronage of academia and universities.

The big thesaurus dictionary of the Romanian language, edited by the Romanian Academy, built over more than one century, is planned to be finalized in 2007. The dictionary appeared in two series: the Dictionary of the Academy (DA), published between 1913 and 1949 and including the entries *A-C*, *D-De*, *F-K*, *L-Lojniță*, and the Dictionary of the Romanian Language (DLR), including the remaining entries. There are many reasons for digitizing this important work, among the most important being: the need to open it to the most largest audience (due to dimensions and price, presently, copies are being kept only in the specialized linguistics Institutes of the Academy and few national libraries), the need of being updated (due to the different periods in which the acquisition of terms has been performed, discrepancies appeared between the first worked-on letters in the first series and the subsequent editions), and the need of being used for language processing (to give just one example, the significant set of examples which accompany each sense of a word, configures a valuable sense-annotated corpus on which programs can be trained to disambiguate, using statistical means, semantically ambiguous words).

This paper is a programmatic document that advocates for an electronic form of this outstanding piece of cultural heritage, describes the work that has been recently initiated for the digitization of DA-DLR and the subsequent processing operations over the digital form. In order to make the distinction among different parts and formats, we will use the following abbreviations, supplementary to those already mentioned (DA and DLR): eDA and eDLR to denote the electronic version of the DA and, respectively, DLR, DTLR for the union of DA with DLR, and eDTLR (Dictionary Thesaurus of the Romanian Language in electronic form) for the electronic version of these two parts, when no content changes have been operated. Furthermore, we will talk about the updated eDTLR, in which eDA has been upgraded to reflect the current language, as eDTLR+.

The next section presents some computerized dictionaries of other languages, beginning with Romance, which have similar structure/profile with DA-DLR. Section 3 gives a brief description of DLR and the advantages of an electronic version of DLR. In section 4, we concentrate on the solutions adopted for the computerized acquisition and use of the dictionary. The last section outlines some perspectives of work in Romanian lexicography, by means of computer resources and tools.

## 2. COMPUTERIZED DICTIONARIES IN THE WORLD

Although there are known 126 languages for which Internet reports monolingual or multilingual dictionaries in digital form[4], only very few languages have a thesaurus dictionary available in an electronic format.

For English, the language with the highest "presence" on the Internet, there exist many electronic dictionaries, either on-line or not. Well-known are the Oxford dictionaries for British and American English. The *Oxford Advanced Learner's Dictionary*, available on-line[5], is based on the British National Corpus[6], a 100 million word collection of samples of written and spoken language from a wide range of sources, representative for the spoken and written British English.

The Collins dictionaries[7] are created using the databases of existing dictionaries and the Collins Word Web, one of the largest databases of English language containing over 520 million words and their contexts. Each edition of the Collins dictionaries is based on a permanently update of the "new words" that appear in language use: new meanings, loanwords, idioms and catchphrases or changes in pronunciation. In December

[4] http://www.lexilogos.com/
[5] http://www.oup.com/elt/catalogue/teachersites/oald7/?cc=fr
[6] http://www.natcorp.ox.ac.uk/
[7] http://www.collins.co.uk/

2004 the dictionary publisher has launched Collins Word Exchange[8], a website where new words are added by anyone, and when the majority of users agree on the definition of the word and there is some evidence for its use on the Internet or in the Collins Word Web, then it is considered for the next edition of the dictionary.

For American English, at Merriam-Webster[9] the citation database contains more than 15.5 million examples of words used in context and more than 100 million words of electronic text. The dictionary can be used on-line, and for each entry it includes, besides all definitions, information about its pronunciation, part-of speech, usage and etymology.

French is well-known for its famous *Trésor de la Langue Française informatisé* (TLFi[10]), one of the largest on-line dictionaries of the Romance languages, with 100.000 words, 270.000 definitions, and 430.000 examples. The 16 volumes of the reference dictionary were acquired based on three main approaches: compiling a reliable computerized archive containing the text of the printed version of the TLF; retroconversion consisting of transforming the text into a structured text in which the various subjects of the articles (definitions, quotations, synonyms and antonyms, indicators for technical domains, and semantic, etymologic, grammatical, stylistic and historic indicators) are delimited; and developing the software to interrogate the structured text at 3 different consultation levels. The TLFi is integrated with Microsoft Word for Windows, with possibilities to correct or explain a word in your own document or to view it in TLFi, and with various Internet browsers.

