1. Introduction

Within the international context of creating national corpora (American (Ide and Suderman 2004), British (http://www.natcorp.ox.ac.uk/), Bulgarian (Koeva, Stoyanova, Leseva, Dimitrova, Dekova and Tarpomanova 2012), Croatian (Tadić 2002), Czech (Křen, Cvrček, Čapka, Čermáková, Hnátková, Chlumská, Jelínek, Kovářiková, Petkovič, Procházka, Skoumalová, Škrabal, Truneček, Vondříčka and Zasina 2016), Hungarian (Oravecz, Váradi and Sass 2014), Polish (Przepiórkowski, Bańko, Górski, Lewandowska-Tomaszczyk, Łaziński, Pęzik 2011), etc.) and even multilingual corpora (Schmidt and Wörner 2012), we present the national collaboration for building a large Romanian corpus, reflecting both the written and the spoken aspects of the language, as well as various styles, domains and subdomains. It is meant as a representative corpus of contemporary Romanian language, hence called CoRoLa. The results of this project, run by the “Mihai Drăgânescu” Research Institute for Artificial Intelligence in Bucharest and the Institute for Computer Science in Iași, will be a corpus in electronic format, available (online) for free, in order to be used for studies on contemporary language, for processing language, for creating applications that use knowledge extracted from large corpora, for improving translation and for teaching Romanian (Barbu Mititelu and Dumitrescu 2014).
The paper is structured as follows: section 2 presents and discusses some key characteristics of CoRoLa; section 3 briefly describes the process of text collection and cleaning and the metadata associated to each text file; section 4 enumerates and presents the annotation levels of the corpus; section 5 presents DRuKoLA – a project closely related to CoRoLa; section 6 sketches the ways the data in CoRoLa will be accessed through the KorAP interface (as part of the DRuKoLA project); the conclusions section closes the paper.

2. CoRoLa’s characteristics

In this section we focus on several key characteristics of CoRoLa, each of them discussed in a separate subsection.

2.1. Size

The targeted size of the corpus is 500,000,000 words. Although several years ago, when we started this endeavour, this amount could have been perceived as a large corpus, at the moment it is no longer the case: there are already available corpora containing billions of words: for example, the German corpus (Kupietz and Lüngen 2014) counts more than 31 billion tokens (http://www1.ids-mannheim.de/kl/projekte/korpora), the Bulgarian one – 1.2 billion words (http://dcl.bas.bg/bulnc/en/), etc.). Nevertheless, for the Romanian language, this will be the largest corpus ever built. The quality, the type and the origin of the texts it contains (as described below) make CoRoLa even more valuable.

At the moment of this writing, CoRoLa contains 880,975,551 written tokens and about 200 hours of recordings. Their distribution according to various criteria is given in section 2.3 below. Out of the total number of written tokens, only 322,290,278 have already been preprocessed and morpho-syntactically annotated, but the whole of them will have been preprocessed and annotated before the end of the project (December, 2017). Out of the total number of recordings, 57 hours have already been processed and annotated, syllabified and some of the allophones have been characterised, but further work will also increase the representation of oral texts in the final CoRoLa.

2.2. Time span and format of textual sources

Focusing on the contemporary language, CoRoLa includes quite recent texts, original writings and a few translations authored in the period 1945 – today, that have been contributed by their legal owners in electronic forms.

2.3. Texts types coverage

Aiming at being a reference corpus, CoRoLa includes texts from all language styles: scientific, administrative, legal, imaginative, publicistic. The colloquial style is
not targeted, but is reflected especially in the imaginative texts. We added a separate style, called memoirs, which includes journals, letters, etc. Table 1 shows the distribution of the CoRoLa texts across these styles at the moment of this writing. The best represented style is the legal one, followed by the scientific one. Further effort for collecting imaginative, publicistic and administrative texts is necessary. Note that some texts (mainly originating in blog posts) are not assigned any of these styles, and thus included in the category “unclassified”.

<table>
<thead>
<tr>
<th>Style</th>
<th>Number of tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>legal</td>
<td>527,519,345</td>
</tr>
<tr>
<td>scientific</td>
<td>184,761,720</td>
</tr>
<tr>
<td>publicistic</td>
<td>77,277,228</td>
</tr>
<tr>
<td>imaginative</td>
<td>51,617,302</td>
</tr>
<tr>
<td>memoirs</td>
<td>26,135,623</td>
</tr>
<tr>
<td>administrative</td>
<td>11,564,015</td>
</tr>
<tr>
<td>unclassified</td>
<td>2,100,318</td>
</tr>
<tr>
<td>TOTAL</td>
<td><strong>880,975,551</strong></td>
</tr>
</tbody>
</table>

Table 1. The distribution of CoRoLa written texts according to their style.

