ROMTEXT – A FUNDAMENTAL INSTRUMENT FOR THE NEW EDITION OF THE *DICTIONARY OF THE ROMANIAN LANGUAGE*[*]

MONICA BUSUIOC[1], DAN CARAGEA[2]

Abstract. The article is a short presentation of the ROMTEXT project, a dated and annotated corpus of selected texts from the bibliography of the *Dictionary of the Romanian Language,* from the 16th − 21st centuries. The project aims at supporting the new digital edition of the thesaurus-dictionary, developed by the Lexicology and Lexicography Department of the "Iorgu Iordan − Al. Rosetti" Institute of Linguistics, Romanian Academy (2017–2019). ROMTEXT shall include over 500 literary and non-literary texts, obtained by optical recognition of the best editions, with assisted corrections. Subsequently, these texts shall be annotated from a morphological, syntactic and semantic point of view, by a team of lexicographers, with computer assistance. ROMTEXT shall have two concordance searching interfaces: one for lexicographers and one for the public. Results limitation and selection methods are also provided based on the text metadata. Due to its design and results, ROMTEXT shall be one of the most modern and versatile corpus linguistics available in Romanian.

Keywords: corpus linguistics, *Dictionary of the Romanian Language*; lemmatization; annotated corpus; reference corpus, Romanian language.

The most important work of Romanian lexicography, the *Dictionary of the Romanian Language* (*Dicţionarul limbii române – DLR*), also known as the *Academy Dictionary,* as it appeared under the aegis of the Romanian Academy, and the *Thesaurus-Dictionary*, due to its lexical richness, was completed in 2010. It was published by volumes, in a period of over one hundred years[3], whilst largely preserving the original methodology and drafting principles.

[*] This work was supported by a grant of Ministery of Research and Innovation, CNCS–UEFISCDI, project number PN-III-P4-ID-PCE-2016-0826, within PNCDI III.

[1] "Iorgu Iordan – Alexandru Rosetti" Institute of Linguistics, Romanian Academy. E-mail: monica.busuioc@yahoo.com.

[2] "Iorgu Iordan – Alexandru Rosetti" Institute of Linguistics, Romanian Academy. E-mail: dccaragea@yahoo.com.br.

[3] The draw up of the dictionary was initiated under the coordination of Sextil Puşcariu, in 1906, with the support of King Carol I. The first volumes and clusters were published between 1913 and 1949 (letters A, B, C, D–DE, F, I, J – LOJNIŢĂ). After an interruption of several years, from 1965 to 2010, under several coordinators, the remaining volumes and clusters were finally published. The publishing time line is the following:

A work like this, which has been drawn up under several different social-political regimes, observing various spelling regulations and lexical modifications appeared during this great time span, cannot be considered homogeneous or completed. No dictionary is complete or final.

In 2010, the "Iorgu Iordan – Al. Rosetti" Institute of Linguistics of the Romanian Academy, with the support of the National Bank of Romania, decided to print an anastatic edition, in 19 volumes, so as to give the public access to this great work (over 175,000 title-words and variants in approximately 18,000 pages)[4].

However, there is the stringent need for an updated, revised and completed second edition, in the shortest time possible, so as to keep up with the transformations suffered by the Romanian vocabulary. Within the context of the information society, this new edition shall benefit from technological progress and detach itself from the traditional media, thus becoming a digital lexicographic work. Therefore, the dictionary shall become a *lexicographic database,* with a writing interface and another one for editing and queries.

The main benefits of this transformation are the following: permanent updating of the dictionary, systematic treatment of the entries, controlled treatment of references, avoiding vicious definitions of words belonging to the same family or same domain (unwanted circularity), possibility of *online* publication and free access for the public to this fundamental work.

For these reasons, the coordinators from the "Iorgu Iordan – Al. Rosetti" Institute of Linguistics of the Romanian Academy have studied for several years the famous *Trésor de la langue française*[5] and in particular its web version: *Trésor de la langue française informatisé*, which can be consulted *online* since 2004.

