COROLA – THE REPRESENTATIVE CORPUS OF THE CONTEMPORARY ROMANIAN LANGUAGE. THE INITIAL PHASE

Verginica Barbu Mititelu, PhD, Ștefan Daniel Dumitrescu, PhD, Romanian Academy Research Institute for Artificial Intelligence "Mihai Drăgănescu"

Abstract: We present here the project of creating CoRoLa, a representative corpus of contemporary Romanian (from 1945 onwards). In the international context, the project finds its place among the initiatives of gathering huge collections of texts, of pre-processing and annotating them at several levels, and also of documenting them with metadata. Our project is a joined effort of two institutes of the Romanian Academy. We foresee a corpus of more than 500 million word forms, covering all functional styles of the language. Although the vast majority of texts will be in written form, we target about 300 hours of oral texts, too, obligatorily with associated transcripts. Most of the texts will be from books, while the rest will be harvested from newspapers, booklets, technical reports, etc. The pre-processing includes cleaning the data and harmonising the diacritics, sentence splitting and tokenization. Annotation will be done at a morphological level in a first stage, followed by lemmatization, with the possibility of adding syntactic, semantic and discourse annotation in a later stage. The target users of our corpus will be researchers in linguistics and language processing, teachers of Romanian, students.

Keywords: reference corpus, contemporary Romanian, annotation, metadata, corpus structure.

1. Introduction

The importance of corpora for the study of a language and, implicitly, for the development of further resources for its electronic analysis, with a view to the development of applications involving language processing is widely accepted by researchers nowadays. A proof for this is the effort having been made (and even re-made, see the case of British National Corpus, for which an increase in size is envisaged), for several decades already, by various nations for gathering large collections of real language samples.

The late multilingual projects with focus on several languages (related or not) have led to the inclusion of Romanian among the "important" languages (i.e., languages for whose study and resources development a lot of both human and financial effort has been invested). As a consequence, the lack of various resources (a corpus, inter alia) for Romanian has been felt more intensely and expressed as such in papers, whose authors blamed it on this lack that their research could not be carried out in a desirable way or similar to the research for those languages and they had to appeal to various less convenient (from various perspectives) alternatives.

In order to fill this gap in the pool of linguistic resources for Romanian, the Romanian Academy Research Institute for Artificial Intelligence "Mihai Drăgănescu" started a project of creating a representative corpus of contemporary Romanian in 2012. One of the preliminary steps was conducting an online survey meant to find out Romanians' expectations from such a

project. We launched a call for participation to this survey on various email lists. Sixty-five people took part and 57% of them are linguists and 35% are computer scientists. Most of them are researchers (66%) and teachers (54%) (there is an expected overlap of these two professions given the number of people having such jobs in parallel), but 8% are employed in the industrial companies. All of them are well aware of the possible uses of such a corpus, as they enumerated the activities in which they themselves would use it (linguistic research (70%), developing applications involving processing of Romanian (51%), lexicography (42%), teaching Romanian (42%), translation, etc.), and they also defined the type of searches they would perform on the corpus: (lexical or morpho-syntactic) contexts of occurrence of a word or word form (86%, and 70% respectively), meanings of a word in context (77%), relative word frequency (70%), lexicalisations of various syntactic structures (59%), co-occurrences (54%), collocations (52%) and many others. As to the types of texts they would search into, the answers are also diverse: newspaper or magazine articles (82%), news (75%), scientific texts (72%), fiction (62%), administrative texts (54%), texts from various Internet sources, oral texts, short texts from advertisements and many other types.

What becomes obvious from this survey is that great need for such a resource does exist and the potential users are prepared for using it and for grounding their future activities in it. The responsibility of offering them a qualitative resource is overwhelming and further motivated us to increase our efforts. Since 2014 our partner in the implementation of this project has been the Institute of Computer Science of the Iași branch of the Romanian Academy. Even so, the number of people involved is still insufficient and this is even worse when the financial resources for hiring more people are absent.

