LITTLE STROKES FELL GREAT OAKS. CREATING COROLA, THE REFERENCE CORPUS OF CONTEMPORARY ROMANIAN

DAN TUFIŞ[1], VERGINICA BARBU MITITELU [1], ELENA IRIMIA[1], VASILE PĂIŞ[1], RADU ION[1], NILS DIEWALD[2], MARIA MITROFAN[1], MIHAELA ONOFREI[3]

Abstract. The paper presents the quite long-standing tradition of Romanian corpus acquisition and processing, which reaches its peak with the reference **co**rpus of contemporary **Ro**manian **la**nguage (CoRoLa). The paper describes decisions behind the kinds of texts collected, as well as processing and annotation steps, highlighting the structure and importance of metadata to the corpus. The reader is also introduced to the three ways in which (s)he can plunge into the rich linguistic data of the corpus, waiting to be discovered. Besides querying the corpus, word embeddings extracted from it are useful to various natural language processing applications and for linguists, when user-friendly interfaces offer them the possibility to exploit the data.

Keywords: Romanian corpus, acquisition, metadata, annotation, query.

1. INTRODUCTION

Collecting language data is not a recent enterprise, but with the advent of information technology, this has become a more systematic activity, subject to more and more precise rules of compiling and documenting. There are three major types of machine-readable data collections: archives, electronic text libraries (ELT) and corpora (Atkins *et al*. 1992). An archive is a repository of readable electronic texts not linked in any coordinated way. An ELT is a collection of electronic texts in a standardized format with certain conventions relating to content, metadata, etc., but without rigorous selection constraints. A corpus is a particular type of an ETL, built according to explicit design criteria for a specific purpose: the texts in the corpus are interesting and useful for the theoretical or computational study of language (not only great works of literature, but also works of other writers, or transcriptions of ordinary conversations). In a landmark report of EAGLES (Expert Advisory Group for Language Engineering Standards), Sinclair (1996) introduced

[1] Romanian Academy Research Institute for Artificial Intelligence (ICIA), {tufis, vergi, elena, vasile, radu, maria}@racai.com.

[2] Leibniz Institute for the German Language, Mannheim, diewald@ids-mannheim.de

[3] Iaşi Branch of the Romanian Academy, Institute for Computer Science, mihaela.plamada. onofrei@gmail.com

RRL, **LXIV**, *3*, p. 227–240, Bucureşti, 2019

an authoritative corpus typology, providing definitions and characterizing different types of corpora (spoken, reference, monitor, parallel, comparable corpora). According to this study, a reference corpus provides comprehensive information about a language, aiming to be large enough to represent all the relevant varieties of the language, the characteristic vocabulary, "as a basis for reliable grammars, dictionaries, thesauri and other language reference materials". A reference corpus may also be hierarchically structured and have subcorpora (Sinclair 1996).

In the present-day understanding, a corpus is a (very) large collection of language data, pre-processed at multiple levels, represented in standardized and interoperable formats and documented following precise specifications. Building and maintaining a corpus is an institutional, long-time job (i.e. it has to be maintained over an indefinite period of time), it is scientifically exciting, calling for multidisciplinarity, and has a major cultural dimension.

2. PREVIOUS STAGES

At ICIA, corpus construction was mainly motivated by language engineering goals, such as developing and evaluating language processing tools (taggers, parsers, aligners, information extraction, etc.). In the TELRI European project (1995-1997), among other relevant actions, a multilingual parallel corpus, including Romanian, was built based on translations in 21 languages of Plato's *Republic*. All the translations were encoded in SGML according to the TEI standard recommendations[4]. This was the first dataset which allowed research on alignment and translation technologies for Eastern-European languages. A similar multilingual parallel corpus was developed during the Multext-East European project (1995–1997), this time based on George Orwell's *1984*. Language resources were developed for Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene, as well as for English – as the pivot language of this project. Existing tools and standards were adapted to these languages. The encoding standard was conformant to CES (Corpus Encoding Standard) and the annotation language was also SGML, replaced by XML in 2000, when XCES guidelines were adopted (Erjavec 2012). The number of translation languages contributed by volunteers, reached 16 (in 2011), and these new sub-corpora followed the same annotation principles as the initial Multext-East corpus. This was and continues to be one of the most influential parallel corpora, because of its quality: it is manually tagged (using the same tagset) and lemmatized following the XCES encoding specifications. Additionally, the initial 7–language version is word-aligned (towards English), which was highly instrumental in the development of the project BalkaNet (2001-2003) that resulted in the first core wordnets for Bulgarian, Czech, Greek, Romanian, Serbian, and Turkish (Tufiş *et al.* 2004). Most of the development and validation of the BalkaNet wordnets (fully aligned at the synset-level with the Princeton WordNet) and associated tools were based on the aligned parallel corpus *1984*.