The historical dictionary of the Italian language, *Tesoro della Lingua Italiana delle origini* (TLIO)[11], is compiled based on the *Opera del Vocabolario Italiano* (OVI). This database, with its 1849 vernacular texts (21.2 million words, 479,000 unique forms) is build, managed and 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*[12] is continuously updated by the Real Academia Española (RAE) and 21 Hispano-American academies based mainly on the synchronic (CREA - *Corpus de Referencia del Español Actual*) and diachronic (CORDE - *Corpus Diacrónico del Español*) 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 the *Língua Portuguesa On-Line*[13].

The Romanian language has an explanatory dictionary available on-line[14], obtained by contribution of web-volunteers which reproduce entries form the original version of the Romanian Academy Explanatory Dictionary (DEX, 1998), dictionaries of synonyms and antonyms. Currently it has almost 300.000 definitions.

## 3. THE DICTIONARY OF THE 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 the 'Dictionary of Romanian Language. New series' (DLR). The already printed 23 volumes of DLR, with their more than 10.000 pages, containing 15 letters (out of 28) of the Romanian alphabet and more than 93,000 entries, continues the old series of DA which in their 5 volumes, include more than 3,000 pages and cca. 45,000 entries. Statistics including both series of the dictionary are illustrated in Table 1. The new series will include the whole letter *D* and the whole letter *L*, as

[8] http://www.collins.co.uk/wordexchange/
[9] http://www.m-w.com/
[10] http://atilf.atilf.fr/
[11] http://tlio.ovi.cnr.it/TLIO/ricindex.html
[12] http://buscon.rae.es/diccionario/drae.htm
[13] http://www.priberam.pt/dlpo/dlpo.aspx
[14] http://dexonline.ro/

such opening the way towards the upgrade of DA, since the sequences *D – De* and *L – Lojniță* will be reworked. The dictionary was created in the traditional way till the nineties, when the editing and publication has started to be done with the help of computers.

**Table 1.** The Dictionary of the Academy and the Dictionary of Romanian Language – statistics

| Publication year | Tom | Part | Letter | No. of pages | No. of word entries |
|---|---|---|---|---|---|
| 1913 | I | **1** | ***A – B*** | 716 | cca 10000 |
| 1940 | I | **2** | ***C*** | 1064 | cca 16000 |
| 1949 | I | **3** | ***D – de*** | 90 | cca 1300 |
| 1934 | II | **1** | ***F – I*** | 936 | cca 14000 |
| 1937 | II | **2** | ***J*** | 66 | cca 990 |
| 1937-1949 | II | **2** | ***L – lojniță*** | 174 | cca 2600 |
| 1965-1968 | VI | | ***M*** | 1076 | 9653 |
| 1971 | VII | **1** | ***N*** | 584 | 5493 |
| 1969 | VII | **2** | ***O*** | 400 | 3622 |
| 1972 | VIII | **1** | ***P*** | 357 | 4006 |
| 1974 | VIII | **2** | ***P*** | 334 | 3783 |
| 1977 | VIII | **3** | ***P*** | 253 | 2727 |
| 1980 | VIII | **4** | ***P*** | 393 | 4537 |
| 1984 | VIII | **5** | ***P*** | 523 | 4680 |
| 1975 | IX | | ***R*** | 641 | 7255 |
| 1986 | X | **1** | ***S*** | 388 | 3540 |
| 1987 | X | **2** | ***S*** | 300 | 2212 |
| 1990 | X | **3** | ***S*** | 347 | 2692 |
| 1992 | X | **4** | ***S*** | 371 | 2757 |
| 1994 | X | **5** | ***S*** | 726 | 5725 |
| 1978 | XI | **1** | ***Ș*** | 271 | 4528 |
| 1982 | XI | **2** | ***T*** | 376 | 5027 |
| 1983 | XI | **3** | ***T*** | 387 | 4202 |
| 1994 | XII | **1** | ***Ț*** | 240 | 3856 |
| 2002 | XII | **2** | ***U*** | 468 | 2347 |
| 1997 | XIII | **1** | ***V*** | 326 | 1747 |
| 2002 | XIII | **2** | ***V*** | 426 | 2396 |
| 2005 | XIII | **3** | ***V, W, X, Y*** | 588 | 2365 |
| 2000 | XIV | | ***Z*** | 409 | 4088 |
| *TOTAL* | | | | **cca 13230** | **cca 138128** |

As a general dictionary, DLR records and explains most of the words which are attested in popular, literary and artistic speech. Special terminologies are present in DLR only if they are attested in at least two stylistic registers. Apart for completeness, DLR claims to gain a historical dimension by including all possible regionalisms, archaisms and popular technical terms. The argotic terms and personal creations of Romanian authors were included only if they are used in familiar or artistic speech; the personal creations of Romanian authors were included also if used in the general literary terms. The compounds have separate entries if they are perfectly fused together, if one of the components does not exist independently, or if they were borrowed in this form. All other cases of 'compounds' are listed under their first constituent. The derivatives are also treated as separate entries. Each homonym has a separate entry, with its own etymon. The words used in specialized languages (children games, riddles, magic spells), phrases, expressions, proverbs and sayings are listed in DLR if they are explained in their original sources, meaning that the occurrence is the one attested in the most representatives (literary, folklore, lexicographic and artistic) texts, linguistic atlases, etc.

