of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3). Mannheim: Institut für Deutsche Sprache, pp. 5-10

CoRoLa Starts Blooming – An update on the Reference Corpus of Contemporary Romanian Language

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

Horia Nicolai Teodorescu, Dan Cristea, Andrei Scutelnicu, Cecilia Bolea, Alex Moruz, Laura Pistol
Institute for Computer Science, Iași
2 T. Codrescu St, 700481, Iași, Romania
hteodor@etti.tuiasi.ro, dcristea@info.uaic.ro, andreiscutelnicu@gmail.com, ceciliabolea@iit.academiaromana-is.ro, mmoruz@info.uaic.ro
laura.pistol@iit.academiaromana-is.ro

Abstract

This article reports on the on-going CoRoLa project, aiming at creating a reference corpus of contemporary Romanian (from 1945 onwards), opened for online free exploitation by researchers in linguistics and language processing, teachers of Romanian, students. We invest serious efforts in persuading large publishing houses and other owners of IPR on relevant language data to join us and contribute the project with selections of their text and speech repositories. The CoRoLa project is coordinated by two Computer Science institutes of the Romanian Academy, but enjoys cooperation of and consulting from professional linguists from other institutes of the Romanian Academy. We foresee a written component of the corpus of more than 500 million word forms, and a speech component of about 300 hours of recordings. The entire collection of texts (covering all functional styles of the language) will be pre-processed and annotated at several levels, and also documented with standardized metadata. The pre-processing includes cleaning the data and harmonising the diacritics, sentence splitting and tokenization. Annotation will include morpho-lexical tagging and lemmatization in the first stage, followed by syntactic, semantic and discourse annotation in a later stage.

1 Introduction

In 2012 the Romanian Academy Research Institute for Artificial Intelligence "Mihai Drăgănescu" from Bucharest (RACAI) finalized the Romanian Balanced Corpus (ROMBAC[1]) (Ion et. al, 2012) containing 44,117,360 tokens covering four domains (News, Medical, Legal, Biographic and Fiction). The nucleus of ROMBAC was represented by the RoCo_News corpus (Tufiș and Irimia, 2006), a hand validated corpus of almost 7 million tokens from the weekly magazine Agenda (2003-2006).

Since 2014 the concern for creating a bigger corpus 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 time span covered by the project is 1945-present, with two subperiods (1945-1990, 1990-present), with clear differences, mainly at the lexical level. From this perspective, a big challenge for us is the collection of electronic texts to cover the whole period. For the last couple of decades there is an important amount of such texts available. However, in the case of the texts from previous decades considerable effort needs to be done for finding the owners of the texts IPR, for scanning, OCRizing and correcting the

[1] http://www.meta-net.eu/meta-share

texts. This could imply raising the awareness of main libraries about the cultural responsibility of digitizing even contemporary books, not only the old ones.

2 Objectives

When finished, CoRoLa will be a medium to large corpus (more than 500 million word forms), IPR cleared, in which all functional styles will be represented: 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 imaginative writing. The provisional structure of the corpus is described in some details in Barbu Mititelu and Irimia (2014). Unlike its predecessor, CoRoLa will include a syntactically annotated sub-corpus (treebank) and an oral component. All textual data will be morpho-lexically processed (tokenized, POS-tagged and lemmatized). The treebank (we target 10,000 hand validated sentences) and the oral component (targeted: 300 hours of transcribed recorded speech) have additional annotations (dependency links, respectively speech segmentation at sentence level, pauses, non-lexical sounds and partial explicit marking of the accent).

Particular attention is paid to data documentation, i.e. associating it with standardized metadata. We adopted the CMDI (Component MetaData Infrastructure)[2] approach for the creation of our metadata.

3 Data Collection and Cleaning

The resource we are building will have two important attributes: it will be representative for the language stage, thus covering all language registers and styles; it will be IPR cleared, which is a challenging task, triggered by the need to observe the intellectual property law. The categories of content excepted by this law are: political, legislative, administrative and judicial. Therefore, without the written accept from IPR owners, from the other kinds of texts only tiny fragments of no more than 10,000 characters can be used. We must also consider only texts written with correct diacritics (otherwise, the linguistic annotation will be highly incorrect).