The experience gained with multilingual corpus construction and corpora-based applications, as well as the introduction of advanced machine learning technologies, outlined the necessity to enlarge the typology of Romanian texts and to consider much

[4] https://tei-c.org/

more textual data. In 2004, at ICIA, these motivated the launching of a new two-year project, RoCo-News, aiming at producing a larger corpus of Romanian. We began with news data, donated by a weekly magazine published in Timişoara (*Agenda*). The data was organized on topics which allowed us to generate useful metadata and structure the entire collection as a news corpus. The RoCo-News corpus (Tufiş and Irimia 2006) is significantly larger than the previous literature-based corpora, containing about 7 million tokens. The corpus was carefully processed (tokenized, tagged and lemmatized) and hand validated for better serving the training of language processing tools. By the same time, a great collection of parallel texts has been released by the European Commission and it became rapidly a must-have-resource for everybody working in multilingual technologies (cross-language information retrieval, text alignment, machine translation, etc.). The JRC-Acquis parallel corpus (Steinberger et al. 2006) contained in its first release about 8000 documents in 20 languages (including Romanian), with an average of 9 million words per language. The parallel documents were paragraph and sentence aligned, 190 language-pairs being available in the distributed data. The ICIA team cleaned, processed and validated the Romanian data (sentence splitting, tokenization, tagging, lemmatization and alignment to English texts). Later, by taking advantage of the previously developed alignment technologies (Tufiş *et al*. 2005, 2006) we refined the alignment of Romanian-English sub-corpus of JRC-Acquis at word level, turning it into a unique gold corpus for this pair of languages. With the corpora created during the previous years, the idea of compiling a balanced corpus for the Romanian language naturally emerged, and by 2012, within the project METANET, we released the ROMBAC corpus (Ion *et al*. 2012). The corpus contains about 36,000,000 words evenly distributed into five genres: journalistic (news and editorials), pharmaceutical and medical short texts, legalese, biographies of Romanian writers, critical reviews of their works, fiction (both original and translated novels and poetry). The texts are tokenized, morpho-syntactically tagged, lemmatized, shallow-parsed (chunked), XCES-compliant encoded and accompanied by metadata.

In 2012 a METANET analysis of the informatization status of the European languages revealed the fragmentary support that Romanian had for speech and text resources (among other categories analyzed). In the same year, the Romanian Academy approved a two-year project at ICIA for beginning a national corpus for Romanian. The first version was a compilation of the previously constructed corpora, harmonizing the annotations and metadata, correcting detected errors and acquiring more textual data. In 2014, the project was joined by IIT-Iaşi and it became a priority project of the Romanian Academy. Named CoRoLa, it aimed at producing a fully IPR-cleared reference corpus of written and spoken Romanian (Barbu Mititelu *et al*. 2014). Its first phase ended in December 2017 with a successful public opening and was extended for two more years (Tufiş *et al*. 2016). In 2016, the CoRoLa consortium with the University of Bucharest and the Leibniz Institute for the German Language (IDS) in Mannheim started a partnership programme funded by the Alexander von Humboldt Foundation. This partnership, besides knowledge transfer from IDS, creators of the largest linguistically motivated collection corpus of contemporary German (DeReKo, comprising more than 40 billion words), supported the acquisition of two powerful hardware servers, as well as data and services transfer to the KorAP corpus management platform developed by IDS (Bański *et al.* 2012). KorAP is one of the most powerful corpora management environments, able to deal with

tens of billions of words in a very fast and efficient way, with a plethora of querying facilities (Cristea et al., in this volume).

3. COLLECTED TEXTS

CoRoLa reflects both written and spoken Romanian. There has been a concern for the diversity of texts to be included in the corpus throughout the project, covering several aspects, like:

- source type: texts collected from publishing houses, radio stations, newspapers and magazines, journals, websites, blogs;
- source[5] names;
- document type: books, book chapters, newspaper/magazine articles, scientific articles, Wikipedia articles, news, interviews, blog posts, letters, reports, etc.;
- style: imaginative, journalistic, scientific, legal, administrative, memoirs;
- domains: arts and culture, science, society, nature;
- subdomains: we were able to cover around 70 subdomains;
- author;
- year of publication; in this case, the diversity is hindered by the extremely limited availability of electronic texts older than about 20 years;
- place of publication. We targeted sources from geographic regions where Romanian is spoken, including the language of the Romanian diaspora. However, there are no texts from the Republic of Moldova, where Romanian is also spoken. Romanian from the diaspora has a smaller representation in the corpus.

Oral texts are either read or spoken texts, recorded in better or poorer conditions (such as radio broadcast, texts read in professional recording studios, texts read in non-professional medium, etc.), with one or several speakers per document. We have not included instantaneous speech. The targeted diversity of the texts was limited by the difficulties of getting access, in many cases due to copyright law. In an I(ntellectual) P(roperty)R(ight) strictly regulated society, gathering large quantities of text and speech data representative for a language is not an easy task and implies requiring written consent from the IPR holders for storing, processing offered texts and making them queryable.