2. International context

Corpora around the world have been created for many languages: the Mannheim German National Corpus (http://www1.ids-mannheim.de/kl/projekte/korpora/archiv.html), the Russian National Corpus (http://ruscorpora.ru), the Czech National Corpus (http://ucnk.ff.cuni.cz), the Bulgarian National Corpus (Koeva et al., 2012) are only a few such examples. Given the size of the enterprise, the effort required and the underlying national interest, some projects have been developed by consortia comprising important institutions: see the British National Corpus (http://www.natcorp.ox.ac.uk/).

There is quite a significant number of reference corpora around the world, among which we mention:

- the reference corpus of contemporary Spanish (http://www.rae.es) – containing electronic written and oral texts from 1975 to 2004, totalling 160 million word forms, belonging to a very wide range of genres and domains; the texts are not annotated;
- the reference corpus of Estonian (http://www.keeletehnoloogia.ee/projects-1/the-reference-corpus-of-the-estonian-language) – containing electronic written text, totalling 245 million word forms, 75% of them coming from newspapers; the texts are morphologically annotated (Kaalep et al., 2010);
- the German reference corpus DeReKo (Kupietz et al., 2010) – containing already tens of billions of words, morpho-syntactically annotated; the developers did not aim at having a representative corpus, let alone a balanced one; all available texts are harvested and it is the user who selects the components (s)he wants to base his/her research on;

- the reference corpus of contemporary Portuguese (http://www.clul.ul.pt/en/research-teams/183-reference-corpus-of-contemporary-portuguese-crpc) – containing more than 310 million word forms in written and oral texts, covering a wide range of text genres and of language varieties; the texts are morphologically annotated.

It is obvious that they do not share the same principles of corpus design, thus becoming mandatory for each developer to make their working principles known to the community and the users.

Analysing the structure of such corpora (Barbu Mititelu and Irimia, 2014), we noticed that the oral aspect of language was of interest in many initiatives. This component is present in various percents, with 10% the highest degree of representation, in British National Corpus and Polish National Corpus. When oral components are included in the corpus, they are either transcribed (and the transcription is subject to (roughly) the same processing and annotation as the written text) or left unprocessed.

A first classification of texts is usually made into informative versus imaginative ones, the latter representing, on average, around a fifth of the whole corpus, although the Russian corpus contains almost 40% of fictional texts.

Considering the medium from which the corpus is taken, the practice is very diverse: the British favoured the books, which provide 60% of the texts, whereas newspapers and magazines provide 30%; the rest of 10% comes from brochures, leaflets, manuals, advertisements, letters, memos, reports, minutes, essays, etc. The Bulgarians could harvest only 2.5% of texts from books, newspapers and magazines, while the Internet provided the vast majority of texts included in their national corpus (97.5%). Half of the Polish texts represent the journalistic style and less than a quarter of the texts come from books. Other corpora have other distributions of their texts from the medium perspective.

3. Our aims

The representative corpus of the contemporary Romanian language (CoRoLa) will be a big corpus (we target at least 500 million word forms), in which all functional styles will be represented. It will contain both written and oral texts. They will be pre-processed and annotated (at least at the morphological level, but we also envisage a syntactic and even semantic and discourse annotation).

The vast part of the corpus will contain texts originally written in Romanian, although a part of the final corpus could be represented by translations from various domains, with the shortcomings specific to such texts, such as the foreign influence especially at the lexical or the syntactic levels.

All functional styles should be covered: scientific, official, publicistic and imaginative. The colloquial style will definitely be included due to its use in imaginative writing, although we foresee problems in its processing and annotation, given its linguistic characteristics.

We will collect texts from all domains that we will have access to. Most texts will be extracted from books, but newspaper articles, booklets, theses and technical reports will not be left aside.

The oral component will be represented by around 300 hours of recordings obligatorily accompanied by their transcript. The transcripts will be processed in a way similar to the processing of written texts. For the oral recordings, we will automatically generate speech segmentation at phoneme level. This will be auxiliary to any annotation and segmentation

already present in the corpora and will enable research in the fields of prosody and speech analysis.