A word entry in DLR has the general format illustrated in Figure 2. A special symbol is used to introduce a new sense of a word: the ◊ symbol introduces a closely related sense whereas the ♦ symbol introduces a more distant sense. Both symbols can appear in no matter what position in a word entry as depicted in Figure 1 and there is no rule to dictate their use.

**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 ...
- Etymological information

**Figure 1**. A word entry in DLR – general scheme.

DLR is an explanatory dictionary (DLR VI, 1965) as the words are defined and explained for all their senses, no matter their frequency, geographical area of provenance, or use. The linguistic facts were generally defined through *genus proximus* and *differentia specifica*, but also by delimiting the semantic values through synonyms or oppositions, by using pattern definitions for words that appear in semantic series (names of seasons, days, months, relatives, military degrees, dances, etc.), in the majority of abstract nouns, in diminutives, augmentatives, names of animals and plants. Prepositions, conjunctions, adverbs, some pronouns and numerals do not have definitions at all, as their place in language was described on the basis of the grammatical function (DLR VI, 1965). 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. The principle of the etymon is used when establishing the order of one word senses. Even though the first written attestations of a word have semantic values developed in Romania, 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.

### 3.2. eDTLR – a Necessity and a Goal

The acquisition of DA-DLR in electronic form, to which, as mentioned in section 1, we will refer as eDTLR, has the following advantages:

For computational lexicography:

– The existence of DA-DLR in electronic form represents the only solution to re-editing the dictionary, from A to Z, which implies the unification, correction and updating of the whole material of the dictionary – new and old series.

– The vast collection of texts/attestations used to exemplify words and senses of DLR (approximately 3.200.000 examples, representing about 88% from the whole text) could be used as a first source for updating the articles of the entries belonging to the old series of the dictionary (published between 1913 and 1949), as well as those of the first volumes of DLR (Florescu, 2006).

– The acquisition and maintenance software, fuelled by the eDTLR database, can be used as such and extended up to a sophisticated environment dedicated to lexicographers, which will contain a rich palette of updating possibilities, minimising the manual typing, and inclining the balance from low-level classical indexing operations on paper cards towards a manner of work in which lexicographers' expert knowledge is used in a creative and challenging way.

– The inconsistencies and differences between the volumes of DA and DLR, as well as differences between different DLR volumes, which came up as a normal consequence of the fact that different

generations of experts have worked in three different centres, will be easier detected and corrected, most of the time in an automatic way.

– It will be easier to define standard formats for certain types of entries, such as certain geological terms, months and days of the year, names 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 to help the lexicographer to fill-in the fields by always respecting the standards;

– It will offer substantial support to lexicographers for generation of various types of dictionaries: orthographic, pronunciation, frequency, etymological; valences, collocations, phraseological, proverbs, citations; onomasiological (thesaurus, dictionary of synonyms) or semasiological (dictionary of word families, retrograde, rhyme); neologisms, loan-word/foreign-word, jargon/slang, onomastic or depending on the group of users (general or language learner's dictionary). In those parts where the content of these dictionaries is not different from the original in eDTLR, the printing formats for these dictionaries can be generated automatically at no cost.

– It will open the only thinkable way for a continuous process of keeping updated the dictionary thesaurus of Romanian, in the rhythm in which the vivid language accepts new terms and senses, and forgets obsolete terms and senses, while also recording for the generations to come all terms which have been active once and are no longer.

For computational morphology:

– The close to exhaustive collection of Romanian terms will allow the completion of the computational morphology for Romanian up to the point where any term will be mastered in analysis and generation with respect to its flexing paradigm;

– The rich collection of examples could then supplement the corpora used for learning a Romanian language model, extensively used for POS-tagging and lemmatisation.

– For computational lexicology and computational semantics:

– The dictionary will make possible the alignment of word senses with those belonging to other dictionaries. Such an alignment was tested (Tănăsescu, 2004) for a part of the letter V and the Romanian WordNet (RoWN), the Romanian version of the English WordNet (Fellbaum, 1998), as it shaped at the level of the year 2004 (Tufis et al., 2004). Between these two resources, 73 common word entries have been found, for which DLR includes 2,300 senses and RoWN, at the time of the experiment, only 100 synsets.