4. METADATA

Information describing the content of the actual data from various perspectives is essential in organizing and exploring the corpus. This was crucial for CoRoLa as well, from its very beginning. Metadata can comprise different types of information, from the most

[5] A comprehensive list of text providers (i.e. sources) is available on the corpus website, corola.racai.ro.

general, like the institution or group developing the resource, to the most specific, such as the author of a certain document in the resource. We focused on document-level metadata generation, the document being the main unit of organizing textual data. For a text to be introduced as a document in the corpus, it must have a specific title and authorship (Bibiri et al. 2015): e.g., an article in a journal, a poem in a poetry volume or a chapter in an edited book are all registered as different documents in our collection.

Metadata is essential in retrieving data from the corpus, in organizing it in sub-corpora, in obtaining statistics on different types of criteria. The metadata model we used was inspired by the CMDI (Component MetaData Infrastructure; Broeder et al. 2012) approach, but we designed a simplified version, containing the following attributes: DocumentTitle and ArticleTitle (which, in some cases can be the same, e.g. for a novel, and in others are different, e.g. all the articles in a magazine have the same DocumentTitle, i.e. the title of the magazine, and each of them has its own ArticleTitle), PublicationDate, the Source type and the SourceName, the AuthorName and the TranslatorName (when applicable), the Medium (Written or Oral), DocumentTextStyle, DocumentTextDomain and DocumentTextSubdomain, CollectionDate (document collection year), SubjectLanguage (which is Romanian, but could have other values when we decide to introduce parallel documents in the corpus) and ISSN-ISBN.

The metadata creation was done automatically for texts crawled from the web: specific classifications of data existent on specific websites were exploited to automatically extract values for metadata attributes such as ArticleTitle, AuthorName, DocumentTextDomain, while more general attributes like DocumentTextStyle, Source, SourceName, CollectionDate, etc. were provided by the person responsible for collecting the data (Gîfu et al., in this volume).

In order to enable the metadata created in CoRoLa to be accessible and exploitable in KorAP, we have mapped all categories to I5 metadata categories (Kupietz and Lüngen 2014), the only supported metadata scheme at the time. Support for arbitrary metadata fields was later introduced to KorAP in version 0.58.4 of the backend component Krill (Diewald and Margaretha 2017). Based on these indexed metadata fields, virtual subcorpora for KorAP can be created (see Cristea et al., in this volume) by combining field constraints and boolean operations to restrict research to all documents in the corpus, that, for example, were published before 1980 or written by a certain author. Different relational operations can be used to express field constraints in KorAP, depending on the chosen metadata field type (string, integer, tokenized text, date, or keywords). Therefore, the metadata field type, as well as the semantic similarity were taken into account for mapping the metadata categories.

5. TEXT PROCESSING

The diversity of texts targeted in the project brought along a diversity of challenges posed by content and format. The former required removal of some parts (all information on the title page, figures and tables and their captions, page numbers, headers and footers, etc.), replacement of other elements (non-standard characters in the UTF-8 encoding),

markup of others, such as foot- and endnotes, gaps, etc. (Bibiri *et al*. 2015). Irrespective of the file format provided to us, all texts end up in the txt format, which does not recognize columns, pages, etc. The conversion was mainly automatic (see Gîfu *et al*., in this volume), but also manual, in a small number of cases.

The speech data gathered in this project are larger than what is made available to the users (we collected about 300 hours of recordings). Given the limitations imposed by the speech involving applications (Boroș and Dumitrescu 2015) and by the corpus query possibilities, only recordings for which transcriptions were made are reported as part of the corpus. The audio files are pre-processed to eliminate noise, when present.

All written texts were normalized prior to being annotated. The following actions were taken: elimination of all texts which are not UTF-8 encoded, automatic insertion of Romanian diacritics in texts that lack them (where the percent of invalid words due to missing diacritics is more than 98%), automatic elimination of the hyphen character when used for splitting a word into syllables at the end of a line, automatic splitting of glued words (only two glued words are checked), automatic sentence filtering that removes sentences containing word material such as units of measure, table drawing characters, punctuation, foreign (non-Romanian) characters, etc. The texts collected display two orthographic norms, and this requires further normalization.

6. ANNOTATION LEVELS

All written texts (transcriptions included) are annotated with the TTL tool (Ion 2007), which automatically identifies sentences and tokens (words, punctuation, other symbols), morpho-syntactically annotates words and lemmatizes them.

All text files are segmented at the sentence level. The sentences from transcriptions are aligned with the corresponding audio stream, which means that the audio files are also split into shorter files, containing the uttered sentence. Recordings were also phonetically transcribed; words being divided into syllables and aligned at the phoneme level with the audio files.

The annotation of CoRoLa is entirely automatic, except for a medical corpus, MoNERo, extracted from the BioRo corpus (Mitrofan and Tufiș 2018). MoNERo contains 154,825 tokens in 4,987 sentences from three medical subdomains (cardiology, endocrinology and diabetes). It was manually validated at the morphological level and manually annotated with medical named entities belonging to four semantic groups from UMLS (anatomy, procedure, chemicals and drugs, disorders) (Mitrofan et al. 2018).