In parallel with the revision activity and updating of the bibliography, as well as of the methodological principles, besides the effort of implementing the lexical database, the coordinators of the Lexicology and Lexicography Department have decided to develop a research corpus, following the example of the French correspondent, FRANTEXT. The project manager named it ROMTEXT.

Of course, the issue of *DLR* articles is based, throughout its history, on a corpus of texts. Today, the handling of a conventional corpus, the equivalent of a library, would

---

Old series (under the DA logo): 1913, A – B; 1934, F – I/Î; 1937–1948, J – LOJNIȚĂ; 1940, C; 1949, D – DE.

New series (under the DLR logo): 1965 – 1968, M; 1969, O; 1971, N; 1972, P – PĂZUI; 1974, PE – PÎNAR; 1975, R; 1977, PÎNĂ – POGRIBANIE; 1978, Ş; 1980, POGRIJENIE – PRESIMȚIRE; 1982, T – TOCĂLIȚĂ; 1983, TOCĂNA – TWIST; 1984, PRESIN – PUZZOLANĂ; 1986, S – SCLABUC; 1987, SCLADĂ – SEMINȚĂRIE; 1990, SEMN – SÎVEICĂ; 1992, SLAB – SPONGHIOS; 1994, SPONGIAR – SWING; 1994, Ț; 1997, V – VENI; 2000, Z; 2002, U; 2002, VENIAL – VIZURINĂ; 2005, VÎCLĂ – VUZUM; W, X, Y; 2006, D – DEÎNMULȚIT; 2006, DEJA – DEȚINERE; 2007, DEȚINUT – DISCOPOTIRIU; 2008, L – LHERZOLITĂ; 2008, LI – LUZULĂ; 2009, DISCORD – DYKE; 2009, E – ERZAȚ; 2010, ES – EZREDEŞ; 2010, J, K, Q.

[4] Academia Română, *Dicționarul limbii române,* 19 volumes, Bucureşti, Editura Academiei Române, 2010 (ISBN: 973-27-1977-0).

[5] ***, *Trésor de la langue française, dictionnaire de la langue du XIX^e et du XX^e siècle (1789–1960)*, 16 vol., Paris, CNRS, 1971–1994.

request an enormous amount of time, therefore the development of the corpus was carried out in stages, based on the texts included in the *Dictionary's* bibliography.

The first stage was represented by the *CNR*[6] project. This first attempt was based on texts from *DLR* bibliography. These texts were obtained by direct digitalisation. The project comprised a total of 250 texts. These are completed by another corpus developed within the Lexicology and Lexicography Department, which includes almost all bibliographic texts, obtained by optical recognition, without correction.

The drawbacks recorded in the previous period, considered a trial period, led to the initiation of a new, completely redesigned project, ROMTEXT, started in 2017[7].

Within the limits described in this project, ROMTEXT is a nuclear corpus, composed of over 500 texts (80% literary; 20% non-literary), with the following characteristics:

A. From the point of view of the nature of the text:

1. Includes completely analysed texts from the *DLR* bibliography, observing the requirements of a reference corpus. The internal version and link with the original page is preserved due to lexicographic requirements, for confirmation and quoting, according to *DLR* methodology.

2. Scanned texts are subject to image correction and cleaning procedures, as well as to optical character recognition and subsequent correction of the resulting text with ArqCorr[8], aspiring to identity with the original.

3. The obtained texts were processed by elimination of parts modified by editors, isolation of alotexts belonging to other authors (mottoes, forewords, quotes etc.), included in the work, as well as explanatory notes belonging to the author, integral part of the work.

4. Where applicable, the text structure was also marked, so as to avoid indexing the name and number of the chapters, sub-chapters etc., which would lead to an inflation of forms (letters, numbers, the noun "chapter" etc.), thus invaliding certain queries by an overabundance of answers.

5. Poems, articles, studies etc. reunited in volumes, collections, anthologies etc. have been considered autonomous texts. Therefore, the associated metadata can be used to carry out research at cycle, group and volume level, similar to a structure in chapters.

6. The scope is that this nuclear corpus include texts that illustrate the entire textual variety of the bibliography ("typological coverage"), in a derived proportion, to serve as guidance for further input, until exhaustion of the bibliography.

