BULETINUL INSTITUTULUI POLITEHNIC DIN IAŞI
Publicat de
Universitatea Tehnică „Gheorghe Asachi" din Iaşi
Tomul LV (LIX), Fasc. 1, 2009
Secţia
MATEMATICĂ. MECANICĂ TEORETICĂ. FIZICĂ

STEPS IN BUILDING THE ELECTRONIC VERSION OF THE THESAURUS DICTIONARY OF THE ROMANIAN LANGUAGE

BY

DAN CRISTEA[*, **], MARIUS RASCHIP[*], ALEX MORUZ[*, **]

*** ALEXANDRU IOAN CUZA UNIVERSITYOF IASI, FACULTY OF COMPUTER SCIENCE**

**** ROMANIAN ACADEMY, INSTITUTE FOR THEORETICAL COMPUTER SCIENCE**

{dcristea,mraschip,mmoruz}@info.uaic.ro

Abstract. This paper describes a project aiming to develop the electronic version of a paper dictionary edited and printed by the Romanian Academy ever since 1913. We start by describing similar realisations and the characteristics of the paper Dictionary under study. Then we present the steps in the realisation of this project (scanning, optical character recognition, collaborative proofreading done by both novices and experts, parsing and linking the entries onto sources). We finish with some conclusions and estimations for further work.

Key words: computational lexicography, digital resources, electronic dictionaries

1. Introduction

Dictionaries in electronic form are a valuable source of information for mastering the diversity of language knowledge. In addition to their paper equivalents, they offer quick and diversified access, especially when the look-up criteria are of a complex nature. When the realisation of such resources is heavily limited in time due to contractual constrains, methods of collaborative working are tremendously precious. In this paper we report on such an endeavour, aiming to build the electronic form of the biggest lexicographic resource for Romanian language, also one of the richest in the world.

The dictionary we describe is the Thesaurus Dictionary of the Romanian Language, built and published on paper by the Romanian Academy between 1913 and 2009 (1).

The Web gives access nowadays to extremely many digital and on-line dictionaries[1]. Some of them are intended to be used both by humans and machines. We could say that the dictionaries are moving from paper to the Web. Here are some notorious examples: the *Oxford Advanced Learner's Dictionary*[2], *Collins Word Exchange*[3], *Merriam-Webster* dictionaries[4] for American English, the famous *Trésor de la Langue Française informatisé* (TLFi)[5] (one of the largest on-line dictionaries of the Romance languages: 100.000 words, 270.000 definitions, and 430.000 examples), *Tesoro della Lingua Italiana delle origini* (TLIO)[6], *Diccionario de la Lengua Española*[7], etc.

The change in format has brought also a change of vision of the way content could be provided. Examples of collaborative approaches in building dictionaries are: *Wiktionary*[8] (a multilingual Web-based project aiming to build a free content dictionary, available in over 150 languages), the *Kamusi* project[9] (a Swahili-English dictionary), the *Inuktitut Living Dictionary*[10] (a collaborative initiative of building an English-French-Inuktitut dictionary).

2. DA, DLR and eDTLR

The process of building and publishing the Thesaurus Dictionary of the Romanian Language took almost one century. The old series, known as the Dictionary of the Academy (DA) includes 5 volumes with 3,146 pages and 44.890 entries, and has been developed between 1913 and 1947. After an interruption, the work was restarted in the middle of the 7th decade of the last century with the new series, known as the Dictionary of Romanian Language (DLR). The last volume was finally published by the Editing House of the Romanian Academy at the beginning of 2009. In all, DA and DLR have 33

[1] For instance *A Web of On-line Dictionaries* at http://ling.kgw.tu-berlin.de/call/webofdic/diction4.html stores links to 800 dictionaries in 160 languages.
[2] http://www.oup.com/elt/catalogue/teachersites/oald7/?cc=fr
[3] http://www.collins.co.uk/wordexchange/
[4] http://www.m-w.com/
[5] http://atilf.atilf.fr/
[6] http://tlio.ovi.cnr.it/TLIO/ricindex.html
[7] http://buscon.rae.es/diccionario/drae.htm
[8] http://www.wiktionary.org/
[9] Initiated by the Council on African Studies at the Yale University Center for International and Area Studies http://research.yale.edu/cgi-bin/swahili/main.cgi
[10] http://www.livingdictionary.com/backgroundandhistory.jsp

volumes, more than 15,000 pages and about 175,000 entries. The dictionary was created in the traditional pencil-and-paper way, with citations collected from more than 2,500 volumes of the written Romanian literature.