7. STATISTICS

The distribution of texts according to several criteria, valid at the moment of writing this paper, is presented in Table 1. The numbers represent tokens in the texts.

Table 1

Statistics about CoRoLa

Source type		***Domain***	
Blog	68164515	**Art and Culture**	86191512
Journal	54091238	**Science**	122818526
Publishing house	203212135	**Society**	614653080
Website	614294877	**Nature**	2045244
Other	1441404	**Other**	115495807
TOTAL	941204169	**TOTAL**	941204169

Document type		***Style***	
inCollection	24927365	**Blogpost**	66593706
Book	61895389	**Journalistic**	62442634
Newspaper article	35250538	**Imaginative**	57683325
Booklet	444965	**Science**	168465155
Manual	2359736	**Memoirs**	20195522
inProceedings	818201	**Law**	553924238
inBook	9910869	**Administrative**	10328525
Techreport	134474	**other**	1571064
Blogposts	1568425	**TOTAL**	941204169
wikiArticle	43264253		
Unpublished	467		
Master/PhD Theses	61209		
other	760568278		
TOTAL	941204169		

The oral part of CoRoLa currently contains 49.989 aligned audio files, totalizing over 103 hours. The transcribed recordings contain 821,294 tokens, out of which 45,300 are distinct tokens, as they were identified by the TTL tool. From these tokens, 20,833 have a single occurrence in the corpus (hapax legomena).

8. INTERROGATION TOOLS

There are three tools (two for the written, and one for the oral component) that allow the user to query the corpus. A link to each of them is available on the corpus website (corola.racai.ro). These tools are KorAP, NLP-CQP, and OCQP.

The entire written CoRoLa is indexed by KorAP (Bański et al. 2012), which also allows for its querying. For different ways in which this can be done, see Cristea et al. (in this volume). Users are strongly encouraged to use this tool for their searches. On the one hand, the whole corpus is accessible here, on the other hand, this is a powerful platform.

A part of the written CoRoLa is also accessible with NLP-CQP. This is an application that tries to automatically convert queries formulated in constrained Romanian phrases (describing token succession and annotation) into a format specific to C(orpus)Q(uery)P(rocessor) (Hardie 2012). It does this by automatically identifying search predicates and their arguments in the request. The following token properties may be mentioned in the Romanian search phrase: part-of-speech, lemma, word form and syntactic group in which the token may be embedded (i.e. noun phrase, adjectival/adverbial phrase, verb phrase and prepositional phrase). For instance, the Romanian search phrase (see Fig 1): *100 de fraze în care lema "mașină"* ***este urmată imediat*** *de un grup prepozițional* ('100 sentences in which the lemma "car" **is immediately followed** by a PP') will be automatically translated into the following CQP query:

set Context s; [lemma = "maşină"] <pp> cut 100;

The underlined phrases represent the natural language terms, the bolded phrase, the search predicate. These are automatically detected by NL2CQP using predefined patterns. The request above is about obtaining 100 sentences ("100 de fraze") in which the singular form of the lemma "maşină" ('car') is immediately followed ("este urmată imediat") by a prepositional phrase ("un grup prepozițional").

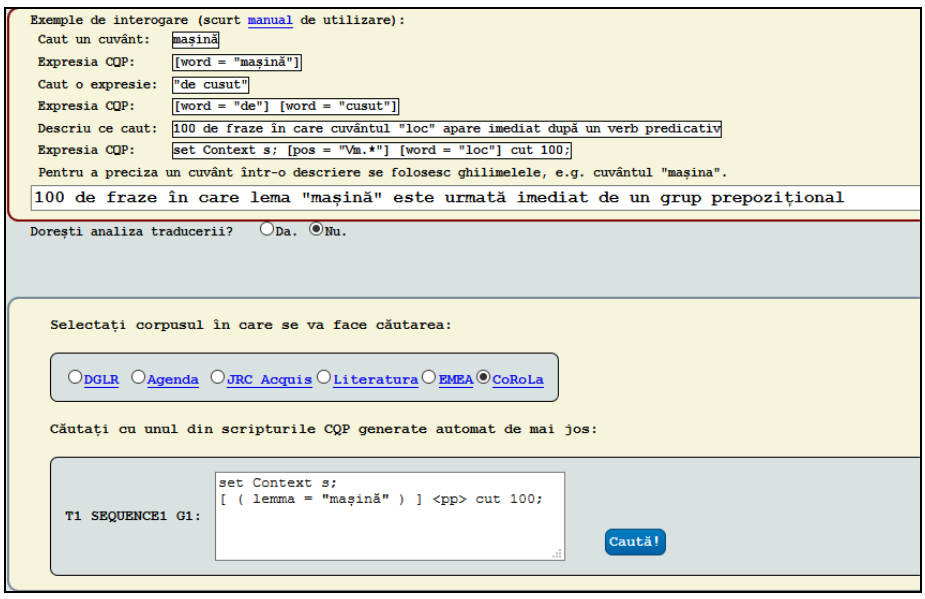

Fig. 1. Query formulation part of the NLP-CQP interface.