The collecting process will be accompanied by metadata creation. We will devote special attention to the specification of the metadata schemes for corpus and document level description, following standards recommended in the community.

CoRoLa will be developed and refined in successive steps and the automatic processing chain of the texts to be included has to conform to the format requested by the indexing and searching platform (in our case, tabular codification, with XML-type annotations).

4. Steps taken so far

A reference corpus is designed to provide comprehensive information about a language (Sinclair, 1996). In order to attain this aim, it has to contain all relevant language varieties and the characteristic vocabulary. We want our corpus to be representative for the contemporary phase of the language.

4.1. Corpus design

In designing CoRoLa's structure we considered mainly the other corpora existing in the world (as presented in section **2.** above), correlated with the results of a survey about Romanians' reading preferences (http://ivox.ro/download/get/f/raport-cat-cum-si-ce-citesc-romanii-2012, accessed on August 29th, 2014): according to it, most people read books (in our terms, imaginative writing) (28.47%) (correlated with the main reason for reading identified in this study, namely for pleasure and relaxation, expressed by 38.85% of the subjects, and with the leading group of 15.59% of people who love reading fiction), a slightly lower percent of people (27.4%) read articles from online magazines, 21.49 percent of people read printed newspapers and magazines, 10.37% of people read online scientific articles and 9.69% of the readers read bloggs.

We aim at the following structure for CoRoLa (for details and a more refined classification, see Barbu Mititelu and Irimia, 2014):

- 10% oral texts and 90% written ones. The former will reflect continuous speech and will have the transcribed counterpart, as well;
- the majority of texts will come from books (60%), almost a third from newspapers and magazines (30%), while other sources (such as blog posts) will contribute 10%;
- the distribution of functional styles is presented in Table 1. Memoirs are not recognized as a functional style, but given their characteristics, we chose to treat them separately. The last column of the table contains the feeders of each style, i.e. those that have already signed a written agreement to allow us to introduce their texts in the corpus, to process them and make them available for searching for those interested;
- as far as the domains are concerned, we propose the distinction in 4 main domains with specific subdomains:
    - Arts & Culture: Literature, Art History, Folklore, Film, Architecture, Sculpture, Painting & Drawing, Design, Fashion, Theatre, Music, Dance, Others;

- Society: Politics, Law, Administration, Economy, Army, Health, Sport, Family, Gossip, Social Events, Education, Social Movements, Tourism, Religion, Entertainment, Others;
- Nature: Environment, Natural Disasters, Universe, Natural Resources, Others;
- Science: Mathematics, Informatics, Logics, Standards, Medicine, Archaeology, Engineering, Architecture, Technics/technology, Aeronautics, Agronomy, Metrology, Criminalistics, Constructions, Military Science, Pharmacology, Oenology, Pedagogy, Geography, Economy, History, Psychology, Sociology, Ethnology, Anthropology, Religious Studies and Theology, Juridical Sciences, Linguistics, Political Sciences, Philosophy, Philology, Biology, Physics, Astronomy, Chemistry.

Table 1. Styles distribution in the written component of CoRoLa and their feeders.

Style	**Percent in the written component**	**Feeders**
Imaginative	25	Humanitas, Polirom, România literară, the journal of Colegiul Național „Unirea" from Focșani, Destine literare
Memoirs	5	Humanitas, Polirom, Editura PIM
Law	10	
Administrative	10	
Science	30	Humanitas, Polirom, Editura Academiei, Editura Universității din București, Editura Economică, Editura Simetria, Muzica, România literară, Editura PIM
Journalistic	20	DCNEWS, România literară, Actualitatea muzicală, România literară, Destine literare

Both great and modest names occur in our list of contributors. It is normal for us to target the important publishing houses, as the readers focus mainly on fiction from books. They can offer "big names" as far as the list of authors is concerned, as well as quality texts, as far as orthography and text format is concerned. However, as the process of persuading publishing houses and media to become our partners in this project is sometimes quite slow (for reasons varying from one potential partner to another), we welcome whomever offers or is easily persuaded by our team or our collaborators to join our efforts.

