Memoirs of the Scientific Sections of the Romanian Academy
Tome XL, 2017

COMPUTER SCIENCE

AN INSIGHT INTO THE CORPUS OF CONTEMPORARY ROMANIAN

DAN CRISTEA[1,2], DANIELA GÎFU[1,2], MIHAI-ALEX MORUZ[1,2], MIHAELA ONOFREI[1], LAURA PISTOL[1], LIVIU-ANDREI SCUTELNICU[1,2] and SPERANȚA-CECILIA BOLEA[1]

[1]Institute of Computer Science, Romanian Academy, Iaşi Branch, Iaşi, Romania
[2]"Alexandru Ioan Cuza" University of Iaşi, Faculty of Computer Science, Iaşi, Romania
Corresponding authors: danu.cristea@gmail.com, cecilia.bolea@iitacademiaromana-is.ro

This paper presents the almost final results of a priority project of the Romanian Academy – the Corpus of **Co**ntemporary **Ro**manian **La**nguage (CoRoLa). The Corpus includes data in both written and spoken forms of the language. The textual collection is made up of publications covering the period from the 2nd World War to our days, while the spoken collection includes only recent recordings.

Keywords: Romanian language, data collections, text and speech, metadata, interfaces for acquiring linguistic data and access.

INTRODUCTION

PLANNING COROLA

Answering to a necessity, which has been noticed already for some time in modern Romanian linguistics, in 2014, the Presidium of the Romanian Academy commissioned to two of its institutes that are known for their research in the domains of computational linguistics and natural language processing – the "Mihai Drăgănescu" Research Institute for Artificial Intelligence in Bucharest (ICIA) and the Institute for Computer Science in Iaşi (IIT) – the elaboration of a large Romanian corpus of texts and voice recording. In the vision of the Romanian Academy, this corpus would have had to display samples of written and spoken contemporary Romanian language, therefore not older than the end of the Second World War, be large enough to include occurrences of the large majority of Romanian words used in the contemporary language, be quasi-exhaustive in the use of the senses of words and be openly available on the internet. These requirements configure the primary characteristics of the Corpus: in digital form, online and free, contemporary language and representative. The representativeness is parametrized through the dimension of the corpus (planned to include 500 million textual words) and the distribution of its texts over a quasi-exhaustive range of domains and literary styles. Other important features used in planning the corpus have been: all linguistic items included (pieces of texts and voice recordings) have to be accompanied by the written accept of the legal owners, they have to be "documented", in the sense of having associated metadata, have to be annotated, and the texts, at least, segmented at the sentence and token levels, with lemmas and part-of-speech (POS) marked, while the speech recordings – minimally accompanied by their transcribed textual content and aligned with the text at sentences and words. The metadata and the annotations were thought such as to allow the access to the corpus of expert linguists but also of the large public, through an easy-to-use interface. This interface should give access to the corpus content by formulating complex queries that include details contained in the metadata (such as source, author, year of publishing, domain, style, etc.) and in the annotation (at the lexical and morphological levels).

THE MAKING-UP OF COROLA

The project was started on January 2014, with the kick-off and dissemination meeting held in the Aula of the Romanian Academy Library, in Bucharest, on 3rd of February 2014, and will finish at the end of 2017, with a closing and dissemination session, to be held in the same Aula, on 14th December 2017.

Apart from the members of the project consortium and the support of experts in linguistics from University of Bucharest (Departments of Linguistics and Foreign Languages), "Alexandru Philippide" Institute of Philology of the Romanian Academy, Iaşi branch, and computer scientists from the Institute of German Language in Mannheim, Germany, the project was voluntarily contributed also by students from "Alexandru Ioan Cuza" University of Iaşi, University "Politehnica" of Bucharest, University of Bucharest, Technical University of Cluj-Napoca and University of Craiova. The students contributed mainly in cleaning the texts, filling-in metadata and doing voice recordings.

Our presentation will follow roughly the process followed in the building of the Corpus.

COPYRIGHTS

In order to develop the corpus, we have contacted representatives of important publishing houses and editorial offices that publish Romanian contemporary writers, as well as bloggers, individual authors, Radio and TV channels, and even theaters. So far, we have signed written agreements with 17 publishing houses, 16 magazines, 16 bloggers, and 40 individual authors, 3 Radio channels and one TV channel.

Their willingness to get involved and to negotiate the conditions for our collaboration was like finding a gold mine in the benefit of the Corpus. We have mentioned all text providers in the list of Table 1, irrespective of the size of their contributions. The process of persuading publishing houses and media to collaborate in the elaboration of CoRoLa was very laborious and, sometimes, it was quite slow. In the case of documents available online, we have developed crawlers that extracted automatically the raw TXT format out of the HTML format. As for the text files received from providers, mainly in the form of PDF and DOC files, they were also converted to TXT. But we included in our collection also texts that are outside the scope of the copyright law, such as those belonging to the law and administrative domains, which have been simply downloaded and added to CoRoLa.

Table 1

The complete list of text providers for CoRoLa

Style	**Providers**
Imaginative	Editura Humanitas, Editura Polirom, România literară, Destine literare, The journal of Colegiul Național „Unirea" Focşani, Editura PIM, Editura Institului European, Editura Adenium, Casa Editorială Demiurg, Editura ARS Longa
Memoirs	Editura Humanitas, Editura Polirom, Editura PIM
Law	Wikipedia, Romanian Legislation, EU Legislation
Administrative	Wikipedia, Editura PIM, Editura Institului European
Science	Editura Humanitas, Editura Polirom, Editura Academiei Române, Editura Universității din Bucureşti, Editura Economică, Editura Simetria, Editura Muzica, România literară, Editura PIM, Wikipedia
Journalistic	România literară, Actualitatea muzicală, Destine literare, Ziarul Agend, Revista UZP, Revista Medicală Română, Revista Balcanii şi Europa, Revista Candela de Montreal, Revista Timpul, DCNEWS
Blog	www.opiniastudenteasca.ro, www.zilesinopti.ro, www.uzp.org.ro, www.scriitorii romani.com, www.printreranduri.eu, www.belva.ro, www.simonatache.ro, www.blog deparinti.ro, www.simonatache.ro, www.ramonacervenciuc.ro, www.andreeaignat.ro,
Authors	Luminița Cărăuşu, Zeno Fodor, Corneliu Leu, Liviu Petcu, Andrei Anton Popescu, Adina Ciubotariu, Daniela Gîfu, Dan Cristea, Alexandru Iliescu, Alexandru Sălăvăstru, Pânzariu Anca, Mihaela Bețа, Ana-Maria Timofciuc, Ana-Maria Crețu, Andreea Țigănescu, Diana – Alexandra Soponaru, Ana-Maria Lungu, Alina Leonte, Mădălina Maria Bîrzu, Cosmin-Constantin Andrei, Lavinia Maria Băisan, Andreea Evelina Levițchi, Cristian Radu, Purice Gabriela-Diana, Căciulă Ionuț Răducu, Luca Andrei Cristian, Ioan Baciu, Roxana Luminița Belciug, Adriana Moroşan, Ioana Curcă, Cătălina Cojocari. Roxana Hrăniciuc, Adriana Chițac, Gabriela Torică, Evelina Zaporojanu, Maria-Tereza Barnea, Adina Zaharia, Sacaloş Francesca, Amalia Maria Tanasă, Monica Pălimariu

METADATA AND CLEANING

METADATA

Metadata are standardized information blocks attached to primary data. Their role is to make explicit information regarding the title of the document, the list of authors, the location and other important information about the original files. Metadata are essential for indexing the corpus and can be used to draw sophisticated query criteria by the end users.