The Oral Corpus Query Interface (OCQP) is a custom developed online interface for querying the oral component of CoRoLa (Fig. 2) and visualizing the results (Fig. 3). It allows searching for single words or lemmas, as well as restricting the results to a subset having a specified part-of-speech (formulated either by indicating the C(ategory)TAG or the M(orpho)S(yntactic)D(escription). Display options can be chosen in order to display the words, lemmas, part-of-speech tags in CTAG or MSD notation. The context window can be either 5 tokens (Fig. 2) or the full sentence.

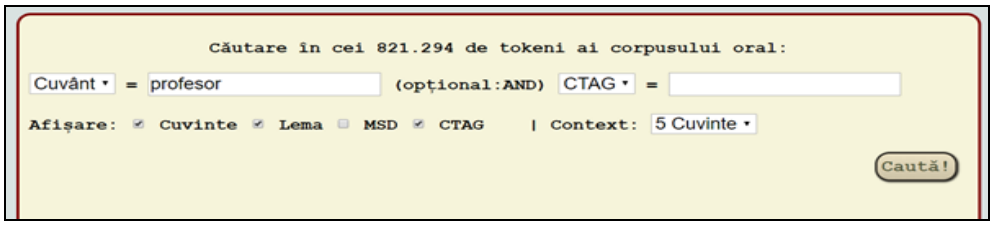

Fig. 2. Query formulation part of the OCQP interface.

Apart from displaying the context, the interface also offers the possibility to listen to either the query word or to the entire sentence. In the case of more results, the interface allows pagination and retrieval of 20 results for each page.

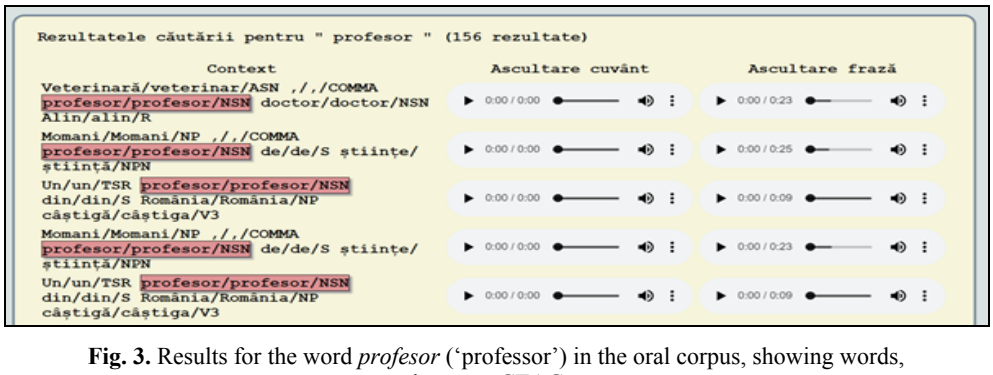

Fig. 3. Results for the word *profesor* ('professor') in the oral corpus, showing words, lemmas, CTAGs.

9. WORD EMBEDDINGS

Given the size and representativeness of the CoRoLa corpus, it became a natural choice for computing distributional representation of words, also known as word embeddings (Mikolov et al 2013, for Romanian see Păiş and Tufiş 2018a). This representation assigns each word a vector of real numbers, having the property that similar words will have corresponding vectors with small cosine distance. It should be noted that similarity is based on corpus co-occurrence here. For visualizing and interacting with the word embeddings representations, several textual and graphical interfaces were constructed. The most fundamental one is based on a simple word query. Given a word, it will show

10 most similar other words, including the computed cosine distance and the associated vectors (if requested by the user). An example for this interface is presented in Fig. 4: for the word *cald* (‘hot’), one can see semantically similar words found in CoRoLa.

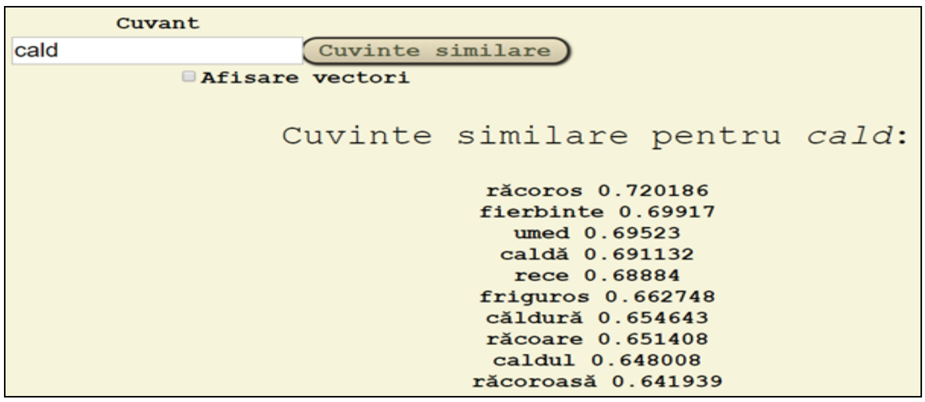

Fig. 4. Basic interface to find similar words using word embeddings.