7. The texts included are written by authors from the 16th–21st centuries, in order to illustrate language progress ("chronological coverage"). Text distribution is based on the proportions derived from the bibliography.

8. We took into consideration manuals, scientific and technical texts, illustrating the various usual domains ("terminology coverage"), trying to obtain a distribution correlated with the bibliographical proportion and non-literary text proportions (20%).

---

[6] ***, *CNR – Corpus de referință al limbii române pentru constituirea de dicționare academice* (*Reference Corpus of the Romanian Language for the Development of Academic Dictionaries*), project financed by CNCSIS, Bucharest, 2007–2008.

[7] *ROMTEXT, digital corpus of Romanian texts, annotated and dated, 16th–21st centuries*, financed by UEFISCDI, under PN III-P4. Execution period: 2017–2019.

[8] Correction software for electronic texts obtained by optical recognition, developed by Arheus.ro (http://www.archeus.ro/lingvistica/main).

9. The texts included shall be accompanied by metadata (publication year, author, volume, domain, text type, working edition etc.), to allow lexicographers, as well as the public, to create micro-corpora for focal studies.

10. The written form of texts from editions published prior to the Romanian Academy Decision published in the *Official Gazette* of Romania, Part I, no. 51/1993, shall be updated according to the standards imposed by this decision and stipulated in the *Orthographic, Orthoepic and Morphological Dictionary of the Romanian Language (Dicţionarul ortografic, ortopepic şi morfologic al limbi române – DOOM,* 2nd edition, 2005). At the same time, we carried out the tacit correction of the mistakes overseen by the proofreaders of the previous editions, such as missing letters. In addition, all hyphens appearing at the end of the rows shall be eliminated. Words shall appear as a whole, according to the international practice applicable to linguistic corpora. The use of brackets or similar signs used by the editors to reconstruct words in old texts shall also be limited. In all these cases, the lexicographers shall have at hand the image of the original page, for collating.

B. From the point of view of lexical, grammatical and semantic annotation:

1. All texts included in the ROMTEXT corpus shall be annotated by the computer assisted method.
2. Inflected forms are managed by a lemmatizer, which allows control and presentation upon searching with a lemma. The lemmatizer is based on hypernyms, allowing the identification of historical, regional or even stylistic variants.
3. The same organisation method shall be extended to classes without inflections.
4. The grammar used for annotations is the traditional Romanian grammar, applied locally.
5. All words[9] shall be annotated morphologically by using the computer assisted annotation technique. For example, non-ambiguous monosemantic words shall be automatically annotated. Annotations refer to morphological class and certain morphological categories. Agglutinated forms (for example nouns articulated with a definite article), are considered sub-classes of the main morphological classes. Composed forms shall receive an additional tag, independently of the number of segments they are composed of.
6. All words shall be annotated syntactically by using the computer assisted annotation technique, based on dependencies. In the future, this shall allow viewing a sentence from the corpus under the form of dependencies.
7. Some semantically "full" words (nouns, adjectives, verbs, adverbs) shall be semantically annotated, allowing the finding of a semantic field by a single search (by one of its members).
8. Besides automated annotation and lemmatization programmes, computer assisted annotation implies the use of various available IT resources, as well as the lexicographer's human intervention, to control and correct ambiguities in the corpus.
9. Special attention shall be paid to old texts. Their annotation shall be made by linking the texts with indices and terms glossaries of the editions in question.

[9] A *word* represents a combination of alphabetic characters forming an autonomous graphic segment, separated from other segments by spaces or punctuation signs.

10. Lists of "special" words (forms not recognised by the spell checker) shall also be used for text annotation. They shall be recorded upon each correction and subsequently processed by lexicographers.

11. The objective of these multiple annotations is to facilitate the search and in particular avoid the return of a large number of forms with identical characteristics (redundant results). Therefore, they aim at helping the researcher from the point of view of linguistic coverage and representativeness.

What can we further mention about corpus searches? Of course, they must satisfy the lexicographer's needs to the highest degree. On the other hand, ROMTEXT allows a wide variety of research for a comprehensive and diverse public. The answers shall be optimised depending on the above mentioned coverage principles.