Out of the many schemas developed recently to accommodate metadata of primary resources (the general purpose schema – Dublin Core [8], the Metadata Object Description Schema [16], the Science schema – Darwin Core [17], the NASA's Standard [18]), we have opted to use the CMDI [19] (Component MetaData Infrastructure) format. CMDI offers ready-made sets of metadata elements (components) for various types of resources. They can be edited, modified, and combined to generate personalized metadata schemas (profiles). The CMDI model has close ties to the ISOcat data category registry [20].

```
<root>
  <Metadata>
        <DocumentTitle>Serie Noua, Anul XVIII (90), Nr. 2</DocumentTitle>
        <ArticleTitle>Patima desfrânarii si biruirea ei în viziunea
             spiritualitatii ortodoxe ( I )</ArticleTilte>
        <AuthorName>Liviu Petcu</AuthorName>
        <PublicationData>2008</PublicationData>
        <Source>Journal</Source>
        <SourceName>Revista Teologica</SourceName>
        <TranslatorName>-</TranslatorName>
        <Medium>Written</Medium>
        <DocumentType>Proceedings</DocumentType>
        <DocumentTextGenre>Publicistic</DocumentTextGenre>
        <CollectionDate>2008</CollectionDate>
        <SubjectLanguage>Româna</SubjectLanguage>
        <ISSN-ISBN>1222-9695</ISSN-ISBN>
  </Metadata>
</root>
```

Fig. 1. Example of the metadata schema of a CoRoLa document

CLEANING THE DATA

In order for the textual data to be automatically annotated and indexed, the raw texts extracted from the primary sources had to go through a cleaning process. The cleaning was needed for a variety of reasons: incorrectly encoded text (non-Unicode characters), non-alphabetical sequences in the position of Romanian diacritics (particularly in cases where the primary sources were PDF files), internal word hyphenation at the end of lines, etc. Moreover, any formatting applied to the original document, such as title of the book or article, name of author(s), editor, year of publishing, table of contents, headers and footers, etc., had to be removed. In addition, those sentences that had too many non Romanian characters were also removed, to avoid inclusion of citations in Cyrillic or Greek, less relevant in the context of a modern Romanian corpus. Let's note however, that citations in other languages than Romanian occurring in the Latin alphabet could not be detected in the cleaning phase, but only much later, after the POS-tagging process.

Generally, the process of cleaning the primary textual data can be seen as made up of two steps: an initial cleaning (of the raw text) and a more elaborate cleaning (after the addition of metadata).

All texts were acquired in electronic form, but the format of the basic files and their encoding were largely different: some were MS DOC files, with characters encoded in Windows 1252 charset, others were TXT files, usually encoded in Unicode, and still others were PDF files encoded in ASCII, with special sequences or characters for diacritics. In order to get a uniform encryption of all files in the Unicode format, an automatic correction phase processed each type of file appropriately.

The texts extracted from PDF files gave rise to further complications. Firstly, PDF files do not usually encode Romanian diacritics properly, as the base text is ASCII, and a font is used to show the correct glyph for the character. Furthermore, different PDF files use different fonts (and, of course, different fonts can occur in the same document), such that some ASCII character can represent one letter in some cases, and another one in others. In order to solve this, a tool was built, which, on the basis of a table recording the frequency of the symbol in question, translates, on a case by case basis, all the special ASCII characters to their proper Unicode equivalents [12]. The text extracted from PDF files also contained headers and footers from the initial document. In some cases these could be automatically identified and removed [12], in others they had been removed manually. In addition, the paragraph identification had to be automatically corrected, as most PDF files add a newline symbol at the end of each typographical line and two such characters at the end of each paragraph. After the paragraphs were identified, the end-of-line hyphens separating syllables were removed. This tool uses a glossary of Romanian words to estimate which hyphens should be removed [12].

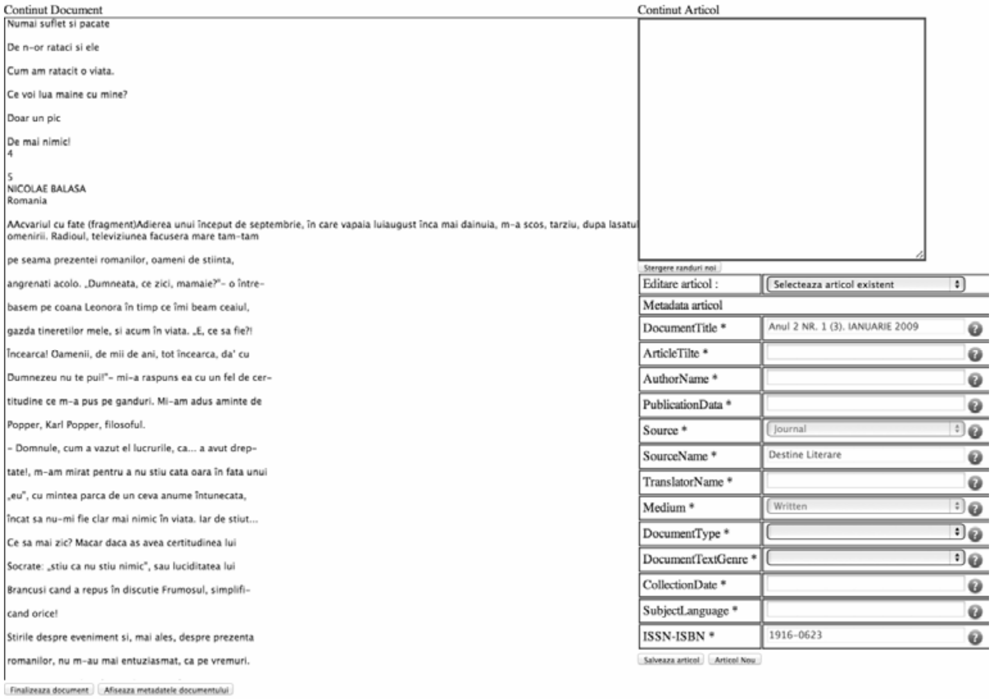
The corrections described above were followed by the partially manual – partially automatic fill-in of the metadata for each document. Human correctors had then a final look over the text and did other manual corrections (removal of tables and formulas, elimination of footnotes and references, etc). After these manual corrections, the resulting raw text needed a new round of processing in order to further remove those elements not needed in the final version of the corpus, as is the title of the book or article, and the list of authors (data already present in the metadata). A specialized tool removes the first occurrence of the title and the list of authors from the beginning of the text. Also at this stage, we have automatically removed all of sentences containing excessive amounts of foreign characters

(Greek, Cyrillic, Coptic, etc.). The threshold over which a sentence is deemed to have too many foreign symbols was determined empirically and set at 40% of the length of the sentence or paragraph.

CODAP: A CORPUS PROCESSING FABLIC

When constructing corpora from different PDF formats, the extraction of relevant text from pages is not a trivial task due to the great amount of irrelevant text, images, tables, formulas, etc. that needs to be identified and removed so that they do not compromise the quality of the corpus. This task, called *boilerplate removal* in the literature, consists of categorizing PDF content as valuable *versus* irrelevant. Figure 2 displays a snapshot from a working session with CoDaP – the CoRoLa Data Cleaning and Metadata Annotation Platform.

Two different files are resulting out of this platform. First, the cleared TXT file, prepared for the linguistic chain and an XML file with the metadata content. Apart from having an important role in querying of the corpus, metadata are relevant also for drawing statistics about the quantity of different components of the corpus, like domains, sub-domains and styles of the texts making up the corpus. As such, the features of representativeness and balance of the corpus can be properly estimated.



Numai suflet si pacate

De n-or rataci si ele

Cum am ratacit o viata.

Ce voi lua maine cu mine?

Doar un pic

De mai nimic!
4

5
NICOLAE BALASA
Romania

AAcvariul cu fate (fragment)Adierea unui început de septembrie, în care vapaia luiaugust înca mai dainuia, m-a scos, tarziu, dupa lasatul
omenirii. Radioul, televiziunea facusera mare tam-tam

pe seama prezentei romanilor, oameni de stiinta,

angrenati acolo. „Dumneata, ce zici, mamaie?"- o între-

basem pe coana Leonora în timp ce îmi beam ceaiul,

gazda tineretilor mele, si acum în viata. „E, ce sa fie?!