– For the computational bilingual lexicography, the existence of eDTLR, will allow and enhance the automatic extraction of translation dictionaries.

– The vast collection of examples for the senses of words in DLR constitute a corpus of words annotated for senses which can be used to train a word sense disambiguation program to recognise senses of words in contexts. The ability to recognise word senses is fundamental in many applications of Natural Language Processing, including machine translation and information extraction.

– The same collection can be used for researches on semantic roles of verbs and nouns derived from verbs, FrameNet (Fillmore et al., 2003) related or not.

For publication, distribution and access:

– The dictionary can be published cheaper by electronic means.

– Sophisticated indexes can be drawn between word occurrences, including links to occurrences outside the dictionary itself, in other linguistic thesauri or in other languages.

– If agreed by the authors, the dictionary or only parts of it can be made available on Internet for the wider possible audience. This way any Romanian speaker and any researcher interested on the Romanian language, independent of the geographical zone s/he is located, will be offered access to it. In the authors view, this is the most significant achievement that is expected as a result of building the electronic version of the dictionary.

## 4. STEPS IN THE CONSTRUCTION OF eDTLR

### 4.1. Scanning, OCR and Initial Correction

Building an electronic form of DTLR is a complex process that requires efforts of linguists and programmers in a coherent collaboration. In this section we will sketch the main activities, their difficulties, the stages realised till now and the expected outcomes.

In the first phase, the volumes printed on paper are scanned. The scanning has already been done, for all volumes in DLR and just one fascicle of DA, using a scanner[15], which is known to protect the paper by its cold light beam[16]. The obtained images, each including two A4 pages of the dictionary, were split, deskewed, cleared of black margins and downscaled from 600 dpi to 300 dpi in preparation for OCR – Optical Character Recognition. We are developing an application on top of Gamera framework[17] for text recognition and segmentation. The difficulty in OCR is given by the multiple alphabets found in the dictionary: Latin, Greek, Cyrillic and Slavonic, which are extended with phonetic notations. Another factor which determines the quality of the OCR's output is the print quality, which varies considerably from the volumes printed during the sixties to those printed nowadays. To minimize errors, the Gamera symbol classifier must be trained on each volume using 5 to 10 pages.

Different strategies to correct the output of the OCR, the most costly and time consuming operation, have been considered. The initial plan was to leave the correction phase entirely on the responsibility of our colleagues, the lexicographers, the only ones capable to take adequate decisions in cases of unknown characters or ambiguity, due to artifacts or dirty zones on the original images. Tests have been performed on a class of master students in Computational Linguistics (in the years 2003-2004 and 2004-2005) in order to approximate the average time required to correct a page of the dictionary. The total amount of time needed by a person to correct the whole OCR-ized DTLR in this way was estimated to 663 days[18]. By being an extremely uninteresting and boring operation for linguists, it is very probable that at the end of this phase lots of errors will still remain. At least a second correction is needed in order to meet the requirements, which roughly translates into a 3-4 years project.

This is why we have thought for a different strategy, which is at this moment in a final stage of implementation. A Web-portal will make available an editing window dedicated, for correction sessions, to a large community of volunteers, students of our University, or even the public at large (see Figure 2). The apparent contradiction between making available to a large number of anonymous contributors the correction operations, while also aiming for a low rate of errors will be solved in three ways:

– Before actually distributing the material on the portal, an initial phase will try to pattern match, over the whole material, the abbreviations of the sources of citations in the examples with a known list. This way we hope to eliminate most of the errors on these fields, which, due to their lack of meaning for a novice user, are very difficult to correct, hence very prone to errors. Where possible to be done with a great deal of confidence on other zones, as for instance the headings of the information given at the end of one entry (grammatical, orthoepical and etimological information), the same procedure will be applied there too.

– The interface offers to the user the possibility to zoom the original scanned image, presented on the left of the screen, for bringing closer to the eyes portions which are difficult to read and understand for correction. The zoom is a real one in the sense that the image file is replaced during this operation with a higher resolution one, which, for reasons of quick access over the web, is used only for such purposes. When the ambiguity cannot be removed this way, the user has the possibility to mark zones of characters on which s/he is uncertain. The marked zones are recorded and, in different sessions, presented to the lexicographers for examination.

– Finally, by employing a large community of volunteers, the authors intend to develop a work-flow that allows for each correction to be performed at least twice, if not even three times, over the whole material of the dictionary. Initial estimates and the experience of the lexicographers show that each correction phase should be applied over a previously corrected version and not over the original[19]. This way, in the subsequent correction phases, the correctors will work on pages which contain less and less errors and it is supposed that this way the correction time will drop significantly. We evaluate now the option to keep a record of the correction operations (which are anyway recorded during each correction session for organising the undo operation) in the database. This way, if a subsequent user working on the same document, will operate a

[15] Type KONICA MINOLTA bizhub PRO 1050.

