CoRoLa – The Reference Corpus of Contemporary Romanian Language

Verginica Barbu Mititelu, Elena Irimia, Dan Tufiș
Research Institute for Artificial Intelligence "Mihai Drăgănescu"
13 Calea 13 Septembrie, 050711, Bucharest, Romania
{vergi, elena, tufis}@racai.ro

Abstract

We present the project of creating CoRoLa, a reference 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 (CMDI). 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. A core of CoRoLa is described in the article. The target users of our corpus will be researchers in linguistics and language processing, teachers of Romanian, students.

Keywords: reference corpus, Romanian, corpus annotation, corpus design, metadata.

1. Introduction

Given the scarcity of public corpora for Romanian and the reduced availability of those existent even for searching purposes, in 2012 the Romanian Academy Research Institute for Artificial Intelligence from Bucharest (RACAI) started a project for defining a powerful infrastructure for collecting texts and speech, annotating them, making them available for searching by those interested, and also making public various statistics based on them. Since 2014 this initiative has been joined by the Institute for Computer Science in Iasi, in a larger priority project of the Romanian Academy: The Reference Corpus of Contemporary Romanian Language. The experience in NLP and Speech (pre-processing and processing it), both with monolingual and with parallel data, the available infrastructure, our positions as institutes of excellence of the Romanian Academy make this consortium an appropriate developer of a reference corpus of contemporary Romanian.

As defined by Sinclair (1996), "a *corpus* is a collection of pieces of language that are selected and ordered according to explicit linguistic criteria in order to be used as a sample of the language". A "*computer corpus* is a corpus which is encoded in a standardized and homogenous way for open-ended retrieval tasks. Its constituent pieces of language are documented as to their origins and provenance". According to EAGLES specifications, the reference corpus is designed so that to "provide comprehensive information about a language". This implies that a reference corpus is great in size, so as to cover all relevant language varieties and the characteristic vocabulary, and reflects the general use of a language, thus being possible to consider it a standard, that is a "basis for reliable grammars, dictionaries, thesauri and other language reference materials".

There are aspects which have not been agreed upon as far as the reference corpus is concerned: which are the relevant language varieties? What is the proportion of their representation in the corpus? Given the annotation tools available, how can we ensure all words and word senses are covered?

Given the circumstances, we can foresee the difficulty of our undertaking and we expect criticism about our design of the corpus.

2. Related Work

Many research groups in the community have dedicated their efforts to constructing very large corpora: 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) and many others. 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/). In fact, this is our strategy, as well: to join efforts with other scientific bodies who express their interest in this project and invite others, which we consider appropriate, to fall in.

As far as reference corpora are concerned, there is quite a significant number of initiatives around the world, out of 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.

3. Objectives

CoRoLa will be a big corpus (more than 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).

Aiming eventually at a reference corpus, we have in view, in a first step, the contemporary literary language. Contemporary Romanian is the last phase in the evolution of the language, starting, according to specialists, after the Second World War. Due to historic reasons, to the different political, economic and social transformations that marked the community of speakers, we can further divide this period in a communist stage and a post-communist or post-revolutionary one. The main differences between them are visible at the vocabulary level: words frequencies, words creation mechanisms, borrowings. We want to represent both periods in the corpus, although it is evident that the latter is more easily attainable, given the existence of texts in electronic format, whereas for the former we need to use printed materials, scan, OCRize and correct them.

The vast part of the corpus will contain texts originally written in Romanian. We intend that a part of the final corpus should be represented by translations from various domains. Although translated texts may be influenced (at the lexical or the syntactic level) by the originals, this is a phenomenon affecting language and it must be recorded.

The texts to be included in the corpus will be selected so that all functional styles should be covered: scientific, official, publicistic and imaginative. Although the colloquial style is not a major concern for us, it will definitely be included due to its use in some imaginative writing.

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 300 hours of recordings 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 relying either on the Hidden Markov Model Toolkit (HTK) or CMU Sphinx. 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 must be accompanied by the imperative task of 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 (see section 6.2).