Încearca! Oamenii, de mii de ani, tot încearca, da' cu

Dumnezeu nu te pui!"- mi-a raspuns ea cu un fel de cer-

titudine ce m-a pus pe ganduri. Mi-am adus aminte de

Popper, Karl Popper, filosoful.

- Domnule, cum a vazut el lucrurile, ca... a avut drep-

tate!, m-am mirat pentru a nu stiu cata oara în fata unui

„eu", cu mintea parca de un ceva anume întunecata,

încat sa nu-mi fie clar mai nimic în viata. Iar de stiut...

Ce sa mai zic? Macar daca as avea certitudinea lui

Socrate: „stiu ca nu stiu nimic", sau luciditatea lui

Brancusi cand a repus în discutie Frumosul, simplifi-

cand orice!

Stirile despre eveniment si, mai ales, despre prezenta

romanilor, nu m-au mai entuziasmat, ca pe vremuri.

Editare articol :	Selecteaza articol existent
Metadata articol	
DocumentTitle *	Anul 2 NR. 1 (3). IANUARIE 2009
ArticleTilte *	
AuthorName *	
PublicationData *	
Source *	Journal
SourceName *	Destine Literare
TranslatorName *	
Medium *	Written
DocumentType *	
DocumentTextGenre *	
CollectionDate *	
SubjectLanguage *	
ISSN-ISBN *	1916-0623

Fig. 2. A working session with CoDaP

One of the first steps involved in processing audio files is their transcription into text according to the Romanian contemporary orthography. The in-house transcriptions have been processed through a free tool for speech processing (Praat) [4], which provides two levels: *transcription* tiers and *speaker* tier (see Fig. 3). In the transcription process, for each speaker, the beginning and the end of each utterance were pointed by means of markers. Other textualized events are: repetitions, hesitations, disfluencies, laughter, breathing, overlapping talk, foreign words, etc.

Because transcription is a time-consuming task we requested help from volunteers. They were asked to transcribe the audio files ignoring the particularities of speech such as laughter or hesitations, and omitting marking the beginning and the end of each utterance. These transcriptions are edited for further processing.

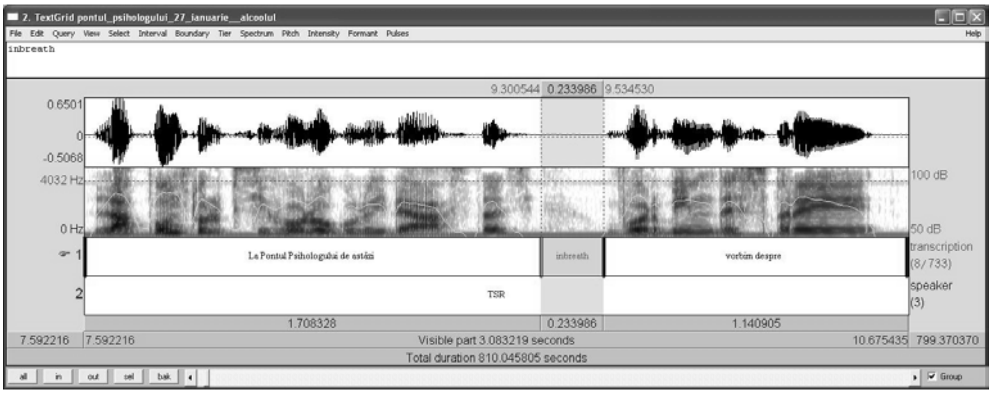
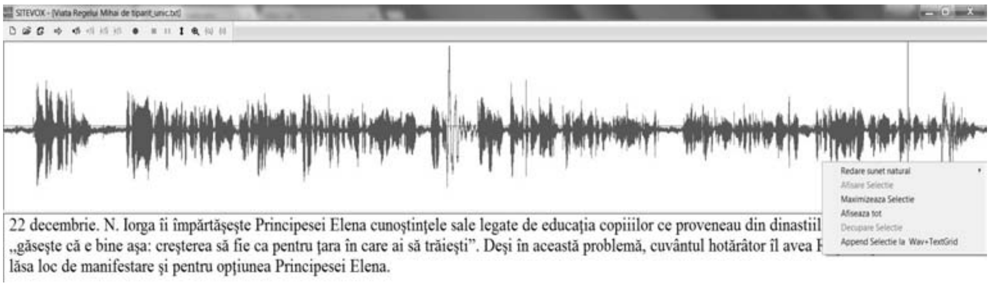

2. TextGrid pontul_psihologului_27_ianuarie__alcoolul
File Edit Query View Select Interval Boundary Tier Spectrum Pitch Intensity Formant Pulses
Help
inbreath
9.300544
0.233986
9.534530
0.6501
0
-0.5068
4032 Hz
0 Hz
100 dB
50 dB
1
La Pontul Psihologului de astăzi
inbreath
vorbim despre
transcription
(8/733)
2
TSR
speaker
(3)
1.708328
0.233986
1.140905
7.592216
7.592216
Visible part 3.083219 seconds
10.675435
799.370370
Total duration 810.045805 seconds
all
in
out
sel
bak
Group

Fig. 3. A Praat transcription session

For the in-house spoken text recordings we used a specially designed interface to align the spoken utterances with their corresponding text (the interface developed in our institute, see Figure 4).



22 decembrie. N. Iorga îi împărtăşeşte Principesei Elena cunoştinţele sale legate de educaţia copiiilor ce proveneau din dinastiil
„găseşte că e bine aşa: creşterea să fie ca pentru ţara în care ai să trăieşti". Deşi în această problemă, cuvântul hotărâtor îl avea F
lăsa loc de manifestare şi pentru opţiunea Principesei Elena.

Fig. 4. An annotation session with the in-house read speech tool

ANNOTATIONS: THE LINGUISTIC CHAIN

The annotation of the cleaned TXT documents is realized in a pipeline of processors, which perform, in sequence: segmentation at the sentence level, tokenization, POS-tagging and lemmatization. All these levels are XML marked, as exemplified in Figure 5.