[16] A temperature higher than 30 Celsius degrees is never reached at the surface in contact with the paper.

[17] http://ldp.library.jhu.edu/projects/gamera/

[18] 8 continuous hours per day

[19] In another way of organizing the correction methodology, each corrector would have worked over the original OCR-ized document, the outputs of different sessions would have been compared and a voting strategy would have been implemented.

second correction on the same sequence of characters, the system will record a message addressed to the expert lexicographer that signals that zone is prone to an inexact match.

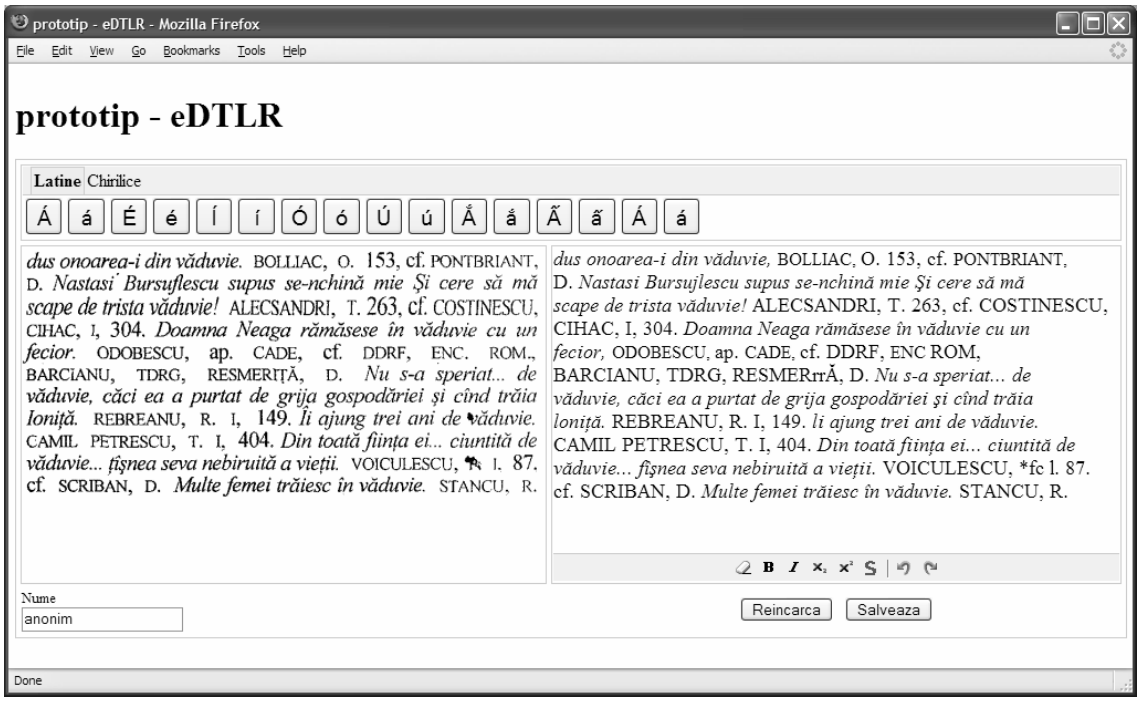


**Figure 2.** Screenshot of the on-line corrector interface
(left – the original scanned image; right – the editing window).

The Academy, which is uncomfortable to facilitate a large dissemination of an unfinished form of the dictionary, has, normally, imposed very strict distribution constrains at this stage. But how will these restrictions fit with using a large community of volunteers for correction operations? We have found an interesting solution to this problem by allowing a user to see and work only on an extract of a column[20] during each of her/his editing sessions using the portal. The window contains a piece of text which is no larger than the IPR reproduction limit and which is assigned randomly at the time of password-based log-on. After the correction session, the small piece is re-integrated in the whole by the portal, like in a game of puzzle. This strategy will make practically impossible for a user to re-assemble larger portions of the dictionary, as with 13,230 pages, the total amount of extracts goes to almost 160,000. A rough estimation for the probability to obtain one page from 12 extracts has the order of $10^{-55}$.

### 4.2. From the Corrected Input to a Database Format

In order to make possible any kind of search, browsing and re-printing operations over the electronic format of the dictionary, its corrected files have to be transformed into a database format. Inspired by the successful elaboration of a lexicographic grammar for DEX (Tufiş et al., 1999), it is possible that such a grammar could also be defined for DA-DLR, which would guide the parsing and help in the recognition of format errors. Such an enterprise is expected however to be much more difficult than in the case of DEX, due to great variations in the volumes' formats, as they have been elaborated over a long period.