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). The platform we chose is IMS Open Corpus Workbench (CWB, http://cwb.sourceforge.net/), an open source medium that allows complex searching with multiple criteria and support for regular expression. It also offers the user the following possibilities: to choose the (sub)corpus/(sub)corpora with which to work (choose from among the domains and subdomains, but also from the available authors), to find out words frequencies in a (specified) (sub)corpus, to search for a word or a word form, to search for more words (either consequent or permitting intervening words), to find words collocations and co-occurrences (within a window of a pre-established size), to find lexicalization of specified morphological or/and syntactic structures, n-gram models, etc. It was already installed and tested on ROMBAC corpus (Ion et al., 2012) at RACAI and coupled with our processing chain which produces the adequate annotated format for morphological and shallow syntactic searches. The processing chain currently includes the TTL web services (Tufiş et al., 2008), available at http://ws.racai.ro/ttlws.wsdl, offering, at the moment of this writing, the following specific functionalities: sentence splitting, tokenisation, tagging, lemmatising and chunking. Future services regarding processing and query facilities for discourse (Cristea & Pistol, 2012) will be provided. CoRoLa will be automatically annotated, but a fragment of it (~2%) will be manually validated.

The corpus users that we have in mind include, but may not be limited to: linguists, students, textbook authors and language engineers. At a survey we launched two years ago, 65 potential users participated and expressed their need to search the corpus for lexical contexts of words, for their meanings in contexts, for words relative frequencies, for morpho-syntactic contexts, lexicalizations of a morpho-syntactic structure, for morpho-lexical realization of a syntactic function, co-occurrences, collocations, word families, n-gram models, etc.; such

information is relevant for their research, teaching or language processing. The search interface will allow them to choose the subcomponents relevant for their study.

4. Text Collection

Collecting texts to include in the corpus is, as easily imaginable, a difficult task, given the intellectual property law. 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 novels and scientific books, for example. 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 (March 2014), we have signed agreements with the following publishing houses: Humanitas, Polirom, Romanian Academy Publishing House, Bucharest University Press, "Casa Cărții de Știință" Publishing House, "Editura Economică". Some magazines and newspapers have also agreed to help our project by providing access to the text of their articles: România literară, Muzica, Actualitatea muzicală, DCNEWS, the school 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). Their readiness to get involved was a very pleasant surprise for us. We established together the conditions for our collaboration. A very important aspect is the fact that our access to these texts is free.

As established in the agreements signed with the publishing houses and editorial offices representatives, the annotated text fragments cannot be made available for download. Access to the corpus will be possible only through the CWB interface, which displays small fragments of linguistically annotated text. We will ensure secured access to the corpora, in order to prevent misuse and vandalism.

The informed users may search the current version of the corpus using the CQP query language (Hardie, 2012). For users not familiar with CQP, a natural language interface[1] (constraint Romanian) offers most frequent search facilities. For instance, the interface accepts requests like the ones below, translates them into CQP queries, displays the translation for editing, validation (or just for learning purposes), and finally executes the CQP code:

<list 10 sentences containing in any order the lemmas "car" and "drive" >, or

<list all the sentences containing lemma "drive" followed by lemma "car" at most 3 words distance>

The interface allows the user to browse and/or save the results of his queries.

5. Current Statistics

At the moment, the corpus contains the data presented in Table 1, where one can notice the domain distribution of the texts, as well as quantitative data related to each domain: number of sentences, tokens (word forms and punctuation) and words:

	Sentences	Tokens	Words	Content words
News	651,872	10,294,016	8,558,619	4,662,528
Medical	603,161	10,950,271	9,163,029	5,226,837
Legal	659,646	9,067,516	7,482,484	4,247,737
Biogr.	314,368	5,802,961	4,298,493	2,567,427
Fiction	517,803	8,002,596	6,773,648	3,531,156
Total	**2,746,850**	**44,117,360**	**36,276,273**	**20,235,685**

Table 1. Domain distribution and quantitative data.

Wiki-Ro, the Romanian part of a big collection of sentences extracted from Wikipedia within the ACCURAT European project (http://www.accurat-project.eu/), is sentence split and tokenized and contains 2,747,411 sentences and 30,992,034 words. The documents will be classified using the Wikipedia categories graph so as to match the list of domains represented in CoRoLa.

Besides the written element, the corpus also contains an oral subpart, made up of prosodic annotations (Boroș et al., 2014) for 7022 syllables.

6. Data pre-processing and annotation of the written documents

6.1. Pre-Processing

The first step of the pre-processing stage is the conversion of the documents provided by our partners. Depending on the type of text, e.g. an article in the pdf version of a magazine or on a website, a fragment in a book or a whole book, the work of extracting the information is more or less tedious and has to be done manually, giving the discrepancies in the structure of the materials received. Moreover, the manual extraction ensures better quality of the data, reducing the possibility of unwanted boilerplates to occur in the texts, and allows for the collection of detailed information about the document for the metadata creation in this initial pre-processing stage. When copied from their original sources, the content is converted into the UTF-8 encoding and saved as plain text documents.