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<POS_Output>
   <S id="1" offset="0">
   <W Case="direct" Definiteness="no" Gender="feminine" LEMMA="prefață" MSD="Ncfsrn"
     Number="singular" POS="NOUN" Type="common" id="1.1" offset="0">Prefață</W>
   <W Case="direct" Gender="feminine" LEMMA="acest" MSD="Dd3fsr---e" Number="singular"
     POS="DETERMINER" Person="third" Positioning="prenominal" Type="demonstrative" id="1.2"
     offset="8">Această</W>
   <W Case="direct" Definiteness="no" Gender="feminine" LEMMA="carte" MSD="Ncfsrn"
     Number="singular" POS="NOUN" Type="common" id="1.3" offset="16">carte</W>
   <W EXTRA="tranzitiv" LEMMA="prezenta" MSD="Vmip3p" Mood="indicative" Number="plural"
POS="VERB"
     Person="third" Tense="present" Type="predicative" id="1.4" offset="22">prezintă</W>
   <W Case="direct" Definiteness="no" Gender="feminine" LEMMA="noțiune" MSD="Ncfprn"
     Number="plural" POS="NOUN" Type="common" id="1.5" offset="31">noțiuni</W>
   <W LEMMA="de" MSD="Sp" POS="ADPOSITION" id="1.6" offset="39">de</W>
   <W Case="direct" Definiteness="no" Gender="feminine" LEMMA="bază" MSD="Ncfsrn"
Number="singular"
     POS="NOUN" Type="common" id="1.7" offset="42">bază</W>
   <W LEMMA="din" MSD="Sp" POS="ADPOSITION" id="1.8" offset="47">din</W>
...
</S>
</POS_Output>
```

Fig. 5. An annotated fragment of text

STATISTICS REGARDING THE CORPUS

THE TEXTUAL COMPONENT

Two months before the end of the project, the original target assumed (a dimension of the textual component of 500.000.000 words), was accomplished, as CoRoLa counts now a total of 1.257.745.725 words in 386.501 files. As explained already, all files are paired by metadata, and annotated at the sentence and morpho-lexical levels.

Figures 6 and 7 show the actual number of words in all the domains and styles, respectively. The huge difference between the Society domain, respectively the Law style, and all others is due to the inclusion of a large collection of legal texts.

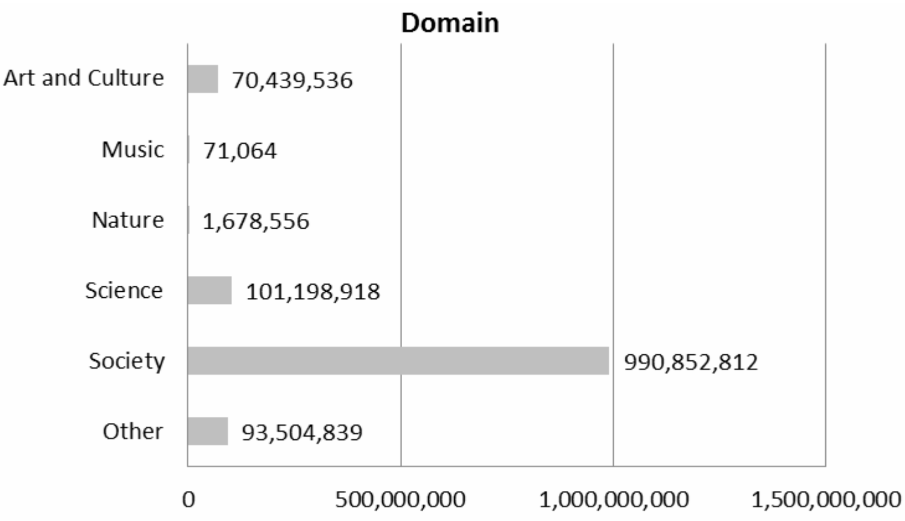


Fig. 6. No. of words by Domains

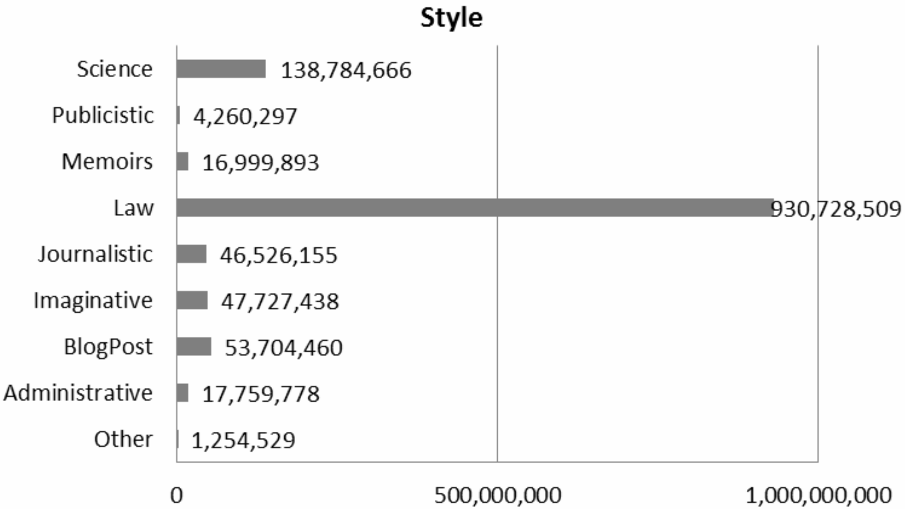


Fig. 7. No. of words by Styles

Figures 8 and 9 show the proportions after a possible reduction of the quantity of legal texts that would still maintain the promised target of 500 million words. Even slashed this way we would still have to accept a distortion of approx. 10 to 1 between the Society domain and the largest collection of the other domains – the Science texts. However, in terms of the Style, the unbalance reduces to only 1.25. It is clear that a simultaneous realization of representativeness and balance is still a target to go for in the future.

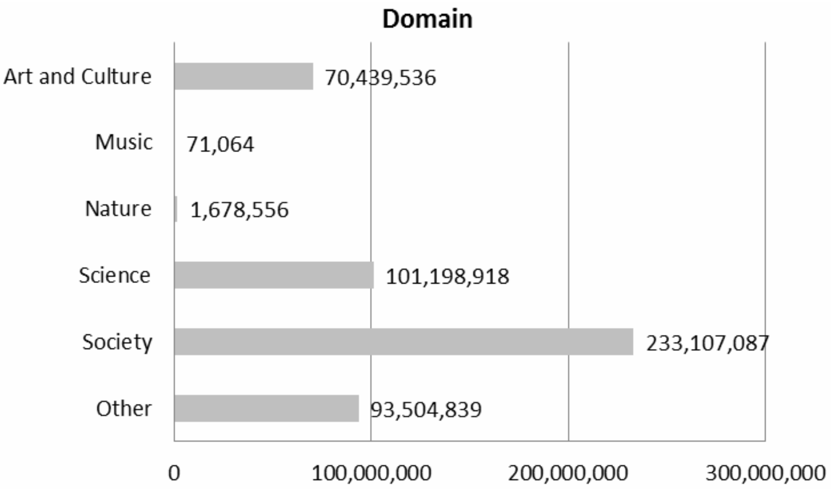


Fig. 8. No. of words by Domain after reduction

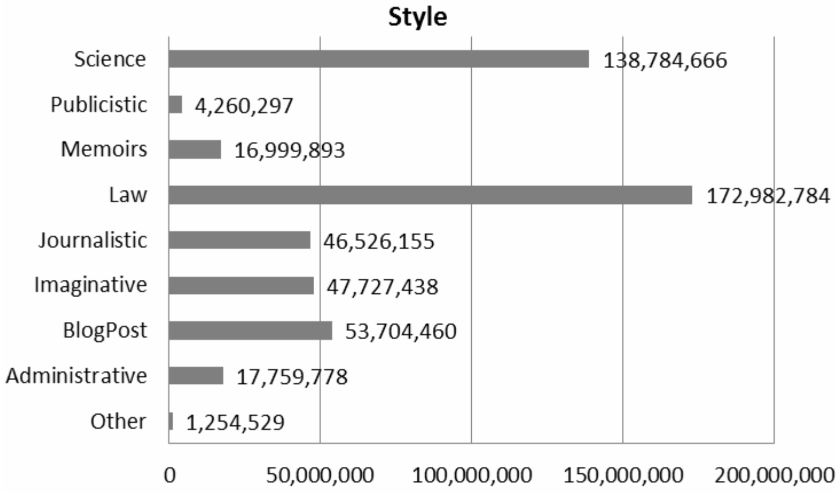


Fig. 9. No. of words by Style after reduction

THE SPOKEN COMPONENT