Another criterion for classifying texts in a corpus is the domain to which they belong. For CoRoLa we have defined four domains: arts and culture, nature, society and science. These domains are assigned to texts belonging to the scientific and publicistic styles. The texts from the styles legal, memoirs, administrative, imaginative and “unclassified” are assigned to the domain “unclassified”. Table 2 shows the distribution of the CoRoLa texts across these domains.

<table>
<thead>
<tr>
<th>Domain</th>
<th>Number of tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>arts and culture</td>
<td>27,697,861</td>
</tr>
<tr>
<td>nature</td>
<td>1,831,275</td>
</tr>
<tr>
<td>society</td>
<td>119,150,171</td>
</tr>
<tr>
<td>science</td>
<td>160,309,410</td>
</tr>
<tr>
<td>unclassified</td>
<td>571,986,834</td>
</tr>
<tr>
<td>TOTAL</td>
<td><strong>880,975,551</strong></td>
</tr>
</tbody>
</table>

Table 2. The distribution of CoRoLa written texts according to their domain.

2.4. IPR-cleared texts

In Romania, Law 8/1996 concerning the author’s rights has scope, as far as our project is concerned, over literary, publicistic, scientific works, be they written or oral. As a consequence, such texts can be included in the corpus provided that we have the agreement of the IPR owner(s). To achieve the coverage mentioned above, we have contacted authors, publishing houses and media representatives and got commonly agreed upon protocols allowing us to store such texts, process and annotate them and offer users the possibility to query them.
The texts outside the scope of this law, considered of public and open use, namely the administrative and legal texts, as well as those belonging to the Romanian Wikipedia, have simply been downloaded from the source, stored, processed and annotated, and they can also be queried by users.

The IPR clearness of CoRoLa is one great asset of this corpus, made possible by the open-mindedness of IPR owners, by their generosity and willingness to take part into this national cultural act.

2.5. Access to CoRoLa

All texts are stored on the developers’ servers, in Bucharest and Iași, and a query interface (see section 6 below) will offer the users the possibility to search for words, word combinations, even by specifying morphologic or syntactic constraints, in the corpus and visualise the results.

3. Texts – collection, cleaning and associated metadata

For being processed and further annotated, written texts must be in txt format. However, when offered to us or harvested automatically, they were in various formats: pdf, doc, html. If automatic extraction of the txt equivalent from the html and doc files does not raise serious issues, in the case of the pdf files things are far from trivial. Some commonly occurring problems are: diacritics are lost or different letters with diacritics (usually ş and ţ) are recognized as the same character (very frequently a symbol instead of a letter), lines from parallel columns are intermingled with each other, footers, headers and page numbers are inserted in text without any delineation, etc. Consequently, a partially manual, partially automatic time- and effort-consuming cleaning phase was necessary in their case.

Oral texts are of several types (Tufiș, Barbu Mititelu, Irimia, Dumitrescu, Boroș 2016): news, radio interviews, theatre, randomly selected Romanian Wikipedia sentences recorded by anonymous volunteers, texts (selected so as to cover most of the interesting phonetic structures of Romanian) read by professionals and recorded in acoustic rooms, and even fragments of audiobooks. All these recordings are accompanied by their corresponding written texts.

All texts (written and oral) have a metadata file associated, which contains information that can be used by the user (either alone or in combination) in her/his corpus research (see section 6). The title of the text, its author(s), its publication date and medium, the style of the text, the domain and subdomain when applicable, the annotation levels are elements of the metadata files.
4. Texts processing and annotation levels in CoRoLa

All written texts are automatically split into sentences and then tokenized: inside each sentence, words and punctuation are separated as individual tokens. As such, they enter a processing phase (in which diacritics are automatically restored, syllabification hyphens are eliminated, words written as one token are separated into different tokens) and then the annotation phase (which consists in a morphologic and syntactic analysis of the words). We present below the analysis of the sentence “Era o carte fără titlu.” (Was a book without title. ‘It was a book without a title.’), processed and annotated at all levels available at the moment for the corpus. It has 6 tokens: 5 words and one punctuation mark, each numbered according to its order in the sentence (the first column in the example below). Each word is lemmatised (the third column), the part of speech is indicated, in an encoding specific for the Romanian language inherited from the MULTEXT EAST project (http://nl.ijs.si/ME/V4/msd/html/msd-ro.html) (the fourth column), and a full morphologic analysis is provided (the fifth column). For example, the first word, *Era* (“Was”) is analysed as a verb (V), the main one (m), in the indicative mood (i), the imperfect tense (i), the third person (3), in singular (s); the word *o* (“a”) is analysed as an article (T), indefinite (i), in the feminine case (f), singular number (s) and a direct case (r). The fourth column contains a shorter tag corresponding to the one in the fifth column. The syntactic annotation follows the dependency framework, more exactly the Universal Dependency (UD) version 1.4 (Nivre, Agić, Ahrenberg et al. 2016). For each token, it offers information about its head (see the seventh column, where numbers refer those in the first column) and the relation it establishes with that head (the eighth column). The root of this sentence (according to the UD principles) is the word *carte* (“book”), thus it does not have a head in the sentence (notice the number 0 in its seventh column). This root is the head of the words *Era* (“Was”), *o* (“a”), *titlu* (“title”), as well as of the final punctuation of the sentence. In its turn, *titlu* is the head of the preposition preceding it, namely *fără* (“without”).