The law and administrative texts are outside the scope of the copyright law, so we can freely download such texts and add them to the corpus.

4.2. Data harvesting

Texts collection is a difficult task when the intellectual property law applies. The categories of content excepted by the law are: political, legislative, administrative and judicial. For the other domains, we can freely use fragments of no more than 10,000 characters. However, this is a small amount of text if we think of the large part of the fiction (novels) and scientific books, for example, in our corpus. Given the type of facilities we want to offer to users, we need continuous fragments from larger texts, instead of short fragments from different parts of a long text. Moreover, we must consider only texts written with diacritics (otherwise, the linguistic

annotation will be highly incorrect) and we need to ensure ourselves that only the correct type of diacritics is used, especially that the standard was changed several years ago.

To ensure the volume and quality of the texts to be included in the corpus, as well as copyright agreements on these texts, our endeavour was to contact publishing houses and editorial offices representatives and to find solutions for collaboration. We targeted important publishing houses, which publish Romanian contemporary writers. So far (October 2014), we have signed agreements with the following publishing houses: Humanitas, Romanian Academy Publishing House, Bucharest University Press, Polirom, "Editura Economică", Simetria, PIM. Some magazines and newspapers have also agreed to help our project by providing access to the text of their articles: Destine literare, România literară, Muzica, Actualitatea muzicală, DCNEWS, the magazine of Unirea National College from Focșani. Until now two bloggers have also agreed to allow us to include some of their posts in the corpus: Simona Tache (http://www.simonatache.ro) and Dragoș Bucurenci (http://bucurenci.ro). Oral texts (read news, live transmissions and live interviews) (one hour per working day) are provided by Rador, the press agency of Radio Romania. Their readiness to get involved was a very pleasant surprise for us. We negotiated the conditions for our collaboration, very important aspects being our free access to these texts and the possibility of disseminating our work and results.

For texts available online (from DCNEWS, România literară and blogs) we have developed crawlers that extracted the texts of interest for us. Otherwise, we received mainly .pdf files, but also a few .doc files. The oral texts are received as .mp3 files and their transcriptions as .doc files.

4.3. Metadata creation

The importance of metadata creation for the documentation of the corpus content is straightforward. Metadata contain general information such as the creators of the corpus, the availability and the licence, the development status, the projects and cooperation agreements that support the creation, etc. and specific information at the document level like the author of the metadata and of the manual pre-processing work, annotation details (tools, level of annotation, validation of annotation, etc.), the author, source, type and genre of the text, the number of words and other statistics for the document. Some of the information specified in the metadata at the document level is essential for the indexing of the corpus and the facilitation of the searching process for the end users.

As we get texts either as electronic files of by crawling them, we have created metadata in two ways: manually and automatically. The manual method used mainly an application for both creating metadata and extracting text from .pdf files (Moruz and Scutelnicu, 2014), but while this tool was under construction or debugging, we used Arbil (https://tla.mpi.nl/tools/tla-tools/arbil/), which is only a metadata editor, as an alternative, while the text was simply copied from the pdf file in a text editor or automatic online conversion was made when possible (i.e., when diacritics or other characters were not affected by the process of automatic conversion). We assured that the output metadata files format was the same for both applications.

For the text crawled from DCNEWS, România literară and the blogs, we created the metadata automatically, at the crawling moment. We took advantage of the fact that on the news website and in the online magazine texts are already classified. We extracted the classifications and mapped the categories on our own set of categories. Thus, the automatic creation of the metadata file for each crawled article, with the same structure as the metadata created manually,

became possible. A total of 76868 txt files have been crawled so far and have automatically associated metadata.