The alternative agreed upon was to try to extract fields of the original scanned files through pattern-matching operations. An approach investigating this direction was finalised in a graduation paper (Hriţcu, 2004) that reported interesting results. At this moment we believe that this method can be applied

[20] 10 – 11 lines from one column

successfully for the extraction of the entries' fields with a sufficient accuracy. In order to do this, a tool has been designed and implemented, DLRex (Haja et al., 2005).

The functionalities of DLRex include:

– extracts the fields of the dictionary form (represented in RTF) and marks them onto XML;

– offers a browsing and interrogation interface over the resulted XML files;

– offers an updating interface for these fields.

The main functionality of the application is the conversion of the corrected OCR-ized DLR into XML format. The OCR-ized files are saved in the rtf format. When parsing the DLR, the formatting of the text and the presence of special symbols (such as ♦ or ◊) are very important. Parsing an entry is based on the succession of the formatting/styles used for each section of the entry. It has been tested on a broad sample of DLR entries and the results are promising, but at this moment we cannot guarantee that at the end of the extraction phase no manual intervention will be necessary.

By means of DLRex, the dictionary can be interrogated in advanced ways: it is possible to search word entries, definitions, senses, specific examples, etc.

### 4.3. eDTLR at the Basis of a Continuing Updating Effort of the Thesaurus Dictionary

eDTLR, the digitized form of DA+DRL, has the potential to become an important cultural and scientific work of a tremendous importance for the Romanian culture, offered to the public and it can open tremendously many opportunities for linguistic research. We believe however that there are at least two requirements to be fulfilled in order for this challenge to become a reality: first, the two big parts of the dictionary, eDA and eDLR, the electronic versions of DA and DLR, have to be unified in terms of content and format, and secondly, a methodology has to be agreed upon, at the level of the three linguistics institutes of the Academy, with respect to linguistic and lexicographic conventions, in order for the resulted eDTLR+ to be kept updated with the language evolution.

Neither of these enterprises is easy to accomplish. In the following we sketch some of the steps that we think are necessary for these requirements to become facts.

For the integration of eDA with eDLR (with the result – eDTLR+):

– All fields of eDA and eDLR, now XML coded, must be POS-tagged and lemmatized. Applied to title words, synonyms, definitions, and examples, the part-of-speech and lemma tags will allow building an index of the words of both series of the dictionary and, among others, a browsing capability can be developed based on this. The methodology to accomplish tagging and lemmatisation over eDTLR is in itself new, because, although POS-tagging and lemmatisation is successfully realised for Romanian (Tufiş, 2004; Ion, 2007), none of the previous runs have worked over a text of such a linguistic diversity and it is, hence, expected that a great number of items will not be recognised by the existing language models.

– All examples put together, as already shown in section 3.2, make up a very significant collection. Terms recognised in this collection as belonging to the entries in DA could than be sorted and compared against the entries already belonging to eDA. This way new terms, not worked on, as well as new senses of the terms that exist in eDA can be discovered and signalled to the lexicographers. They can use the set of examples of eDLR in which the term have been found to complement those found in other sources.

– The first step forward towards a consistent methodology for updating eDTLR will be accomplished when the texts in the list of sources used to collect examples will exist in electronic form, as well. This implies a significant effort to acquire them, either by the kind agreement of printing houses which do keep the corresponding electronic copies, or by digitising paper documents if no such resources can be found. All these texts will afterwards have to be POS-tagged and lemmatised, while also keeping all document identification information. Let us call this electronic national repository of resources that stay at the basis of eDTLR as the eCorpus.

– As remarked in section 3.2, existing examples can be used to train a program to do word sense disambiguation in context. Then, in order to recognise new senses, therefore not included in eDA, of a word which however belongs to eDA, one should first recuperate all occurrences of that word in eCorpus, operation which can be accomplished during one run of an occurrence finder. Then, the collection of these occurrences (each keeping a sentence-large context around the target word) will be classified corresponding to the senses learned from eCorpus. Those occurrences which are not recognised as corresponding to the

known senses are good candidates for new senses, and should be passed to the expert lexicographer for her/his evaluation.

– With a similar approach, new entries for words which do not exist in eDA21, can be added to the dictionary. The difference from the case of adding new senses is that this time no examples exist for training and therefore all occurrences found in eCorpus will have to be forwarded to the lexicographers for their consideration.