To ensure the volume and quality of the texts in the corpus, as well as copyright agreements on these texts, our endeavour was to establish collaborations with publishing houses and editorial offices. So far (March 2015), we have signed agreements with the following publishing houses: Humanitas, Polirom, Romanian Academy Publishing House, Bucharest University Press, "Editura Economică", ADENIUM Publishing House, DOXOLOGIA Publishing House, the European Institute Publishing House, GAMA Publishing House, PIM Publishing House. Some magazines and newspapers have also agreed to help our project by providing access to their articles: România literară, Muzica, Actualitatea muzicală, Destine literare, DCNEWS, PRESSONLINE.RO, the school magazine of Unirea National College from Focșani, SC INFOIASI SRL, Candela de Montreal. Until now four bloggers have also agreed to allow us to include some of their posts in the corpus: Simona Tache[3], Dragoș Bucurenci[4], Irina Șubredu[5] and Teodora Forăscu[6]. Also, we have signed agreements with the writers Corneliu Leu and Liviu Petcu.Oral texts (read news, live transmissions and live interviews) (one hour per working day) are provided by Rador (the press agency of Radio Romania) and Radio Iaşi – a local broadcasting agency. All data providers readiness to get involved was a very pleasant surprise for us and we express here, again, our gratitude.

Another challenge in corpus creation is to have texts in a clean format, easy to process and annotate. Once our collaborators dispatch a textual resource (usually in unprotected pdf files, rarely in doc files), the first step is to convert it into an adequate format for our pre-processing tools[7].

Given the large amount of texts, we automated a part of the process (Moruz and Scutelnicu, 2014): the text is automatically retrieved from the pdf files, paragraph limits are recuperated, column marking newlines are erased as well as hyphens at the end of the lines. However, a lot of manual work remains to be done: separating articles from periodicals in different files, removal of headers, footers, page numbers, figures, tables, dealing with foot- or end-notes, with text fragments in foreign languages, with excerpts from other authors, etc. When copied from their original sources, the content is converted into the UTF-8 encoding and saved as plain text documents.

[2] http://www.clarin.eu/content/component-metadata

[3] http://www.simonatache.ro

[4] http://bucurenci.ro

[5] http://irina.subredu.name

[6] https://travelearner.wordpress.com

[7] http://www.racai.ro/en/tools/

CoRoLa is 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, IMS Open Corpus Workbench (CWB, http://cwb.sourceforge.net/), an open source medium that allows complex searching with multiple criteria and support for regular expressions. It allows 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. The platform has already been installed and tested on the ROMBAC corpus and coupled with our processing chain which produces the adequate annotated format for morphological and shallow syntactic searches. For the near future, we plan to switch to the more powerful corpus management platform KorAP (Bański et al., 2014).

The TTL (Ion, 2007) processing chain ensures, at the time of this writing, the following specific functionalities: sentence splitting, tokenisation, tiered-tagging (Tufiş, 1999), 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.

4 Current Statistics

4.1 Textual Data

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: tokens (word forms and punctuation).

A finer classification of the documents, according to their sub-domains, outlines the following categories: literature, politics, gossip columns, film, music, economy, health, linguistics, theatre, painting/drawing, law, sport, education, history, religious studies and theology, medicine, technology, chemistry, entertainment, environment, architecture, engineering, pharmacology, art history, administration, oenology, pedagogy, philology, juridical sciences, biology, social, mathematics, social events, philosophy, other.

In parallel with the CoRoLa corpus, at ICIA and UAIC a Romanian treebank is under development (Irimia and Barbu Mititelu, 2015), (Perez, 2014), (Mărănduc and Perez, 2015). Currently each of the two sections of the treebank contains almost 5,000 sentences, which are in the process of being mapped into the UD project specifications[8]. The final version of the CoRoLa corpus will include the Romanian treebank as well.

DOMAIN		STYLE	
arts&culture	32,838,881	journalistic	44,248,356
society	33,582,123	science	26,990,172
others	9,990,383	imaginative	11,945,283
science	19,923,533	others	1,777,475
nature	106,196	memoirs	1,511,676
		administra-tive	865,660
		law	9,102,494
TOTAL[9]	96,441,116	TOTAL	96,441,116

Table 1. Domain and style distribution of textual data.

4.2 Speech data

Speech data collected so far is accompanied by transcriptions (observing the current orthography). Partially (about 10%), it was automatically pre-processed and the transcriptions were XML encoded with mark-up for lemma, part-of-speech and syllabification. Additionally to the XML annotations we provide 3 files which contain the original sentences (".txt" extension) the stripped version (which is obtained by removing all punctuation from the original sentences – useful in training systems such as Sphinx or HTK (Hidden Markov Model Toolkit) – ".lab" extension) and time aligned phonemes (tab separated values which contain each phoneme in the text with its associated start and stop frame – ".phs" extension).

[8] https://code.google.com/p/uni-dep-tb/

[9] Currently more textual data, not included into CoRoLa, has been collected, which may be used for improving models of our statistical processing tools. Among them are Wiki-Ro, the Romanian part of a big collection of sentences extracted from Wikipedia within the ACCURAT European project (http://www.accurat-project.eu/) and the Romanian part of the Acquis-Communautaire (Steinberger et al. 2006). They are already pre-processed and contain more than 50 million words. Similarly, we acquired some audio-books (not IPR clarified and thus, not included into CoRoLa) used only for evaluation of our tools.