4.4. Texts correction

In order to prepare the texts for processing and annotation, we decided to correct the misspellings. They were identified against a comprehensive lexicon of inflected forms. Nevertheless, besides various types of misspellings (missing letters (*priejul* instead of *prilejul*), inversed letters (*entobotanicele* instead of *entobotanicele*), extra letters (*săpătămâna* instead of *săptămâna*), missing dashes (*miam* instead of *mi-am*), missing diacritics (*insemnind* instead of *însemnînd*), missing spaces (*înlimbaromână* instead of *în limba română*), multiple phenomena (*avengură* instead of *anvergură*)), besides the hesitant spelling of recent borrowings for which only one form exists in the lexicon (*cannabis* and *canabis*), besides tendencies specific to this phase of the language (*servici* instead of *serviciu*) and besides the foreign words (*Cronică literară & beyond*), there are also valid word forms that are marked in order to be included in the lexicon (probably after a further filtration): they are: recent borrowings (*dronă*), ad-hoc creations (*umorism, arhicanalia*) (this is the type abounding in the texts from România literară), dialectal words (*peleg*).

5. Results of the initial phase

This one-year work in partnership between the two institutes can boast several achievements: first, the attraction of a growing group of partners ready to offer us free access to electronic versions of different types of documents and recordings; second, we have developed a tool for extracting text from pdf files and creating metadata; third, we have extracted texts from the written documents; fourth, we created metadata files for all these texts; fifth, we have partially corrected these texts. The charts below present some figures according to various criteria: document type, text styles, text domains and text domains that are present so far in our collection of texts. The numbers are to be interpreted as thousands of word forms. It is obvious that, for the moment, most of our texts are newspaper articles, which is in contradiction with our aims. However, for the moment, we have harvested whatever type of texts we could get and the balancing according to our provisioned structure (Barbu Mititelu and Irimia, 2014) will be done in the next phases.

Figure 1. Statistics about the document types.