The same similar words can be displayed graphically in a 2D representation. The projection from the embeddings vector space to the two-dimensional screen is performed using the t-SNE algorithm (van der Maaten and Hinton 2008). This has the advantage of reducing the vector space in a 2D representation while keeping similar words grouped together. An example is given in Fig. 5, using also the word *cald*.

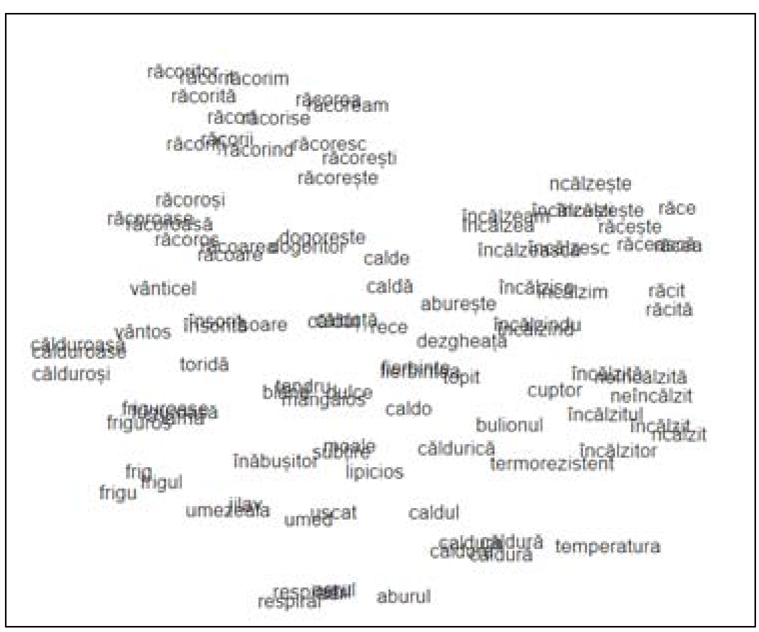

Fig. 5. Graphical representation of words similar to *cald* using the t-SNE algorithm.

Another useful property of word embeddings is finding analogies. This involves searching for a word D which should be similar to another word C in the same way as the word B is similar to word A (in this context, the words A, B, C are known, and D will be the query result). Such a query is resolved by the equation

$$vec(D) = vec(A) – vec(B) + vec(C),$$

where vec(X) is the vector representation of word X. For interacting with the CoRoLa based word embeddings in terms of analogies, two interfaces were created: a basic text query and a graphical representation. The basic analogy interface involves entering the words A, B, C, and the system will provide the word D, with the possibility to also display the corresponding words. An example is given in Fig. 6.

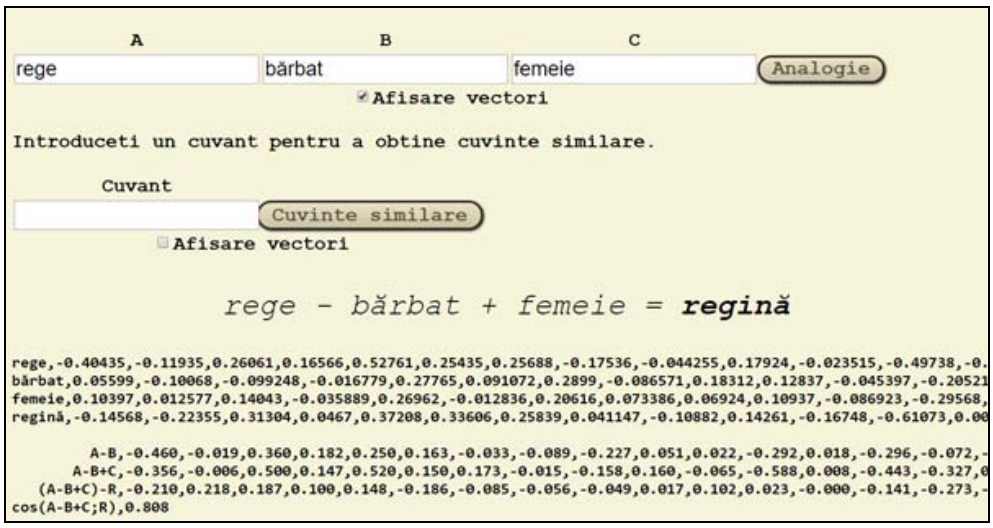

Fig. 6. Simple query interface for analogies.

A graphical interface for representing multiple words on the same graph was constructed, having in mind the analogies. However, this interface can be used to represent any number of words on the same graph. In this case, the nodes are the words, and edges represent similarity between the words involved. An example of this interface is given in Fig. 7.

The graph interface is interactive, in the sense that it can be zoomed or panned using a mouse. Nodes can be clicked, triggering the loading of similar words to the clicked word.