The collected speech corpora are transcribed according to the Romanian contemporary orthography (*i.e.* they are not phonetically transcribed). All speakers have been aware of the presence of the microphone, which means that part of the naturalness of pronunciation could, sometimes, been lost. However, in the case of radio and TV interviews the naturalness of the conversation manifests through events like hesitations, overlaps, pauses, disfluencies, and so on, the interviews being spontaneous, unscripted. Other types of recordings are represented by audio books, actors voice recordings in in-theatre performances or rehearsals and in-house recordings, the last category operated by one of our colleagues and by volunteers. In this case the speakers used a specially designed interface to align the spoken utterances with their corresponding textual parts, the texts being selected from the textual corpus. In all cases, the textual counterparts of the aligned speech-text corpus items have themselves been processed with the same pipeline as all the other texts of the corpus: sentence segmentation, tokenization, lemmatization and POS-tagging. Table 2 gives statistics about the primary voice recordings operated in Iaşi.

Table 2
The content of the speech components

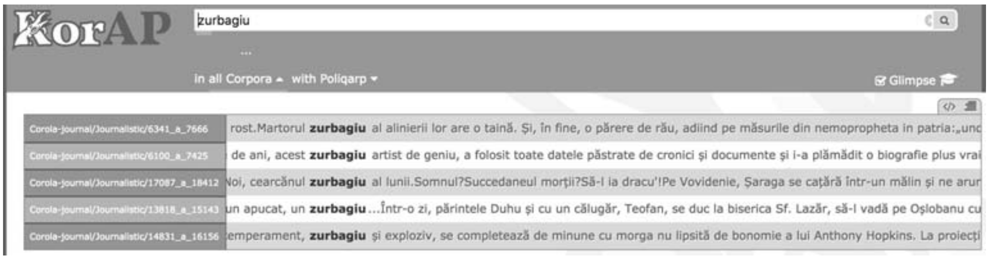
Source of the corpus	No. speakers	Representing	Time length (h)
Roman TV	many	interviews	28
Radio VIVA	many	interviews	4
Radio Iaşi	many	news & interviews	33
Radio Universitas	many	news & interviews	22
Teatrul Național Târgu-Mureş	many	theatre performances & rehearsals	8
Audio books	single	spoken book	40
Prof. Luminita Cărăuşu	many	exercises of intonation	4
In-house read text recordings	many	read speech	11
Total			**150**

USER INTERFACES

KORAP

Bański *et al.* [1] present a scalable corpus analysis platform designed to serve very large datasets, called KorAP. The platform and its interface, functioning as a concordances program, opens to any interested user the possibility to query large, multiply-level annotated corpora. The morphologically processed CoRoLa corpus has been adapted for the requirements of the KorAP interface and can be used with the "drukola" – prefix[1]. Below in this section we present a few examples of interrogations expressed in the Poliqarp Query Language.

I. Searching simple occurrences: `zurbagiu` (Fig. 10)



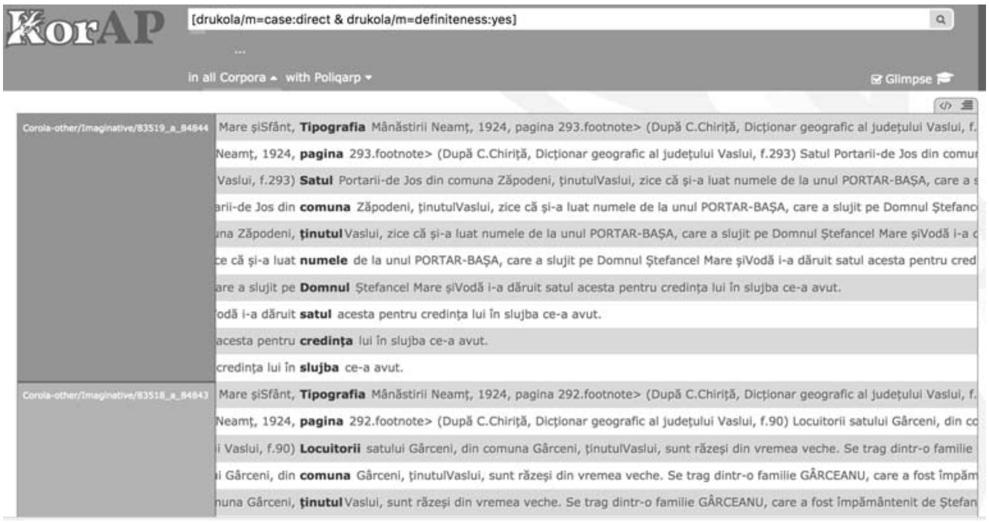
Corola-journal/Journalistic/6341_a_7666	rost.Martorul **zurbagiu** al alinierii lor are o taină. Şi, în fine, o părere de rău, adiind pe măsurile din nemopropheta in patria:„unc
Corola-journal/Journalistic/6100_a_7425	de ani, acest **zurbagiu** artist de geniu, a folosit toate datele păstrate de cronici şi documente şi i-a plămădit o biografie plus vrai
Corola-journal/Journalistic/17087_a_18412	Noi, cearcănul **zurbagiu** al lunii.Somnul?Succedaneul morții?Să-l ia dracu'!Pe Vovidenie, Şaraga se cațără într-un mălin şi ne arur
Corola-journal/Journalistic/13818_a_15143	un apucat, un **zurbagiu**...Într-o zi, părintele Duhu şi cu un călugăr, Teofan, se duc la biserica Sf. Lazăr, să-l vadă pe Oşlobanu cu
Corola-journal/Journalistic/14831_a_16156	emperament, **zurbagiu** şi exploziv, se completează de minune cu morga nu lipsită de bonomie a lui Anthony Hopkins. La proiecți

Fig.10. KorAP interface – simple occurrences

[1] From the URL: http://89.38.230.10:5555/

II.Searching a single definite noun token in nominative/accusative (Fig. 11):