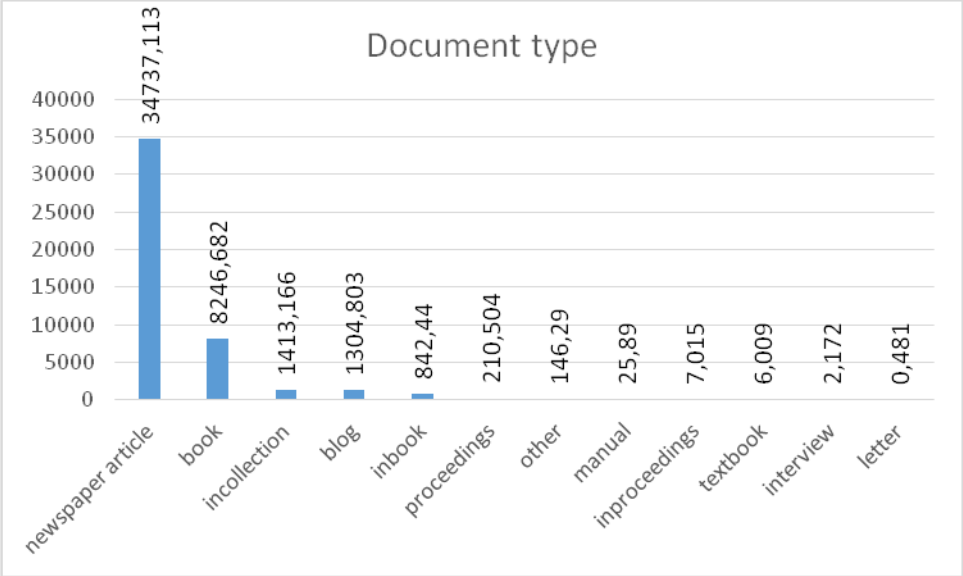
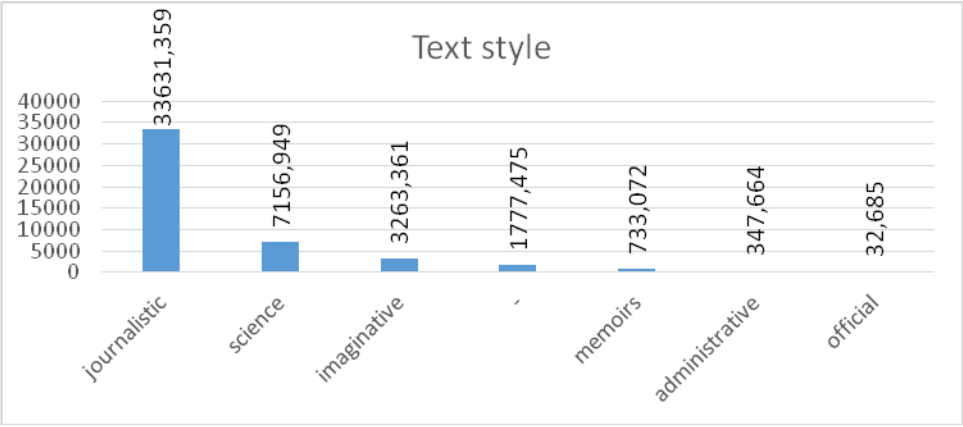
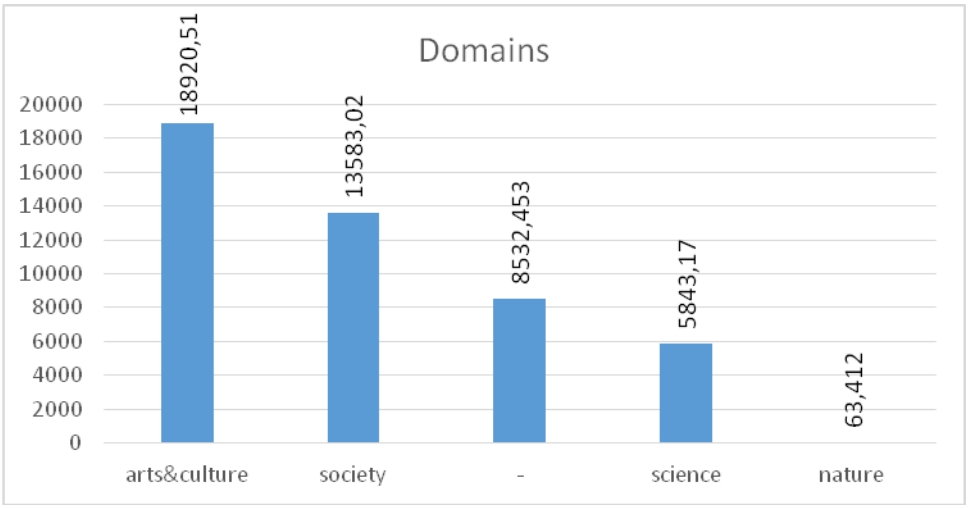


Figure 2. Statistics about the functional styles.



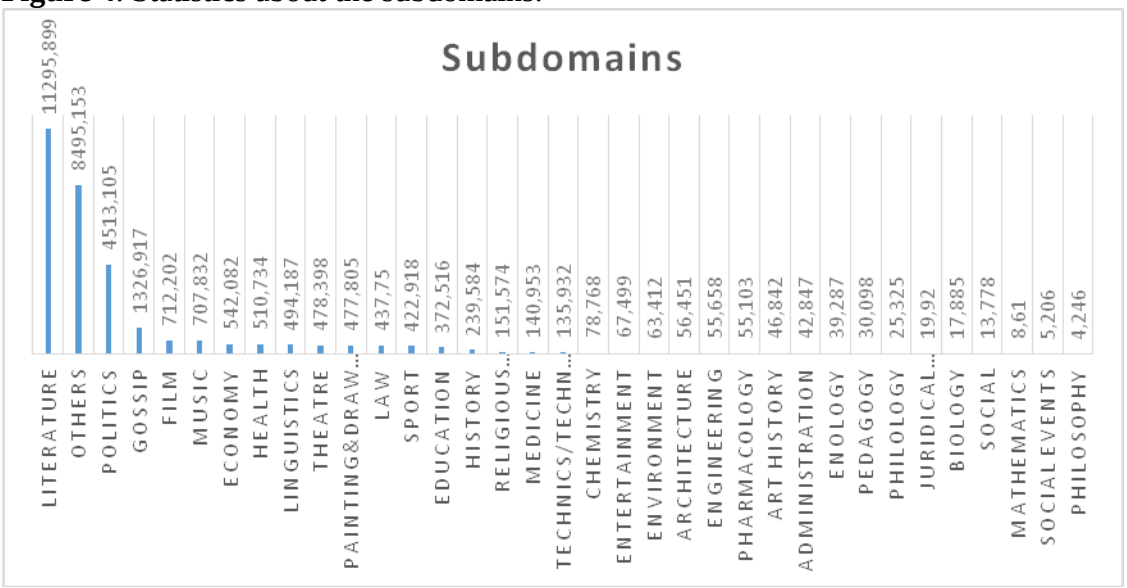
We have managed to cover all functional styles, in completely diverse proportions. We can notice some texts totalling 1777475 word forms are not assigned a certain style. They are mostly texts from blog posts and are kept apart for the moment (at least).