Apart from the basic word embeddings computed using the complete words, additional representations were trained, using lemmas and combination of part-of-speech tags and lemmas. For each of those, new instances of the textual and graphical interfaces were deployed. Part of this work was described by Păiş and Tufiş (2018b).

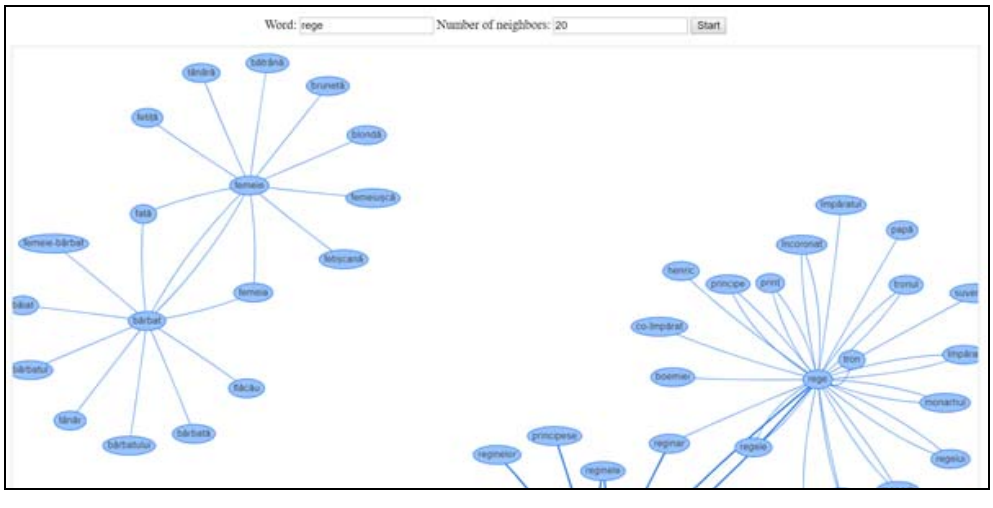

Fig. 7. Example graph representation of words and similarity relations.

10. CONCLUSIONS

After a brief overview of corpora and natural language processing technologies created in time at ICIA, we have focused in this paper on the making of the reference corpus of Romanian. ICIA's corpora, described in this paper, are not the only ones created for the Romanian language. For a presentation of Romanian corpora, we refer a.o. to Barbu Mititelu et al. (2018). The CoRoLa project offers users access to a very large and diverse corpus, which is IPR-cleared, associated with metadata, several layers of annotation and other layers envisaged in short- and medium-term. As mentioned in the paper, creating such a corpus implies long-term maintenance, curation and even further enrichment.

Corpora address users from various domains, be they language scientists or language engineers, students or teachers, native speakers or speakers of other languages, specialists or common people. Given the limitations that describe such a resource, corpora cannot mirror language faithfully. The automatic processing to which they are subject is also limited and never entirely accurate, yet continuously improvable. These are the main reasons why errors will always find their place in corpora, and why linguistic phenomena may sometimes not be encountered at the expected rate. What is more, the interdisciplinary context in which corpora are created asks for a user willing to face the challenge of stepping out of the comfort zone and start using instruments that offer and ease access to the linguistic richness of a corpus.

Creating CoRoLa was an effort that involved many cultural entities in Romania and even abroad. All the contributors are acknowledged on the corpus website. Moreover, there are people who understood the importance of this project and got involved in ways that gave us the possibility to reach out to text providers. We are equally grateful to them all.

REFERENCES

Atkins, S., J. Clear, N. Ostler, 1992, "Corpus Design Criteria", *Literary and Linguistic Computing*, 7. 1–16. 10.1093/llc/7.1.1.

Bański, P., P.M. Fischer, E. Frick, E. Ketzan, M. Kupietz, O. Schonefeld, A. Witt, 2012, "The New IDS Corpus Analysis Platform: Challenges and Prospects", in: N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (eds), *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC 2012), Istanbul, Turkey, 2905–2911.

Barbu Mititelu, V., E. Irimia, D. Tufiş, 2014, "CoRoLa – The Reference Corpus of Contemporary Romanian Language", in N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (eds), *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC 2014), Reykjavik, Iceland, 1235–1239.

Barbu Mititelu, V., D. Tufiş, E. Irimia, 2018, "The Reference Corpus of the Contemporary Romanian Language (CoRoLa)", in N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (eds), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018), Miyazaki, Japan, 1178–1185.

Bibiri, A. D., S. C. Bolea, L. A. Scutelnicu, M. A. Moruz, L. Pistol, D. Cristea, 2015, "Metadata of a Huge Corpus of Contemporary Romanian Data and Organization of the Work", *Proceedings of the 7th Balkan Conference on Informatics Conference*, Craiova, Romania, 35:1–35:8.

Boroş, T., S. D. Dumitrescu, 2015, "Robust deep-learning models for text-to-speech synthesis support on embedded devices", in *Proceedings of the 7th International Conference on Management of computational and collective Intelligence in Digital EcoSystems (MEDES'15)*, Caraguatatuba/Sao Paulo, Brazil, 98–102.