```
[drukola/m=case:direct & drukola/m=definiteness:yes]
```



[drukola/m=case:direct & drukola/m=definiteness:yes]

Corola-other/Imaginative/83519_a_84844

Mare șiSfânt, **Tipografia** Mânăstirii Neamț, 1924, pagina 293.footnote> (După C.Chiriță, Dicționar geografic al județului Vaslui, f.

Neamț, 1924, **pagina** 293.footnote> (După C.Chiriță, Dicționar geografic al județului Vaslui, f.293) Satul Portarii-de Jos din comu

Vaslui, f.293) **Satul** Portarii-de Jos din comuna Zăpodeni, ținutulVaslui, zice că și-a luat numele de la unul PORTAR-BAȘA, care a s

arii-de Jos din **comuna** Zăpodeni, ținutulVaslui, zice că și-a luat numele de la unul PORTAR-BAȘA, care a slujit pe Domnul Ștefanc

una Zăpodeni, **ținutul** Vaslui, zice că și-a luat numele de la unul PORTAR-BAȘA, care a slujit pe Domnul Ștefancel Mare șiVodă i-a c

ce că și-a luat **numele** de la unul PORTAR-BAȘA, care a slujit pe Domnul Ștefancel Mare șiVodă i-a dăruit satul acesta pentru cred

are a slujit pe **Domnul** Ștefancel Mare șiVodă i-a dăruit satul acesta pentru credința lui în slujba ce-a avut.

odă i-a dăruit **satul** acesta pentru credința lui în slujba ce-a avut.

acesta pentru **credința** lui în slujba ce-a avut.

credința lui în **slujba** ce-a avut.

Corola-other/Imaginative/83518_a_84843

Mare șiSfânt, **Tipografia** Mânăstirii Neamț, 1924, pagina 292.footnote> (După C.Chiriță, Dicționar geografic al județului Vaslui, f.

Neamț, 1924, **pagina** 292.footnote> (După C.Chiriță, Dicționar geografic al județului Vaslui, f.90) Locuitorii satului Gârceni, din co

i Vaslui, f.90) **Locuitorii** satului Gârceni, din comuna Gârceni, ținutulVaslui, sunt răzeși din vremea veche. Se trag dintr-o familie

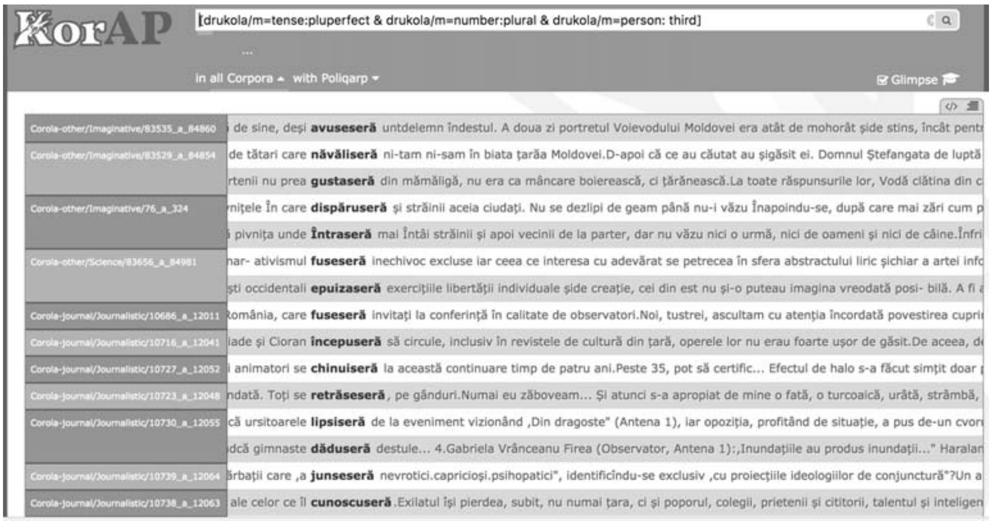
i Gârceni, din **comuna** Gârceni, ținutulVaslui, sunt răzeși din vremea veche. Se trag dintr-o familie GÂRCEANU, care a fost împăm

nuna Gârceni, **ținutul** Vaslui, sunt răzeși din vremea veche. Se trag dintr-o familie GÂRCEANU, care a fost împământenit de Ștefan

Fig. 11. The KorAP interface – definite nouns in the direct case

III. Searching verbs in past perfect, plural, 3rd person:

`[drukola/m=tense:pluperfect & drukola/m=number:plural & drukola/m=person: third]` (Fig. 12)



[drukola/m=tense:pluperfect & drukola/m=number:plural & drukola/m=person: third]

Corola-other/Imaginative/83535_a_84860

de sine, deși **avuseseră** untdelemn îndestul. A doua zi portretul Voievodului Moldovei era atât de mohorât șide stins, încât pent

Corola-other/Imaginative/83529_a_84854

de tătari care **năvăliseră** ni-tam ni-sam în biata țarăa Moldovei.D-apoi că ce au căutat au șigăsit ei. Domnul Ștefangata de luptă

rtenii nu prea **gustaseră** din mămăligă, nu era ca mâncare boierească, ci țărănească.La toate răspunsurile lor, Vodă clătina din c

Corola-other/Imaginative/76_a_324

nițele în care **dispăruseră** și străinii aceia ciudați. Nu se dezlipi de geam până nu-i văzu înapoindu-se, după care mai zări cum p

pivnița unde **întraseră** mai întâi străinii și apoi vecinii de la parter, dar nu văzu nici o urmă, nici de oameni și nici de câine.Înfri

Corola-other/Science/83656_a_84981

nar- ativismul **fuseseră** inechivoc excluse iar ceea ce interesa cu adevărat se petrecea în sfera abstractului liric șichiar a artei info

ști occidentali **epuizaseră** exercițiile libertății individuale șide creație, cei din est nu și-o puteau imagina vreodată posi- bilă. A fi

Corola-journal/Journalistic/10686_a_12011

România, care **fuseseră** invitați la conferință în calitate de observatori.Noi, tustrei, ascultam cu atenția încordată povestirea cupri

Corola-journal/Journalistic/10716_a_12041

iade și Cioran **începuseră** să circule, inclusiv în revistele de cultură din țară, operele lor nu erau foarte ușor de găsit.De aceea, d

Corola-journal/Journalistic/10727_a_12052

i animatori se **chinuiseră** la această continuare timp de patru ani.Peste 35, pot să certific... Efectul de halo s-a făcut simțit doar

Corola-journal/Journalistic/10723_a_12048

ndată. Toți se **retrăseseră**, pe gânduri.Numai eu zăboveam... Și atunci s-a apropiat de mine o fată, o turcoaică, urâtă, strâmbă,

Corola-journal/Journalistic/10730_a_12055

că ursitoarele **lipsiseră** de la eveniment vizionând ,Din dragoste" (Antena 1), iar opoziția, profitând de situație, a pus de-un cvor

dcă gimnaste **dăduseră** destule... 4.Gabriela Vrânceanu Firea (Observator, Antena 1):,Inundațiile au produs inundații..." Haralan

Corola-journal/Journalistic/10739_a_12064

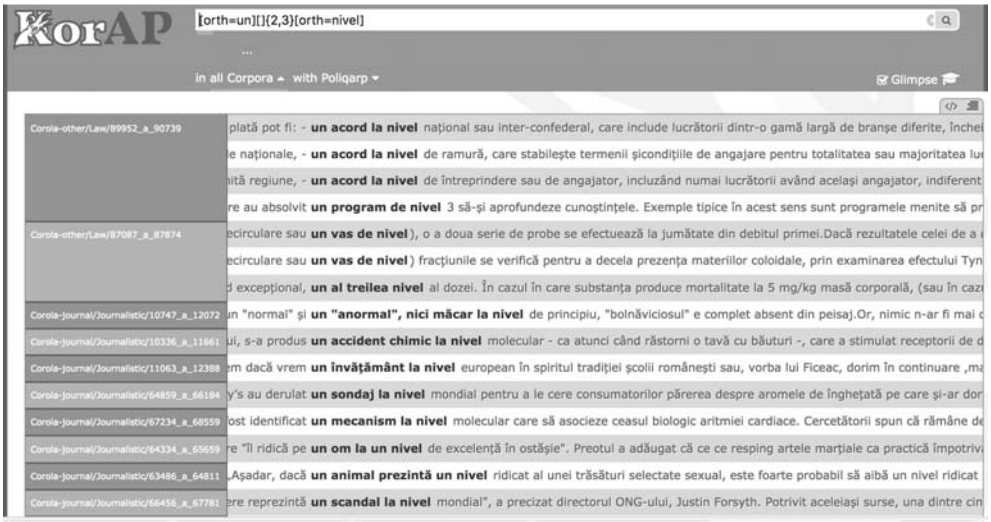
ărbații care ,a **junseseră** nevrotici.capricioși.psihopatici", identificîndu-se exclusiv ,cu proiecțiile ideologiilor de conjunctură"?Un a

Corola-journal/Journalistic/10738_a_12063

ale celor ce îl **cunoscuseră**.Exilatul își pierdea, subit, nu numai țara, ci și poporul, colegii, prietenii și cititorii, talentul și inteligen

Fig. 12. The KorAP interface – searching for verbs complying with some morphological constrains

IV. Defining a distance between tokens in Poliqarp, by using empty tokens: the undefined article `un` preceding at a distance of 2 or 3 items the word `nivel`: `[orth=un][]{2,3}[orth=nivel]` (Fig. 13).



Corola-other/Law/89952_a_90739	plată pot fi: - **un acord la nivel** național sau inter-confederal, care include lucrătorii dintr-o gamă largă de branșe diferite, înche e naționale, - **un acord la nivel** de ramură, care stabilește termenii șicondițiile de angajare pentru totalitatea sau majoritatea lu ită regiune, - **un acord la nivel** de întreprindere sau de angajator, incluzând numai lucrătorii având același angajator, indiferent re au absolvit **un program de nivel** 3 să-și aprofundeze cunoștințele. Exemple tipice în acest sens sunt programele menite să pr
Corola-other/Law/87087_a_87874	ecirculare sau **un vas de nivel**), o a doua serie de probe se efectuează la jumătate din debitul primei.Dacă rezultatele celei de a ecirculare sau **un vas de nivel**) fracțiunile se verifică pentru a decela prezența materiilor coloidale, prin examinarea efectului Tyn d excepțional, **un al treilea nivel** al dozei. În cazul în care substanța produce mortalitate la 5 mg/kg masă corporală, (sau în caz
Corola-journal/Journalistic/10747_a_12072	n "normal" și **un "anormal", nici măcar la nivel** de principiu, "bolnăviciosul" e complet absent din peisaj.Or, nimic n-ar fi mai
Corola-journal/Journalistic/10336_a_11661	ui, s-a produs **un accident chimic la nivel** molecular - ca atunci când răstorni o tavă cu băuturi -, care a stimulat receptorii de d
Corola-journal/Journalistic/11063_a_12388	m dacă vrem **un învățământ la nivel** european în spiritul tradiției școlii românești sau, vorba lui Ficeac, dorim în continuare ,m
Corola-journal/Journalistic/64859_a_66184	y's au derulat **un sondaj la nivel** mondial pentru a le cere consumatorilor părerea despre aromele de înghețată pe care și-ar do
Corola-journal/Journalistic/67234_a_68559	ost identificat **un mecanism la nivel** molecular care să asocieze ceasul biologic aritmiei cardiace. Cercetătorii spun că rămâne d
Corola-journal/Journalistic/64334_a_65659	e "îl ridică pe **un om la un nivel** de excelență în ostășie". Preotul a adăugat că ce ce resping artele marțiale ca practică împotriv
Corola-journal/Journalistic/63486_a_64811	.Așadar, dacă **un animal prezintă un nivel** ridicat al unei trăsături selectate sexual, este foarte probabil să aibă un nivel ridicat