- **RASC** (Romanian Anonymous Speech Corpus) is a crowd-sourcing initiative to record a sample of sentences randomly extracted from Ro-Wikipedia (Tufiş et al., 2014). The corpus is automatically aligned at phoneme/word level.
- **RSS-ToBI** (Romanian Speech Synthesis Corpus) is a collection of high quality recordings compiled by (Stan et al., 2011) and designed for speech synthesis. It was enhanced with a prosodic ToBI-like (Tone and Break Indices) annotation (reference to be added). It is automatically aligned at phoneme/word level.
- **RADOR (**Radio Romania) and **Radio Iaşi** is a collection of radio news and interviews, provided daily by the Romanian Society for Broadcasting and the main Iaşi radio channel. At the time of this writing, the transcriptions are under preprocessing. They are not yet aligned at phoneme/word level.

Corpus	Type	Source	Time length (h:m:s)
RASC	many speakers	RoWikipedia	04:22:02
RSS-ToBI	single speaker	news&fairy tales	03:44:00
RADOR	many speakers	news& interviews	106:52:33
Radio Iaşi	many speakers	interviews	07:00:00 under development
			>121:58:35

Table 2. Speech corpora.

Besides these speech corpora, we contracted professional recordings (about 10 hours) of sentences selected by us from Romanian Wikipedia. These recordings will enlarge the RASC corpus.

Further information on the already processed speech data are given in the table below.

Corpus	sentences	words	phonemes
RASC	2,866	39,489	270,591
RSS-ToBI	3,500	39,041	235,150
	6,266	78,530	505,741

Table 3. Currently pre-processed speech corpora

A special mention deserves the site "Sounds of the Romanian Language" (Feraru et al., 2010), which is a systematically built, explanatory small collection of annotated and documented recordings of phonemes, words, and sentences in Romanian, pronounced repeatedly by several speakers; the corpus also includes as annex materials numerous papers on the topic and several instruments for speech analysis. Sections of the corpus are devoted to emotional speech, to specific processes as the double subject, and to phonetic pathologies. The corpus is maintained by the Institute for Computer Science of the Romanian Academy[10].

5 Metadata Creation

The challenge in CoRoLa is to create a corpus from which more than only concordances to be extracted, i.e. giving the user the possibility to construct his/her own subcorpus to work with, depending on the domain/style/period/author/etc. The only way to obtain this is to document each file with metadata. For documents sent by publishing houses, etc., we created the metadata files manually. For text files crawled from the web (articles, blogs), we automatically created metadata, with a preliminary phase of mapping the existent classifications of texts on those sites onto our classification of texts.

6 Annotation of the data

As mentioned before, a processing chain[11] has been established, consistent with the tabular encoding specific to the CWB platform and comprising more program modules that execute particular functions. The web-service chain provides:

- sentence splitting: it uses regular expressions for the identification of a sentence end;
- tokenization: the words are separated from the adjacent punctuation marks, the compound words are recognized as a single lexical atom and the cliticized words are split as distinct lexical entities;
- POS tiered-tagging with the large MULTEXT-East tag set; its accuracy is above 98%;
- lemmatization: based on the tagged form of the word, it recovers its corresponding lemma from a large (over 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

[10] http://www.etc.tuiasi.ro/sibm/romanian_spoken_language/en/

[11] http://ws.racai.ro/ttlws.wsdl

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%. A better lemma-guessing (about 93%) is ensured by a new neural network based-tagger (Boroș et al., 2013), not yet integrated in the processing chain for CWB.

- chunking: for each lexical unit previously 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: syntactic parsing, semantic annotation and discourse analysis.

The annotation of the speech data includes, additionally, the syllabation and accent mark-up plus the grapheme to phoneme alignment.

7 Annotation correction

In our previous experiments (Tufiș and Irimia, 2006) 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 errors identified in this manner can also be automatically corrected. This validation procedure was used in the past to correct tagging and lemmatization errors for the journalistic corpus RoCo_News and for ROMBAC 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 two 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 mentioned methodology to improve the quality of the entire corpus, we intend to manually validate a limited part 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.

8 Conclusions

In the international context of growing interest for creating large language resources, we presented here the current phase in the creation of a reference corpus of contemporary Romanian. It is a joined effort of two academic institutes, greatly helped by publishing houses and editorial offices, which kindly accepted the inclusion of their texts at no costs. The corpus will be available for search for all those interested in the study or processing of the Romanian language.