eDTLR is the name of the digital form of DA+DLR, including its sources in digital form and the software to access them, but also the acronym of a three years project (2)[11]. It focuses on three main activities: transposing onto digital format the two parts of the dictionary, as well as the majority of its sources, correcting the digital format, and building browsing capabilities, including direct access from the dictionary citations onto the images of pages of the original sources.

3. Steps in Building eDTLR

Building an electronic form of DTLR is a complex process that requires efforts of linguists and computer scientists in a coherent collaboration. In this section we will sketch the main activities for achieving this goal.

3.1 Preliminary processing applied to the paper version

The project begun with scanning the printed dictionary volumes which have not been published by using electronic technology (as the very last few). The images thus obtained, 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. The difficulty in OCR is given by the multiple alphabets used in the dictionary: Latin, Greek, and Cyrillic. Another factor which influenced the quality of the OCR's output is the print quality, which varies considerably among the volumes.

In addition to OCR applied to the dictionary itself, we have started to process the huge number of bibliographical sources of the Dictionary. More than half of the sources have been scanned already and about 1/4th are OCR-ed, verified and stored on the eDTLR server.

3.2 The initial correction phases

In order to remove the errors introduced by the OCR on the dictionary files, we have decided for a combined novice-expert two steps proofreading process. A Web-portal[12] was open to a large community of volunteers[13].

---

[11] Running between 2007 and 2010 and financed by the National Centre for Program Management of the Romanian Ministry of Education, Research and Youth.
[12] https://consilr.info.uaic.ro/edtlr/

The Academy is reluctant to disseminate unfinished versions of the Dictionary as produced by intermediary steps. The solution was to allow a user to see and work only on an extract of a column (approximately 1/12 of a page) during each of the editing sessions using the eDTLR portal. The interface displays a piece of text which is no larger than the IPR reproduction limit. The segment is assigned randomly at opening the session and each time the user saves the work over a segment. After the correction session, the portal re-integrates the small pieces in the whole, like in a game of puzzle. This strategy makes practically impossible for a user to re-assemble larger portions of the dictionary. Then the program uses some heuristics to limit the fraud (applied only to novices): if a page is saved without any key touch during the editing phase, it is most probably fake and should be ignored; if the difference between the number of characters of the segment in input (as given by the OCR) and in output (as saved by the user) is too high, than the correction is ignored again.

3.3 Parsing the entries

After these two correction phases, the text is considered conform to the original. However, the Word format (as saved by the proofreading interface) is not compatible with the requirements of advanced search, because it records only typographic markings (CAPS, bold, italic, paragraphs) not also the significance of the entry fields. The structure of the entry should be recorded in a format compatible to lexicographic standards (for instance XML-TEI) and this is obtained by a specially designed parser (4). The parsing method developed, both efficient and robust, is done in three steps. In the first step a configuration of markers is used to identify the different types of fields in the entries (e.g., the etymological sections). The sense tree of each entry is then determined by another level of markers and, finally, the atomic definitions (fine-grained senses) are extracted by means of a third level of markers. As such, the parser first identifies the sense markers in a breadth-first manner, and only afterwards builds the sense tree. This strategy, although resembling other known approaches for dictionary parsing (for instance the one used by the French TLFI), brings as a novelty the hierarchical arrangement of the markers, declaratively separated from the parser code.

3.4 Correcting the structure

The entry structure obtained as a result of the parsing process is forwarded to a new correcting phase, which concentrates only on identifying structural mismatches. A specially designed interface highlights entry fields by using different colors and tags and displays two visions of the entry, a

[13] To our calls responded 500 professors and students from the Universities of Iaşi, Suceava, Bacău, Baia Mare, Galaţi, etc. as well as from the Republic of Moldova.

schematic tree-like one and a detailed, full text, one. The editing window allows the user to operate corrections relative to the limits of fields and to their hierarchical placement.