Corola-journal/Journalistic/66456_a_67781	ere reprezintă **un scandal la nivel** mondial", a precizat directorul ONG-ului, Justin Forsyth. Potrivit aceleiași surse, una dintre cin

Fig. 13. The KorAP interface – searching for tokens

OTHER INTERFACES

NL2CQP is an interface developed by our colleagues in ICIA Bucureşti, which allows the interrogation of the Corpus by expressing queries in a controlled (Romanian) language. Examples are: *5 fraze în care cuvântul “lumină” apare după verbul “a aprinde”* (*5 sentences in which the word “light” occurs after the verb “to lit”*), *100 fraze în care prepoziția “de” este urmată imediat de un verb la participiu* (*100 sentences in which the preposition “de” is immediately followed by a verb in participle*)).

The speech component contains about 300 hours of recordings (WAV files), transcripts (TXT format) and annotations (XML format) for part of speech, token and lemma. CoRoLa’s speech query interface[2] allows the interrogation of speech corpus by “word”, “lemma” and the morpho-syntactic tag. It also gives the possibility to choose the display method of the results and to listen to the recording of the chosen word, cut from its context of occurrence, or the whole sentence it belongs to.

[2] Accessible at http://89.38.230.23/corola_sound_search/index.php

CONCLUSIONS

COROLA IN THE CONTEXT OF INTERNATIONAL PRECCUPATIONS FOR REPRESENTATIVE CORPORA

In the field of corpus linguistics there is no other scholar more entitled to be named the father of this domain than professor John Sinclair, former director of the Collins series of dictionaries, who developed the first corpus of British English in the sixties. He was the one to show, by accessing original data, that a word does not carry meaning by itself, but only in the context of the surrounding words. He established the typology of corpora [13], and since then, people found out the treasures they can discover in well mastered examples, able to bring to light patterns of language use.

One of the largest corpora in the world is the contemporary corpus of German language, created starting with 1964 at IDS Mannheim. Since 2004, the reference corpus of the German language is called DeReKo. It includes now more than 25 billion words [9, 10] and is updated and expanded continuously.

For four years, CoRoLa has been a project of national priority for the Romanian Academy. Since 2014, when it started, various tools necessary in developing the corpus, CoDaP – the CoRoLa Data Cleaning and Metadata Annotation Platform [3], the website and the speech interface were created. The corpus was developed continuously during the project lifetime, constantly acquiring texts and speeches [2].

In 2016, the DruKoLa[3] project started, centered on finding correlations between the German language (as reveled in DeReKo [5, 10]) and the Romanian language (due to CoRoLa [15]). DRuKoLa is a transdisciplinary project involving corpus linguistics, computational linguistics and cross-linguistic studies. One of its important objectives is the construction, provisioning and harmonization of corpora in the German and Romanian languages.

EuReCo (European conference Corpus) [11] is a joint project which aims to harmonize three European corpora, DeReKo, Corola and the Hungarian National Corpus, with respect to metadata, annotation conventions and query interfaces.

COROLA AT THE END OF THE PROJECT

In this paper we presented the technology of development and the final statistics at the end of the elaboration of CoRoLa, the Computational Representative Corpus of Contemporary Romanian Language.

Comparing the goals established at the moment of the start of the project, January 2014, with the achieved results, one month before its ending, we can posit

[3] DRuKoLA is funded by the Alexander von Humboldt-Foundation, and includes the University of Bucharest, IDS Mannheim, RACAI and IIT as parteners.

a successful enterprise. The visionary goal of 500 million occurrences in all functional styles and domains was exceeded: occurrences count now over 1.2 billion. All texts were cleaned, accompanied with metadata and pre-processed at the lexical and morphological levels.

The corpus has the following main features:

- it is representative, because it includes texts of all genres (prose, poetry, drama, science, journalism, etc.) and all fields (literature, science, politics, humanism, etc.);
- it reflects contemporary Romanian language, since it includes only text and voice recordings from the Second World War to the present day;
- it is IPR safe, because all proprietary texts and voice recordings were included with the written accept of authors;
- when open to the public[4], it will be free for online consultation.

A LIFE-TIME ENTERPRISE – TOWARDS A DIACHRONIC CORPUS OF THE ROMANIAN LANGUAGE

Planning the future should always start from reconsidering the past. This is especially true in the case of language. The Romanian Academy, by its statute, has always been focused primary on the growth and consolidation of the Romanian language, one of its most significant achievements being the elaboration of the Thesaurus Dictionary of the Romanian Language, during one century of intensive activity. As such, it is not surprising that this institution decided to commission to two of its information technology institutes that deal with computational linguistics and natural language processing the construction of the first large corpus of the Romanian language in electronic form.

However, reaching a successful end, the CoRoLa project cannot stop here. Many more things should be added to the already developed platform in order to consolidate this work, up to a level that configures a significant technological support of the Romanian language, one which could constantly be used by specialists, teachers of Romanian and the public at large as a working and research environment. In the following we shortly describe necessary future developments.

Apart from a continuous maintenance activity, which is supposed to keep the corpus running and eliminate most of the errors that still remained, the immediate necessary efforts should take into consideration:

a) new levels of annotation: addition of phrase levels markings – minimally, around nouns and verbs, and a syntactic level – in the functional-dependency formalism, by processes that still need amelioration, and, more distant in time, addition of a semantic level – which would disambiguate the senses for the most productive classes (nouns, verbs and adjectives);

[4] Lounch planned on 14th December 2017.

b) inter-linking with other linguistic resources: indexing CoRoLa's occurrences onto a significant Romanian dictionary, such as the Thesaurus Dictionary of Romanian Language in Electronic Form (eDTLR) [7], and the Romanian WordNet (RoWN) [14, 6] The CoRoLa-eDTLR linking could be done bidirectional at the entries level or, if senses are marked, even in the eDTLR tree of senses. These complex inter-links open spectacular possibilities of complex enquiries, in which more than one resource is mentioned in the query.