In the next stage, the text documents are automatically corrected to eliminate unrecognised characters, empty or delimitation lines, headers, notes, captions and other

[1] The interface has been implemented by Radu Ion.

redaction elements which interrupt the text continuity.
Special attention is dedicated to diacritics, which are essential in Romanian since they often differentiate morphologically and semantically between words. Therefore, we made sure that all the texts introduced in CoRoLa are written with diacritics and that they are normalised. For two of the letters in the Romanian alphabet, *ş* and *ţ*, and their correspondent capital letters, 2 variants of characters were used: one with a comma accent, and another one with a cedilla, due to the fact that initially, the comma accent characters, preferred by the Romanian typography, where not encoded in the Unicode Standard. Since 2009, only the comma accent variant has been recommended by the Romanian National Standardisation Body (ASRO). Accordingly, we automatically convert all the concerned characters to this recommended standard.
As far as orthography law is concerned, we do not normalize the texts, as three different laws were in effect during the period we focus on (from 1945 onwards).

6.2. 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.
Along the years, digital metadata for language resources and tools were created at local or research community level, without the concern for standardisation. Together with the global initiatives for common language resources and tools infrastructures like CLARIN (http://www.clarin.eu/) and MetaShare (http://www.meta-net.eu/meta-share/index_html), the necessity to harmonise the metadata, from common data categories to reusable metadata profiles and schemes, has become evident.
Envisaging future collaboration with international structures, like the integration of our corpus in a transnational infrastructure, we adopted the CMDI (Component MetaData Infrastructure) approach and tools for the creation of our metadata. CMDI, initiated in CLARIN, proposes a component-based approach: the creator of metadata can combine several metadata components (sets of metadata elements) into a self-defined scheme, called "profile". For this purpose, CLARIN created the Component Registry (an online common metadata repository) in which any user can browse already designed components and profiles and can create, edit, register and store its own. The reuse of components and profiles is the key facility that Component Registry offers and the guiding principle of the CMDI initative. Any newly created element must be associated with a data category in ISOcat[2] (and through it to a unique administrative identifier), which is the common scheleton that connects and harmonises all the different components and profiles in the Component Registry.
The metadata editor recommended by CLARIN is Arbil (http://tla.mpi.nl/tools/tla-tools/arbil/). It may be downloaded and installed locally by the metadata creators, but each component and profile made public in the online Component Registry is imediately available for loading by Arbil if the machine is connected to the internet.
Starting from detailed CMDI profiles created in the CLARIN project for annotated text and speech corpora, we have designed profiles tailored to our specific needs.

6.3. Annotation of the data

As mentioned before, a processing chain has been established, consistent with the tabular encoding specific to the CWB platform and comprising more program modules that execute particular functions. The chain is based on the web service TTL (Tufiş et al., 2008), available at http://ws.racai.ro/ttlws.wsdl and it provides:

- sentence splitting: it uses regular expressions for the identification of a sentence end;
- tokenization: the words are released of the adjacent punctuation marks, the compound words are recognized as a single lexical atom and the cliticized words are separated as distinct lexical entities;
- part-of-speech tagging with MULTEXT-East[3] tag set: it is a reimplementation of the HMM statistical tagger (Brants, 2000) for the Tiered Tagging strategy (Tufiş, 1999); its accuracy is more than 98%;
- lemmatization: based on the tagged form of the word, it recovers its corresponding lemma from a large (more than 1,200,000 entries) human-validated Romanian word-form lexicon; the precision of the algorithm measured on running texts is almost 99%; for the unknown words (which are not tagged as proper names), the lemma is provided by a five-gram letter Markov Model-based guesser, trained on lexicon lemmas with the same POS tag as the token being lemmatized. The accuracy of the lemma guesser is about 83%.
- chunking: for each lexical unity previously

[2] ISOcat is an implementation of the ISO 12620:2009 standard (dedicated to the specification of data categories and management of a Data Category Registry for language resources).

[3] see http://nl.ijs.si/ME/V3/msd/html/msd.html

tagged and lemmatized, the algorithm assigns a syntactic phrase, guided by a set of regular expression rules, defined over the morpho-syntactic descriptions.