When this step will be accomplished by the lexicographers, all entries of the Dictionary will be heavily marked in XML.

3.5 Linking the dictionary entries onto sources

The last phase of the project aims to build links between the citations in eDTLR's entries and their corresponding sources (3). At a mouse click on a citation, a context, displaying a segment of the original page from where it has been extracted, will appear. This context could be larger or shorter, in conformity with the limits allowed by the intellectual property rights regulations.

4. Conclusions

The paper describes the eDTLR project, which aims at building one of the biggest digital dictionaries in the world. Indeed there are very few resources of this kind which could be compared with the dictionary described here, from the point of view of coverage (more than 15,000 pages on the original paper format, about 175,000 entries, approximately 1,300,000 examples.) and the main functionality (complex browsing capabilities and indexing of examples in images of the original editions of the written sources).

Benefits of such a large digital dictionary go towards easiness of access, large dissemination for speakers of Romanian, benefits for natural language processing and, not the least, a dramatic change in the manner in which lexicographers' work will be pursued from now on. Also, not negligible, the Dictionary in this form can be published cheaper, while also providing sophisticated indexes between word occurrences, including links to occurrences outside the dictionary itself, in other linguistic thesauri of Romanian or even in other languages.

Acknowledgements

This research is supported by the grant no. 91 _013/18.09.2007 of the Ministry of Education, Research and Youth, through the National Centre for Management of Programs. We thank to all our collaborators, as follows: the Faculty of Computer Science of the Alexandru Ioan Cuza University of Iaşi (coordinator), the Institute of Linguistics "Iorgu Iordan - Alexandru Rosetti" of

the Romanian Academy – Bucharest, the Institute of Romanian Philology ""Alexandru Philippide" of the Romanian Academy – Iaşi, the Institute of Literary History "Sextil Puşcariu" of the Romanian Academy – Cluj-Napoca, the Research Institute of Artificial Intelligence of the Romanian Academy – Bucharest, the Research Institute for Computer Science of the Romanian Academy – Iaşi, and the Faculty of Letters of the Alexandru Ioan Cuza University of Iaşi.

We are grateful also to all our collaborators, professors, students and the public at large, in Romania and the Republic of Moldova, who have contributed to the first proofreading phase in the elaboration of eDTLR.

R E F E R E N C E S

1. Dictionary of Romanian Language. New Series.Tome VI. 1965.
2. Cristea, D., Forascu, C., Raschip, M., Zock, M. (2008). How to Evaluate and Raise the Quality in a Collaborative Lexicographic Approach, Proceedings of LREC-2008, Marrakech.
3. Cristea, D., Răschip, M. (2008): Linking A Digital Dictionary Onto Its Sources, FASSBL Proceedings, Dubrovnik.
4. Curteanu, N., Moruz, A., Trandabăţ, D., Extracting Sense Trees from the Romanian Thesaurus by Sense Segmentation & Dependency Parsing, Proceedings of CogAlex Cognitive Aspects of the Lexicon: Enhancing the Structure, Indexes and Entry Points of Electronic Dictionaries, COLING 2008, pp. 55–63, ISBN 978-1-905593-56-9

Etape în realizarea Dicţionarului Tezaur al Limbii Române în format electronic

(rezumat)

Lucrarea descrie etape în construirea formatului electronic al marelui Dicţionar Tezaur al Limbii Române (dicţionar elaborat de Academia Română între 1913 şi 2009 şi tipărit în 33 de volume). Aceste etape au inclus scanarea volumelor Dicţionarului, transformarea imaginilor în şiruri de caractere, efortul colaborativ de corectare la prima mână cu voluntari, cel de corectare la mâna a doua cu experţi lexicografi, transpunerea automată a formatului de document într-unul adnotat XML, corectarea structurii, scanarea surselor Dicţionarului şi realizarea legăturilor între citate şi imaginile paginilor.