<p>| | | | | | | | | |</p>
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Era</td>
<td>fi</td>
<td>V3</td>
<td>Vmii3s</td>
<td>_</td>
<td>3</td>
<td>cop</td>
<td>_</td>
</tr>
<tr>
<td>2</td>
<td>o</td>
<td>un</td>
<td>TSR</td>
<td>Ti$f$r</td>
<td>_</td>
<td>3</td>
<td>det</td>
<td>_</td>
</tr>
<tr>
<td>3</td>
<td>carte</td>
<td>carte</td>
<td>NSRN</td>
<td>Ncfsrn</td>
<td>_</td>
<td>0</td>
<td>root</td>
<td>_</td>
</tr>
<tr>
<td>4</td>
<td>fără</td>
<td>fără</td>
<td>S</td>
<td>Spsa</td>
<td>_</td>
<td>5</td>
<td>case</td>
<td>_</td>
</tr>
<tr>
<td>5</td>
<td>titlu</td>
<td>titlu</td>
<td>NSN</td>
<td>Nems-n</td>
<td>_</td>
<td>3</td>
<td>nmod</td>
<td>_</td>
</tr>
<tr>
<td>6</td>
<td>.</td>
<td>.</td>
<td>PERIOD</td>
<td>PERIOD</td>
<td>_</td>
<td>3</td>
<td>punct</td>
<td>_</td>
</tr>
</tbody>
</table>

The tokenisation, morphologic tagging and lemmatization of the texts in CoRoLa are made with the TTL tool developed in ICIA (Ion 2007). The syntactic parsing of
the whole corpus is going to be done with a parser trained on a Romanian treebank, which is now in a training phase (Barbu Mititelu, Ion, Simionescu, Scutelnicu and Irimia 2017).

As said, oral texts in CoRoLa are accompanied by their written counterpart. These are either transcripts (in the case of interviews) or original texts, as in the case of theatre plays, registered during hearings or actual performances with actors, or simply read and registered by members of the project or students, in professional studios or in informal environment, thus obtaining the oral form. The existence of both oral and the written components allows for their automatic alignment (at the sentence and word level, more accurate alignments going down to the letter-phoneme level). The textual counterparts of the spoken files follow the same tokenisation/lemmatisation and annotation process at the morphologic level.

5. DRuKoLA

The development of CoRoLa is supported and completed by a parallel running project, DRuKoLA1 (Cosma, Cristea, Kupietz, Tuţuş and Witt 2016), run (between 2016 and 2018) by the Institut für Deutsche Sprache (Mannheim), “Mihai Drăganescu” Research Institute for Artificial Intelligence (Bucharest), the Institute for Computer Science (Iaşi) and the University of Bucharest (the Departments of Germanic Languages and of English). This project aims at building comparable virtual corpora under one analysis tool, the new KorAP, a query and analysis tool of the Institut für Deutsche Sprache. It is a powerful analysis platform able to easily manage big data and gradually targeting towards linking existing resources (Diewald, Hanl, Margaretha, Bingel, Kupietz, Bański and Witt, 2016).

Core elements of this project are the two reference corpora of the Romanian and of the German language (CoRoLa and DeReKo, Kupietz and Lüngen 2014), from which comparable subcorpora would be dynamically extracted, based on text characteristics or annotation features considered to be essential for the individual research topics. DRuKoLA itself is part of a larger project aiming at developing a corpus technology able to share corpora, technical processes and features, as well as

---

1 DRuKoLA is an acronym for Sprachvergleich korputstechnologisch Deutsch-Rumänisch, a Research Group Linkage Programme of the Alexander von Humboldt-Foundation as an alumni programme meant to link a host research institute from Germany with the home institute of the Humboldttian alumna/alumnus (in this case, the University of Bucharest).
research results in a common analysis platform for a European Reference Corpus (EuReCo) (Kupietz, Witt, Bański, Tufiș and Cristea 2017).

The fact that CoRoLa was still at its beginning when planning the DruKoLa project made it much easier to concord on metadata categories and to make sure that the combined collection becomes interoperable. KorAP offers the opportunity to either work within each language on a virtual corpus separately, integrating features of each language, as developed by the respective curator institute, or to work cross-linguistically within a comparable virtual corpus built on individual needs. Adjustments will be made on the basis of an inventory of specific functionalities for contrastive research. A further phase would require testing and exploration of language specific features identified with respect to different parameters and structures.