– In the following, we will mark as eDTLR++ the eDTLR+ which supports a continuous updating process. We see the following steps for the implementation of a process of continuous updating eDTLR++, by automatically processing a flow of Romanian texts, in the rhythm in which they are published (Cristea, 2005):

– The Romanian Parliament initiates and issues a law that recommends / imposes to the publishing houses and newspapers to archive all electronic variants of their publications in printed forms, in order to constitute a national archive / repository of electronic Romanian texts; the creators / owners of these materials will be protected by restricting the use of this archive only to researches dedicated to the Romanian language;

– An eDTLR++ Committee, formed by experts (mainly linguists, lexicographers, etymologists), establishes the selection criteria for the bibliography to be used as authorised sources for eDTLR++;

– The electronic resources are automatically sorted by their register: literary, stylistic, domain, author, publishing date, etc.;

– As new texts are added to the electronic national repository, a program continuously selects the resources recommended by the eDTLR++ Committee to be used in updating the eDTLR++;

– The eDTLR++ Committee establishes criteria for a word / sense to be considered as “new”, or “out-of-use”;

– The selected documents / resources are automatically annotated for part-of-speech and lemma, and the lemmas are sorted according to their frequency of occurrence;

– The criteria to establish the type of a word / sense (see above) are automatically applied and the output of the program is used by the eDTLR++ Committee in order to decide the word / sense type;

– The new words / senses are automatically detected, then passed to lexicographers to validate / reject these new entries;

– A program creates updated versions of eDTLR++;

– Using adequate interfaces, the lexicographers modify the automatically generated dictionary, wherever needed.

All implications of the activities described above are hard to be appreciated realistically at this moment. A short analysis, as described in section 3.2, shows countless possibilities to further use and exploit eDTLR++.

An approach which has some resemblance to the one we describe has been used in the development of the Collins COBUILD dictionaries (Sinclair, 1987). Currently, the dictionary is kept updated by the help and contribution of volunteers interested in such an activity. All decisions are supervised by a special Committee[22].

This way not only the acquisition, exploitation and updating the eDTLR++ will benefit, but also new (electronic) dictionaries can be generated and created in order to keep pace with the language evolution, and to respond to the needs of our society, taken as a separate entity or as a member of the global multilingual society.

## 5. CONCLUSIONS

This paper describes the technology and the first steps towards the realisation of the electronic form of the updated Dictionary Thesaurus of the Romanian Language (eDTLR+). After assembling the electronic versions of both the Dictionary of the Academy (eDA) and the Dictionary of the Romanian Language (eDLR) as eDTLR, in eDTLR+ eDA is updated to comply with the standards and the linguistic resources of

[21] Such misses may occur because of the difference between the list of text sources which have been used for the construction of DA as compared to those used for the construction of DLR.

[22] http://www.collins.co.uk/wordexchange/Default.aspx?pg=91

eDLR, this way making the reunion of the two series a homogeneous piece of work. We show the benefits of this realisation and the opportunities that it opens for interactive consultation and research. Furthermore, we propose a methodology for keeping updated this masterpiece of Romanian culture in line with the language evolution.

The authors are deeply convinced that besides the benefits showed in this paper, making available eDTLR, eDTLR+ and, ultimately, eDTLR++ on Internet will be an effort appreciated not only by Romanian speakers but also by many researchers outside the Romanian area, which are interested on Romanian. It is notorious that the average Romanian speaker hardly uses 30% of the Romanian lexicon. We envision that the richness of this dictionary thesaurus will make possible the revival of a significant segment of our language, at this time hardly known by the general public. This priceless treasure of the Romanian language, which concentrates at the base activities of the most reputed Romanian linguists over a century, will cease to be a pearl hidden in a shelf in exquisite academic libraries. It will finally become an asset given to the humanity.

## ACKNOWLEDGEMENTS

The work described in this paper, presently only partially accomplished, makes now the subject of an agreement between the Romanian Academy, represented by Marius Sala, vice-president, and the "Al.I.Cuza" University of Iasi, represented by its Rector, prof. dr. Dumitru Oprea. Through its interdisciplinary and pluri-institutional character, this agreement is the starting point for one of the most important projects the Romanian Academy has initiated in the last decades. Moreover, the pluri-institutional character of this endeavour emphasises the importance of the linguistic factor in the establishment of the values of cultural identity in the European context.

The authors are grateful to PIM-Iasi[23], the editing company, which, thanks to its director, Marius Petrariu, has offered their high performance scanning services at an incredible low cost.