Broeder, D., M. Windhouwer, D. van Uytvanck, T. Goosen, T. Trippel, 2012, "CMDI: a Component Metadata Infrastructure", *Proceedings of the Workshop "Describing LRs with Metadata: Towards Flexibility and Interoperability in the Documentation of LRs"* at LREC 2012, 1-4.

Cristea, D., N. Diewald, G. Haja, C. Mărănduc, V. Barbu Mititelu, M. Onofrei, 2019, "How to Find a Shining Needle in the Haystack. Querying CoRoLa: Solutions and Perspectives", in this volume.

Diewald, N., E. Margaretha, 2017, "Krill: KorAP search and analysis engine", *Journal for Language Technology and Computational Linguistics* (JLCL), 31, 1, 63–80.

Diewald, N., V. Barbu Mititelu, M. Kupietz, 2019, "The KorAP User Interface. Accessing CoRoLa via KorAP", in this volume.

Erjavec, T., 2012, "MULTEXT-East: morphosyntactic resources for Central and Eastern European languages", *Language Resources and Evaluation*, 46, 1, 131–142.

Gîfu, D., A. Moruz, C. Bolea, A. Bibiri, M. Mitrofan, 2019, "The Methodology of Building CoRoLa", in this volume.

Hardie, A., 2012, "CQPweb – combining power, flexibility and usability in a corpus analysis tool", *International Journal of Corpus Linguistics*, 17, 3, 380–409.

Ion, R., 2007, *Word Sense Disambiguation Methods Applied to English and Romanian*, PhD Thesis, Romanian Academy.

Ion, R., E. Irimia, D. Ştefănescu, D. Tufiş, 2012, "ROMBAC: The Romanian Balanced Annotated Corpus", in: N. Calzolari, K. Choukri, T. Declerck, M. Uğur Doğan, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis (eds), *Proceedings of the 8th International Conference on Language Resources and Evaluation* (LREC 2012), European Language Resources Association (ELRA), Istanbul, Turkey, 339–344.

Kupietz, M., H. Lüngen, 2014, "Recent Developments in DeReKo", in: N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (eds), *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC 2014), Reykjavik, ELRA, 2378–2385.

Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, J. Dean, 2013, "Distributed representations of words and phrases and their compositionality", in *Advances in neural information processing systems*, 3111–3119.

Mitrofan, M., D. Tufiş, 2018, "BioRo: The Biomedical Corpus for the Romanian Language", in: N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (eds), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018), Miyazaki, Japan, 1192–1196.

Mitrofan, M., V. Barbu Mititelu, G. Mitrofan, 2018, "Towards the Construction of a Gold Standard Biomedical Corpus for the Romanian Language", *Data*, 3 (4), 53.

Păiş, V., D. Tufiş, 2018a, "Computing distributed representations of words using the CoRoLa corpus", *Proceedings of the Romanian Academy*, Series A, 19 (2), 403–409.

Păiş, V., D. Tufiş, 2018b, "More Romanian word embeddings from the ReTeRom project", in *Proceedings of the International Conference on Linguistic Resources and Tools for Processing Romanian Language* (ConsILR), Iaşi, Romania, 91–100.

Sinclair, J., 1996, "Reference Corpora", in *Preliminary recommendations on corpus typology*, EAG-TCWG-CTYP/P, EAGLES, http://www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html.

Steinberger, R., B. Pouliquen, A. Widiger, C. Ignat, T. Erjavec, D. Tufiş, 2006, "The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages", in *Proceedings of the 5th LREC Conference*, Genoa, Italy, 2142–2147.

Tufiş, D., D. Cristea, S. Stamou, 2004, "BalkaNet: Aims, Methods, Results and Perspectives. A General Overview", *Romanian Journal of Information Science and Technology,* Special Issue, 7, 1-2, 9–43.

Tufiş, D., R. Ion, Al. Ceauşu, D. Ştefănescu, 2005, "Combined Aligners", *Proceeding of the ACL2005 Workshop on "Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond"*, Ann Arbor, Michigan, 107–110.

Tufiş, D., E. Irimia, 2006, "RoCo_News - A Hand Validated Journalistic Corpus of Romanian", *Proceedings of the 5th LREC Conference*, Genoa, Italy, 869–872.

Tufiş, D., R. Ion, Al. Ceauşu, D. Ştefănescu, 2006, "Improved Lexical Alignment by Combining Multiple Reified Alignments", *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics* (EACL 2006), Trento, Italy, 153–160.

Tufiş, D., V. Barbu Mititelu, E. Irimia, S. D. Dumitrescu, T. Boroş, 2016, "The IPR-cleared Corpus of Contemporary Written and Spoken Romanian Language", in N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (eds), *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016), Portorož, Slovenia, 2516–2521.

van der Maaten, L.J.P., G.E. Hinton, 2008, "Visualizing High-Dimensional Data Using t-SNE", *Journal of Machine Learning Research*, 9, 2579–2605.