6. Accessing the textual data

Work is presently going on to make the textual content of CoRoLa compatible with the KorAP query and analysis tool, thus accomplishing part of the DRuKoLa objectives. A pipeline is under development that will take the 3 original file formats (raw texts, metadata describing them and the stand-off annotation that complements the texts with morpho-syntactic information) and will transfer them to fit the KorAP standards. This process presupposes also a physical transfer of the corpus content from the server hosting the CoRoLa corpus to the ones allocated to the DRuKoLa project (located in the developers’ premises).

KorAP is meant to be scalable, flexible and sustainable, its components being easily extended, replaceable and maintained. Moreover, KorAP should support multiple query languages (i.e. conventions of expressing queries by users). New in this enterprise is the use of KorAP to access corpora in different languages, German and Romanian at this moment, but others in the near future (Kupietz, Witt, Bański, Tufiș and Cristea 2017).

The KorAP query protocol allows the user to formulate two main types of queries: addressing the collection of documents that satisfy certain conditions (in which metadata are searched for) and addressing the texts themselves, with the intention to get occurrences of particular linguistics structures (in which the primary textual data and their annotations on different linguistic layers are used).
7. Conclusions

CoRoLa is the reference corpus for Romanian, for which the texts (both written and oral) have already been harvested, in a larger proportion than promised in the project proposal. This language resource will be open for querying to all those interested starting December 2017. Moreover, due to the international project DruKoLA, it can be used not only for monolingual studies, but also for contrastive ones. Further on, it could be enriched with even more diverse texts and it could even acquire a multilingual component, by adding parallel or comparable texts in other languages (one of which will be Romanian).

Building a corpus may seem a rather easy task, since it may not involve more than putting together a large collection of texts, many of them anyway available on the web. However, this paper tried to show the complexity of this enterprise.

Acknowledgements

We thank all our texts providers: publishing houses (Humanitas, Polirom, Economica, PIM, Editura Academiei Române, Editura Universității din București, Adenium, Doxologia, Gama, Editura Institutul European Iași, Simetria, Editura Casa Cărtii de Știință), mass media representatives (DCNews, Societatea Română de Radiodifuziune, Info Iași, Timpul, România literară, Actualitatea muzicală, Candela de Montreal, republica.ro, Destine literare, Financiarul, Info Iași, Muzica, Balcanii și Europa, Timpul, Radio Iași, Radio Viva, RadioU, RomanTV, Teatrul Național Târgu Mureș, Teatrul Național Cluj-Napoca), the National College Unirea from Focșani, individual authors and contributors (Luminița Cărăuşu, Corneliu Leu, Liviu Petcu, Daniela Coza, Eugen Rațiu) and several bloggers.

The project benefited from the work of students from the University of Bucharest, “Al. I. Cuza” University of Iași, and University POLITEHNICA of Bucharest.

The syntactic annotation of a treebank, which was included in CoRoLa, and the development of the parser based on these texts and used for further annotation of the texts in CoRoLa were two of the objectives of grant of the Romanian National Authority for Scientific Research and Innovation, CNCS-UEFISCDI, project number PN-II-RU-TE-2014-4-1362.

References

Barbu Mițitelu, Verginica, Ștefan Daniel Dumitrescu, 2014, „CoRoLa – the Representative Corpus of the Contemporary Romanian Language. The initial phase”, in


Nivre, Joakim; Agić, Željko; Ahrenberg, Lars; et al., 2016, Universal Dependencies 1.4, LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics, Charles University, http://hdl.handle.net/11234/1-1827.


Tufiș, Dan, Verginica Barbu Mititelu, Elena Irimia, Ştefan Daniel Dumitrescu, Tiberiu Boroş, 2016, „The IPR-cleared Corpus of Contemporary Written and Spoken Romanian Language”, in Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Marko Grobelnik, Bente Maegaard, Joseph Mariani, Asuncion Moreno, Jan Odijk, Stelios Piperidis (eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), 23-28 May, Portorož, Slovenia, p. 2516-2521.

CORPUS OF CONTEMPORARY ROMANIAN. ARCHITECTURE, ANNOTATION LEVELS AND ANALYSIS TOOLS

(Abstract)

The paper presents, on the one hand, the characteristics, the structure, the annotation levels of the reference corpus for the Romanian language (CoRoLa). On the other hand, it presents the accompanying project of harmonizing an analysis tool of the German language to Romanian (project DDrUKoLA), in order to be able to perform contrastive analysis in different languages by means of comparable virtual corpora under one single tool (KorAP), developed by the Institut für Deutsche Sprache (Mannheim). This would represent a first step towards a larger endeavour to link reference corpora of European languages under a common analysis platform in a European Reference Corpus (EuReCo).