The project has been until now supported by the following grants: INTAS 05-104-7633 RolTech[24] (Platform For Romanian Language Technology: Resources, Tools And Interfaces), work in-project coordinated by the Al.I. Cuza University of Iasi; the grant CNCSIS 1815, "The Dictionary of the Romanian Language in electronic format. Studies regarding its acquisition" (in Romanian), which run between 2003-2005 in the Institute for Romanian Philology "A. Philippide" of Iasi (in the project participated also researchers from the Faculty of Computer Science of the Al.I. Cuza University of Iasi); ROTEL – Intelligent systems for the Semantic Web, based on logics, ontologies and language technologies (29 CEEX I 03.10.2005), and InterOb - Interaction with computer systems and robotics (131 CEEX I 02.10.2006).

## REFERENCES

1. CRISTEA D., *Linguistic resources and natural language technologies. The case of Romanian language* (in Romanian). In Prelegeri Academice, Romanian Academy Iasi branch, **Vol. III**, nr. 3, Iasi, ISSN 1583-4514, 2005.
2. *DEX – Explanatory Dictionary of the Romanian Language*, Romanian Academy, Univers Enciclopedic Publishing House, Bucharest, 1998.
3. *Dictionary of Romanian Language. New Series.* (in Romanian: *Dicționarul limbii române. Serie nouă*) (DLR), Romanian Academy Publishing House, București, Romania.
4. (DLR VI, 1965) *Dictionary of Romanian Language. New Series.Tome VI*, 1965.
5. DUPONT C., *The Opera del Vocabolario Italiano Database: Full-Text Searching Early Italian Vernacular Sources on the Web*, Italica 78:4, pp. 526-39, 2001.
6. FELLBAUM C., *WordNet: An Electronic Lexical Database* (ed.). MIT Press, 1998.
7. FILLMORE C., JOHNSON C. R., AND PETRUCK M.R.L., Background to Framenet, *International Journal of Lexicography*, **Vol 16.3**: 235-250, 2003.
8. FLORESCU C., *Renewed Semantic perspectives: how we lexically advance towards DTLR* (in Romanian) In C. Forăscu, D. Tufiş, 9. D. Cristea (eds.) Proceedings of the Workshop Linguistic Resources and Tools for Processing Romanian Language. Iasi, Romania, November 2006. University Al.I. Cuza Publishing House. ISBN 978-973-703-208-9, 2006.
10. HAJA G., DĂNILĂ E., FORĂSCU C., ALDEA B.M., *The dictionary of Romanian Language in electronic format. Acquisition studies.* (in Romanian). Alfa Publishing House, Iasi, Romania. ISBN 973-8278-93-7, 2005.

[23] http://www.pimcopy.ro/
[24] http://consilr.info.uaic.ro/roltech/

11. HRIŢCU A., *Lexicographic frame for processing the Dictionary of Romanian Language (DLR)*. (in Romanian), Diploma Thesis, Faculty of Computer Science, Iasi, Romania, 2004.
12. ION R. *Methods of automatic semantic disambiguation. Applications to Romanian and English*. PhD thesis. Romanian Academy, Buchreast, Romania (forthcoming), 2007.
13. IORIO-FILI D., *Un nuovo software lessicografico: GATTO*. Bollettino dell'Opera del Vocabolario Italiano 2, pp. 259–70, 1997.
14. *Romanian Explanatory Dictionary* (in Romanian: *Dicţionarul explicativ al limbii române*), Romanian Academy, Institute of Linguistics Iorgu Iordan, Univers Enciclopedic Publishing House, Romania (first edition 1984, second edition 1996).
15. SINCLAIR J. M. (Ed.)., *Looking up: An account of the COBUILD project in lexical computing.* London: Collins COBUILD, 1987.
16. TĂNĂSESCU V. I., *Aligning electronic lexical resources, with application to DLR and Romanian Wordnet.* (in Romanian), Diploma Thesis, Faculty of Computer Science, Iasi, Romania, 2004.
17. *Trésor de la Langue Française (T.L.F.)*, CNRS, Gallimard, 1971-1994, 16 vol.
18. TUFIŞ D., CRISTEA D., STAMOU S., *.BalkaNet: Aims, Methods, Results and Perspectives. A General Overview*. In Romanian Journal on Information Science and Technology, Dan Tufiş (ed.) Special Issue on BalkaNet, Romanian Academy, **vol. 7**, no. 2-3, pp. 9-34, ISSN 1453-8245, 2004.
19. TUFIS, D., DRAGOMIRESCU L., *Tiered Tagging Revisited*. In Proceedings of the 4th LREC Conference, Lisabona, 2004.
20. TUFIŞ D., ROTARIU G., BARBU A.M., *Data Sampling, Lemma Selection and a Core Explanatory Dictionary of Romanian*. In Proceedings of the 5th International Workshop on Computational Lexicography COMPLEX, Pecs, Ungaria, pp. 219-228, 1999.