Figure 3. Statistics about the domains.



All domains of the informative styles are represented (arts & culture with the highest number of texts, and nature with the smallest). The dash covers the texts for which domain specification is not applicable: imaginative writings, blog posts (mainly).

Figure 4. Statistics about the subdomains.



These are the subdomains from which we have informative texts with associated metadata. There are many texts for which no subdomain (from our predefined list) could be selected, thus the presence on the third position of the "Others".

6. Future work and Conclusions

After finishing the correction of these texts, they are to enter the processing and annotation phases, which will be followed by manual correction of a small percent of the annotated texts (2%). The correction of the annotation will be meant to help the annotating tool to improve its results. Meanwhile, all the steps mentioned above will be taken again for other texts. New partners are continuously looked for in order to get access to as many texts as possible.

The corpus will be made available for those interested, mainly for search purposes. As established in the agreements signed with the publishing houses and editorial offices representatives, the annotated text fragments cannot be made available for download. However, the results of the search in the texts outside the scope of such restrictions will be downloadable.

Bibliography

Barbu Mititelu, V., Irimia, E., (2014). *The Provisional Structure of the reference Corpus of the Contemporary Romanian Language (CoRoLa)*, in M. Colhon, A. Iftene, V. Barbu Mititelu, D. Cristea, D. Tufiș, *Proceedings of the 10th International Conference "Linguistic resources and Tools for Processing the Romanian Language"*, Craiova, 18-19 September 2014, Editura Universității „Alexandru Ioan Cuza", Iași, pp. 57-66.

Koeva, S., Stoyanova, I., Leseva, S., Dimitrova, T., Dekova, R., Tarpomanova, E. (2012). The Bulgarian National Corpus: Theory and Practice in Corpus Design. *Journal of Language Modelling*, vol 0, issue 1, pp. 65-110.

Kaalep, H.-J., Muischnek, K., Uiboaed, K., Veskis, K. (2010). *The Estonian Reference Corpus: its Composition and Morphology-aware User Interface*. In I. Skandina & A. Vasiljevs (Eds.), *Human Language Technologies The Baltic Perspective: Proceedings of the Fourth International Conference Baltic HLT 2010*, pp. 143-146.

Kupietz, M., Keibel, H. (2009). *The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research*. Working Papers in Corpus-based Linguistics and Language Education, No. 3, Tokyo: Tokyo University of Foreign Studies (TUFS), pp. 53-59.

Moruz, A., Scutelnicu, A. (2014). An Automatic System for Improving Boilerplate Removal for Romanian Texts, *)*, in M. Colhon, A. Iftene, V. Barbu Mititelu, D. Cristea, D. Tufiș, *Proceedings of the 10th International Conference "Linguistic resources and Tools for Processing the Romanian Language"*, Craiova, 18-19 September 2014, Editura Universității „Alexandru Ioan Cuza", Iași, pp. 163-170.

Sinclair, J. (1996). *Preliminary recommendations on Corpus Typology*, Tech. Rep. EAG--TCWG--CTYP/P.