C. From the point of view of corpus interrogation, the following searches can be made:

1. forms (alphabetical or figures or other symbols);
2. hypernyms, with extension to classes without inflections;
3. proper names (also within semantic classifications, such as *nume de localități*)
4. foreign names (Latin, French, German etc.)
5. combinations of forms or lemmas with *x* segments ("noapte argintie"; "păr de abanos");
6. combinations of forms or lemmas with morphological classes ("râu" + adjective);
7. phrases, expressions, collocations ("de-a lungul", "la Paştele cailor")
8. numbers (Arabic or Roman figures)
9. dates ("10 mai", "iunie 1916")
10. a certain part of speech; semantic fields.

The search includes various selection criteria: year, author, domain, text type, number of lemmas, forms etc., as defined in the text metadata sheet.

Conclusions

ROMTEXT is a corpus project for the completion of the new edition of the *Dictionary of the Romanian Language,* which, from many points of view, exceeds the needs of the lexicographers working on the dictionary, thus aiming at the status of *reference corpus* of the Romanian language.

ROMTEXT was described in the project as a *nuclear corpus*. What we mean is that it covers the central area of Romanian vocabulary ("fundamental vocabulary", "representative vocabulary", "essential vocabulary" etc.), thus becoming a model for further development. A series of mini-corpora can be built around it, comprising specialised texts, processed under the project (by computerised procedures and instruments) based on sample-texts for all language varieties described by functional stylistics, terminologies etc.

ROMTEXT is *generational corpus*, meaning that it can indicate growth paths by comparing resulted vocabulary with that of the *DLR* macro-structure. Assuming, nonetheless, that both durability and reliability requirements are controlled and met. And by that we mean coverage and representativeness.

In conclusion, we hope that this project will be appreciated both by colleagues and the general public, as a pre-requisite for its future development.

BIBLIOGRAPHY

***, *FRANTEXT*, 5,118 reference texts, 297,586.781 terms, 10th–21st centuries (in august 2017), http://www.frantext.fr/

***, *Trésor de la langue française informatisée*, ATILF-CNRS & Université de Lorraine, http://atilf.atilf.fr/

***, *Corpus de Referencia del Español Actual (CREA),* http://web.frl.es/CREA/view/inicioExterno.view

***, *PAISÀ (Piattaforma per l'Apprendimento dell'Italiano Su corpora Annotati),* http://www.corpusitaliano.it/

***, *CRPC − Corpus de Referência do Português Contemporâneo,* http://www.clul.ulisboa.pt/pt/23-investigacao/714-crpc-corpus-de-referencia-do-portugues-contemporaneo

Dănilă, E., 2010, "Despre necesitatea realizării unui corpus lexicografic românesc esențial", *Philologica Jassyensia,* VI, 2 (12), 41–49; http://www.philologica-jassyensia.ro/upload/VI_2_Danila.pdf

Irimia, E., 2015, "Accelerarea dezvoltării unui corpus digital adnotat cu relații de dependență pentru limba română utilizând resurse şi instrumente construite pentru alte limbi", *Revista română de informatică şi automatică*, 25, 3; https://rria.ici.ro/wp-content/uploads/2015/09/03-art-1-IRIMIA_ELENA1-OK-2.pdf

Irimia, E., V. Barbu Mititelu, 2015, "RACAI-RoTb: nucleu de corpus de limbă română adnotat sintactic cu relații de dependență", *Revista Română de Interacțiune Om-Calculator,* 8, 2, 101–120; http://rochi.utcluj.ro/rrioc/articole/RRIOC-8-2-Irimia.pdf

Irimia, E., D. Tufiş, f. a., "Tehnici de validare şi corecție focalizată a adnotării morfo-sintactice în corpusuri de mari dimensiuni", Institutul de Cercetări pentru Inteligență Artificială, Academia Română, Bucureşti; http://www.racai.ro/media/Tufis-Irimia-CONSILR2006.pdf

Perez, C.-A., f. a., "Un corpus de texte pentru limba română adnotat sintactic sub formă de arbori", www.diacronia.ro/indexing/details/V1665/pdf