We emphasize the idea that, although large amount of texts are out there on the web, creating an IPR clear reference corpus is quite a challenge, not only due to vast efforts invested in persuading IPR holders to contribute to a cultural action, but also to achieve agreements on what texts and how much of them to include in the corpus. In spite of the decided CoRoLa structure (text types and quantities) of the linguistic data the supplementary data we manage to collect (mainly from the web) is not discarded, but stored for training specialized statistical models to be used in different data-driven applications (CLIR, Q&A, SMT, ASR, TTS).

Acknowledgements

We express here our gratitude to all CoRoLa volunteers, undergraduate, graduate and Ph.D. students, as well as researchers and university staff in computer science and linguistics, who, noble-minded and aware of the tremendous importance that such a corpus will have for the Romanian culture, have generously agreed to help in the process of filling in metadata and cleaning the collection of texts.

References

P. Bański, N. Diewald, M. Hanl, M. Kupietz, A. Witt. 2014. Access Control by Query Rewriting. The Case of KorAP. *Proceedings of the Ninth Conference on International Language Resources and Evaluation (LREC'14)*: 3817-3822.

V. Barbu Mititelu, E. Irimia. 2014. The Provisional Structure of the Reference Corpus of the Contemporary Romanian Language (CoRoLa). In M.Colhon, A. Iftene, V. Barbu Mititelu, D. Tufiș (eds.) *Proceedings of the 10th Intl. Conference "Linguistic Resources and Tools for Processing Romanian Language"*: 57-66.

T. Boroș, R. Ion, D. Tufiș. 2013. Large tagset labeling using Feed Forward Neural Networks. Case study on Romanian Language. *Proceedings of ACL 2013*: 692-700.

T. Boroș, A. Stan, O. Watts, S.D. Dumitrescu. 2014. RSS-TOBI - A Prosodically Enhanced Romanian Speech Corpus. *Proceedings of 9th LREC 2014*: 316-320.

D. Cristea, I.C. Pistol. 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: 228-246.

S.D. Dumitrescu, T. Boroș, R. Ion. 2014. Crowd-Sourced, Automatic Speech-Corpora Collection–Building the Romanian Anonymous Speech Corpus. *CCURL 2014: Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era*: 90-94.

S.M. Feraru, H.N. Teodorescu, M.D. Zbancioc. 2010. SRoL - Web-based Resources for Languages and Language Technology e-Learning. *International Journal of Computers Communications & Control*, Vol. 5, Issue 3: 301-313.

R. Ion. 2007. *Word Sense Disambiguation Methods Applied to English and Romanian*, PhD thesis, Romanian Academy (in Romanian).

R. Ion, E. Irimia, D. Ştefănescu, D. Tufiş. 2012. ROMBAC: The Romanian Balanced Annotated Corpus. In Nicoletta Calzolari et al. (Eds.) *Proceedings of the 8th LREC*: 339-344.

E. Irimia, V. Barbu Mititelu. 2015. Building a Romanian Dependency Treebank, Proceedings of *Corpus Linguistics 2015*.

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

C. Mărănduc, A.C. Perez. 2015. A Romanian dependency treebank. *Proceedings of CICLing 2015*.

A. Moruz, A. Scutelnicu. 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"*: 163-170.

A. Perez. 2014. Resurse lingvistice pentru prelucrarea limbajului natural (Lingvistic Resources For Natural Language Processing). Ph.D. thesis, „Alexandru Ioan Cuza" University of Iași.

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

A. Stan, J. Yamagishi, S. King, M. Aylett. 2011. The Romanian Speech Synthesis (RSS) corpus: building a high quality HMM-based speech synthesis system using a high sampling rate. *Speech Communication, 53*(3): 442-450.

R. Steinberger, B. Pouliquen, A. Widiger, C.Ignat, T. Erjavec, D. Tufiș, D. Varga. The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In Proceedings of the 5th LREC Conference, Genoa, Italy, 22-28 May, 2006, pp.2142-2147, ISBN 2-9517408-2-4.

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

D. Tufiș, E. Irimia. 2006. RoCo_News - A Hand Validated Journalistic Corpus of Romanian. In *Proceedings of the 5th LREC*: 869-872.

D. Tufiș, R. Ion, A. Ceauşu, D. Ştefănescu. 2008. RACAI's Linguistic Web Services. In Nicoletta Calzolari et al. (Eds.) *Proceedings of the 6th LREC*: 327-333.

D. Tufiș, R. Ion, Ș. D. Dumitrescu, D.Ștefănescu. 2014. Large SMT data-sets extracted from Wikipedia. In *Language Resources and Evaluation Conference (LREC 14)*.Reykjavik, Iceland, May 2014