For the further stages in the corpus development, we envisage adding other types of annotations: deep syntactic parsing, semantic annotation and discourse analysis.

6.4. Annotation correction

In our previous experiments with the task of collecting corpora and ensuring a satisfying quality of the resources, we implemented a coherent methodology for the automatic identification of annotation errors. Most of the error identified in this manner can be also automatically corrected. This validation procedure was used in the past to correct tagging and lemmatization errors for the journalistic corpus AGENDA (Tufiş & Irimia, 2006) and for ROMBAC (Ion et al., 2012), the Romanian balanced corpus designed at RACAI and reduced the estimated error rates to around 2%.

The TTL processing workflow explicitly marks the out-of-dictionary words (ODW), excepting proper nouns, abbreviations and named entities. The ODW can be extracted, sorted and counted, then divided into frequency classes. In the past, we concentrated our analysis on the words with at least 2 occurrences in the corpus (assuming that the others are typographic errors or foreign words) and structured them into error classes, thus being able to split them into errors that need human correction and errors that can be dealt with by implementing automatic correction strategies.

Besides using the described methodology to improve the quality of the entire corpus, we intend to human validate a percentage of it (2% i.e 10 million words). As the process of collecting and managing such an important resource is a life-time task, our attention on assuring its quality will continuously accompany this enterprise.

7. Conclusions

In the international context of growing preoccupation for creating huge language resources, we presented here the initial phase of the creation of a reference corpus of contemporary Romanian. It is a joined effort of two academic institutes, which is greatly helped by publishing houses and editorial offices, which kindly accepted to provide us texts free of charge. The corpus will be available for search for all those interested in the study or processing of the Romanian language.

8. References

Boroş, T., Stan, A., Dumitrescu, S.D. (2014). RSS-TOBI - A Prosodically Enhanced Romanian Speech Corpus. Proceedings of 9th LREC 2014, Rejkyavik, Iceland, 25-31 May, 2014, European Language Resources Association (ELRA).

Brants, Th. (2000). TnT – a statistical part-of-speech tagger. 2000. In Proceedings of the 6th Applied NLP Conference, ANLP-2000, Seattle, WA, pp 224-231.

Cristea, D. and Pistol I.C. (2012). Multilingual Linguistic Workflows. In Cristina Vertan and Walther v. Hahn (Eds.) Multilingual Processing in Eastern and Southern EU Languages. Low-resourced Technologies and Translation, Cambridge Scholars Publishing, UK.

Forăscu, C. and Tufiş D. (2012). Romanian TimeBank: An Annotated Parallel Corpus for Temporal Information. In Nicoletta Calzolari et al. (Eds.) Proceedings of the 8th LREC, Istanbul, Turkey, 21-27 May, 2012, European Language Resources Association (ELRA).

Ion, R., Irimia, E., Ştefănescu, D. and Tufiş. D. (2012). ROMBAC: The Romanian Balanced Annotated Corpus. In Nicoletta Calzolari et al. (Eds.) Proceedings of the 8th LREC, Istanbul, Turkey, 21-27 May, 2012, European Language Resources Association (ELRA).

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

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.

Koeva, S., Stoyanova, I., Leseva, S., Dimitrova, Ts., Dekova, R., Tarpomanova, E. (2012). The Bulgarian National Corpus: Theory and Practice in Corpus Design. *Journal of Language Modelling*, Vol. 0, No. 1, pp. 65--110.

Kupietz, M., Belica, C., Keibel, H., Witt, A. (2010). The German Reference Corpus DeReKo: A Primordial Sample for Linguistic Research. In Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10), pp. 1848--1854.

Sinclair, J. *EAGLES – Preliminary recommendations on Corpus Typology* EAG--TCWC--CTYP/P, 1996.

Tufiş, D. (1999). Tiered Tagging and Combined Classifiers. In F. Jelinek, E. Nöth (eds) Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence 1692, Springer, 1999, pp. 28-33

Tufiş, Dan and Irimia, Elena. (2006). RoCo_News - A Hand Validated Journalistic Corpus of Romanian. In Proceedings of the 5th LREC Conference. Genoa, Italy, May 2006, pp. 869-872.

Tufiş, D., Ion, R., Ceauşu, A., Ştefănescu D. (2008). RACAI's Linguistic Web Services. In Nicoletta Calzolari et al. (Eds.) Proceedings of the 6th LREC, Marrakech, Morocco, May 2008, European Language Resources Association (ELRA).