On the temporal dimension, the project could be extended indefinitely, onto two directions: the future and the past of the language. A continuous acquisition of primary textual documents and records of the moment could seize the evolution of the language. As such, CoRoLa could become a monitor corpus. To obtain diachronicity is more difficult, because indexing lexical tokens in old documents presupposes transcription of Cyrillic Romanian onto Latin Romanian, including the manuscripts. Building a technology able to recognize printed, semi-uncial and hand written documents should then be combined with modeling a diachronic paradigmatic morphology, that would disambiguate the lemmas and morphological features of occurrences in a temporal-free space, irrespective of the chronology of the original document, of their belonging to an obsolete lexicon, or to their variations due to distribution in space, to authors idiosyncratic preferences or to the lack of grammar norms.

Acknowledgements. CoRoLa was a priority program of the Romanian Academy (2014–2017), entirely supported by this institution. DRuKoLa is founded by Alexander von Humboldt Foundation and is a join project involving Institut für Deutsche Sprache (IDS) Mannheim, University of Bucharest (UB), Research Institute for Artificial Intelligence "Mihai Drăgănescu" of the Romanian Academy, Bucharest (RACAI) and the Institute of Computer Science of the Iaşi branch of the Romanian Academy (ARFI-IIT). It is running until the end of 2018. Our colleagues from RACAI have participated in all phases of the acquisition, annotation and development of processing chains for the CoRoLa corpus. The interfaces shortly mentioned in the section 6.2 are developed within this team, authored respectively by Radu Ion and Vasile Pais. Accessing CoRoLa through the KoRAP interface was made possible due to the DRuKoLA project. The interface has been developed in IDS. We express our warm gratitude to all providers of original texts and voice recordings mentioned in Table 1, Section 2. Many students in the Faculties of Letters and Computer Science of the "Alexandru Ioan Cuza" University of Iaşi, the Faculty of Automation, Control and Computers of the Politehnica University of Bucharest and the Faculty of Sciences of the University of Craiova have been involved in voluntary work to clean the texts, to fill in metadata and to record their voices. To all of them, too many to be mentioned here (their list can be found on the CoDaP pages), we express our most grateful thanks. We thank assoc. prof. dr. Ruxandra Cosma from the University of Bucharest and officer of the Humboldt Foundation in Romania for her continuous support and professional collaboration, for establishing working links with the IDS team and for opening doors for accessing Humboldt funds.

Author contributions: DC made sections 1, 7; DG made section 6; AM made subsections 3.2; 3.3; MOP made section 2; LP made subsections 3.3, 5.2; AS made subsection 5.1; CB made subsection 3.1.

REFERENCES

1. BAŃSKI P., DIEWALD N., HANL M., KUPIETZ M., WITT A., *Access Control by Query Rewriting. The Case of KorAP.* Proceedings of the Ninth Conference on International Language Resources and Evaluation (LREC'14): 3817–3822, 2014.
2. BARBU MITITELU V., CRISTEA D., COSMA R. *Corpus of Contemporary Romanian. Architecture, Annotation Levels and Analysis Tools*. The 16th International Conference of the Department of Linguistics: Romanian Linguistics, Romance Linguistics, Bucharest, Nov. 25–26, 2016, https://profs.info.uaic.ro/~dcristea/papers/CoRoLa&DRuKoLA-MititeluCristeaCosma.pdf.
3. BIBIRI D., BOLEA C., SCUTELNICU L. A., MORUZ A., PISTOL L., CRISTEA D., *Metadata of a Huge Corpus of Contemporary Romanian. Data and Organization of the Work*, articol prezentat la Conferința „7th Balkan Conference in Informatics", BCI 2015, Craiova, Romania, Septembrie, 2015.
4. BOERSMA P., WEENINK D., *Praat: doing phonetics by computer* [Computer program]. Version 6.0.35, retrieved 12 October 2017 from http://www.praat.org.
5. COSMA R., CRISTEA D., KUPIETZ M., TUFIŞ D., WITT A., *DRuKoLa – Towards Contrastive German-Romanian Research based on Comparable Corpora*, in Proceedings of the Workshop on the Challenges in the Management of Large Corpora (CMLC-4), Language Resources and Evaluation Conference (LREC), 28 May, Portoroz, 2016.
6. CRISTEA D., MIHĂILĂ C., FORĂSCU C., TRANDABĂȚ D., HUSARCIUC M., HAJA G., POSTOLACHE O., *Mapping Princeton WordNet synsets onto Romanian wordnet synsets*. In Romanian Journal of Information Science and Technology, Dan Tufis (ed.) Special Issue on BalkaNet, Romanian Academy, Bucharest, Romania, special issue on Balkanet, July, ISSN 1453-8245, pages 125–145, 2004.
7. CRISTEA D., HAJA G., RĂSCHIP M., MORUZ A. M., PĂTRAŞCU M. (2011), *Statistici parțiale la încheierea proiectului eDTLR – Dicționarul Tezaur al Limbii Române în format electronic.* În volumul *Lucrările conferinței naționale „Limba română: ipostaze ale variației lingvistice"*, 3−4 decembrie 2010, Catedra de Română, Universitatea din Bucureşti.
8. HILLMANN D. *Using Dublin Core,* http://dublincore.org/documents/2005/08/15/usageguide/, 2005.
9. KUPIETZ M., KEIBEL H., *The Mannheim German Reference Corpus (DeReKo) as a basis for empirical linguistic research.* In: Makoto Minegishi and Yuji Kawaguchi (Eds): Working Papers in Corpus-based Linguistics and Language Education, No. 3. – Tokyo: Tokyo University of Foreign Studies, 53–59. 2009.
10. KUPIETZ M., BELICA C., KEIBEL H., WITT A., *The German Reference Corpus DeReKo: A primordial sample for linguistic research*. In: Calzolari, N. *et al*. (eds.): Proceedings of LREC 2010. 1848-1854, 2010.
11. KUPIETZ M., WITT A., BAŃSKI P., TUFIŞ D., CRISTEA D., *EuReCo – Joining Forces for a European Reference Corpus as a sustainable base for cross-linguistic research.* Published in: Bański, Piotr/Kupietz, Marc/Lüngen, Harald/Rayson, Paul/Biber, Hanno/Breiteneder, Evelyn/ Clematide, Simon/Mariani, John/stevenson, Mark/Sick, Theresa (eds.): Proceedings of the Workshop on Challenges in the Management of Large Corpora and Big Data and Natural Language Processing (CMLC-5+BigNLP) 2017 including the papers from the Web-as-Corpus (WAC–XI) guest section. Birmingham, 24 July 2017. Publisher Institut für Deutsche Sprache, Mannheim, pp. 15–19, 2017.
12. MORUZ A., SCUTELNICU A. L., *An Automatic System for Improving Boilerplate Removal for Romanian Texts*, in Mihaela Colhon, Adrian Iftene, Verginica Barbu Mititelu, Dan Cristea, Dan Tufiş (eds.) Proceedings of the 10th International Conference "Linguistic Resources And Tools For Processing The Romanian Language", Craiova, 17-18 September 2014, "Alexandru Ioan Cuza" University Publishing House, ISSN 1843-911X, pages 163–170.

13. SINCLAIR J., *Developing Linguistic Corpora: a Guide to Good Practice.* Corpus and Text – Basic Principles, Tuscan Word Centre, 2004 https://ota.ox.ac.uk/documents/creating/dlc/index.htm
14. TUFIŞ D., CRISTEA D., STAMOU S., *BalkaNet: Aims, Methods, Results and Perspectives. A General Overview.* In Romanian Journal of Information Science and Technology, Romanian Academy, Bucharest, Romania, Dan Tufis (ed.) Special Issue on BalkaNet, July, 7(1–2), ISSN 1453-8245, pages 9–43, 2004.
15. TUFIŞ D., BARBU MITITELU V., IRIMIA E., DUMITRESCU S. D., BOROŞ, TEODORESCU H.N., CRISTEA D., SCUTELNICU A., BOLEA C., MORUZ A., PISTOL L., *CoRoLa Starts Blooming – An Update on the Reference Corpus of Contemporary Romanian Language*, In Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora (CMLC-3), Lancaster, United Kingdom, 2015.
16. ⁂http://www.loc.gov/standards/mods/
17. ⁂http://www.dcc.ac.uk/resources/metadata-standards/darwin-core
18. ⁂https://wiki.earthdata.nasa.gov/display/NASAISO/Addressing+NASA+Metadata+Requirements+with+ISO+Standards
19. ⁂https://www.clarin.eu/content/component-metadata
20. ⁂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).