THE METHODOLOGY OF BUILDING COROLA

DANIELA GÎFU[1] , ALEX MORUZ[2], CECILIA BOLEA[3],
ANCA BIBIRI[4], MARIA MITROFAN[5]

Abstract. We briefly describe in this paper the process of building CoRoLa – *The Reference Corpus for Contemporary Romanian Language*. The process of its creation included legal acquisition of primary documents, cleaning of textual and speech data, and semi-automatic completion of metadata. The paper also describes CoDaP – the web platform specially built to support these operations in an interactive and collaborative manner and the crawling methodology adopted and sketches some thoughts for further corpus development.

Keywords: Primary documents, IPR, CoRoLa, metadata, boilerplate process, CoDaP platform.

1. INTRODUCTION

In the process of building Corola, The Reference Corpus for Contemporary Romanian Language (see Barbu Mititelu *et al*. 2014), two institutes of the Romanian Academy (RA) – the Institute of Computer Science, Iaşi, and the "Mihai Drăgănescu", Research Institute for Artificial Intelligence, Bucharest, have been primarily involved. They were accompanied in this process by many volunteers and contributors. In this paper, we mainly present the work done in the Iaşi institute, which comprised legal acquisition of primary documents, cleaning of textual and speech data and semi-automatic completion of metadata. In order to accomplish this process, we have built CoDaP (Cristea *et al.* 2017), a web-platform specially designed to support these operations in an interactive and collaborative manner. Any corpus building enterprise aims to satisfy investigation queries of almost any language patterns – of phonology, the lexical, the lexico-grammatical, morphological, syntactical, and the discourse level (Armstrong and Ferguson 2010). The paper of Cristea *et al*. (in this volume) shows that the corpus of the Romanian language

[1] Institute of Computer Science, Romanian Academy, Iaşi Branch, Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iaşi, Cognos Business Consulting S.R.L., Bucharest, daniela.gifu@info.uaic.ro
[2] Faculty of Computer Science, "Alexandru Ioan Cuza" University of Iaşi, mmoruz@info.uaic.ro
[3] Institute of Computer Science, Romanian Academy, Iaşi Branch, cecilia.bolea@iit.academiaromana-is.ro
[4] Social Sciences and Humanities Research Department, Institute for Interdisciplinary Research, "Alexandru Ioan Cuza" University of Iaşi, anca.bibiri@gmail.com
[5] Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy maria@racai.ro

RRL, **LXIV**, *3*, p. 241–253, Bucureşti, 2019

BDD-A30401 © 2019 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 19:21:48 UTC)

succeeded to accomplish this aim to a large extent, but some features are still to be realised in the near future.

Two dissemination meetings marked the official beginning and the end of the CoRoLa project, both held in the same Aula of the Romanian Academy Library in Bucharest, on the 3rd of February 2014 and on the 14th of December 2017. During these almost 4 years, the two institutes of the Romanian Academy organised the process, designed and implemented the processing technology, negotiated with owners of written and spoken Romanian documents, put up a vast network of volunteers to clean the data and fill in metadata information, established and pursued the collaboration with the Leibniz-Institute of the German Language in Mannheim (IDS) for using a common access infrastructure.

The paper is structured as follows: Section 2 presents a brief review of the main linguistic corpora across the world comprising at least 100 million words; Section 3 refers to data acquisition, which respected the ethics of data protection from various IPR holders we have concluded agreements with. Section 4 briefly describes preliminary data processing, while section 5 presents an integrated corpus development technology, including the CoDaP platform, developed in the Institute of Computer Science of Iași. In the last section, we draw some conclusions and mention future directions for all those interested in the processing of the Romanian language.

2. INTERNATIONAL LEVEL OF CORPUS CREATION

Corpora represent the most important category of linguistic electronic resources. When they are properly structured, they become a significant tool for the study of languages in their written and spoken forms. Computer corpora are perfect benchmarks for observing the language evolution (Sinclair 1991, 1992) and can be very useful for researchers, students and for the general public.

Corpora are generally designed for particular purposes and are often assembled to be representative of some language or text type (Leech 1992). A corpus should include a representative selection of texts that can be described by several features (Dash 2005). The corpus should be (1) compatible to the access initiated by man and computer; (2) operational in research and application; (3) representative for a language or a variety of it; (4) possible of being processed by both man and machine; (5) unlimited, concerning the amount of data; (6) systematic, both in form and representation.

Our corpus refers to the period after 1945 and is designed to include both written and spoken forms of the language. It intends to be a mirror of the cultivated Romanian language, large enough to include, along with significant contexts, the greatest part of its vocabulary. CoRoLa is designed to be useful to different categories of users: researchers working on Romanian language, computational linguistics and learners of Romanian as a foreign language, teachers and students (Barbu Mititelu *et al*. 2018).

To complete this project, we identified in the literature of linguistic corpora the following models, which include at least 100 million words: the BNC (*British National Corpus*)[6], the Mannheim German National Corpus DeReKo[7], the COCA (*Corpus of*

[6] http://www.natcorp.ox.ac.uk/.

BDD-A30401 © 2019 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 19:21:48 UTC)

Contemporary American English)[8], the ANC (*American National Corpus*)[9], the Russian National Corpus[10], the NKJP (*Narodowy Korpus Języka Polskiego*)[11], the CNC (*Czech National Corpus*)[12], the HNC (*Hungarian National Corpus*)[13], the CRPC (*Reference Corpus of Contemporary Portuguese*)[14], the BulNC (*Bulgarian National Corpus*)[15], the HNK (*Hrvatski Nacionalni Korpus*)[16].

The corpora mentioned above have some similarities and differences regarding the annotation process: the National Corpus of Polish (NKJP) (Przepiórkowski *et al*. 2010) is annotated at word-level segmentation, sentence segmentation, word sense disambiguation and morpho-syntax; the American National Corpus (Ide and Mcleod 2001) is marked for parts of speech, paragraphs and sentence boundaries, and CoRoLa for parts of speech, tokens, sentences. The search tools are based on the Poliqarp language (Janus and Przepiórkowski 2007 a, b) in NJKP, and this open-source corpus indexer and search engine is also used in CoRoLa.

3. DATA ACQUISITION

In the process of building CoRoLa, much attention was given to IPR (*Intellectual Property Rights*). The law prohibits detaining any collection of language data (oral or written) without the written accept of publishing houses, editorial offices, media channels, individual authors, whoever might be their legal owner or author. This preoccupation has encumbered a great deal of effort from our side[17]. The two CoRoLa teams have signed agreement contracts with representatives of important Romanian publishing houses (17) and editorial offices (16), with bloggers (16), individual authors (40), Radio and TV channels (4), etc. Metadata completion and text cleaning activities have been amplified by attracting volunteers[18].

[7] http://www.ids-mannheim.de/kl/projekte/korpora/archiv.html.

[8] https://www.english-corpora.org/coca/

[9] http://www.anc.org/

[10] http://ruscorpora.ru/

[11] http://nkjp.pl/index.php?page=0&lang=1

[12] http://www.czech-language.cz/korpus/korpus.html

[13] http://corpus.nytud.hu/mnsz/index_eng.html

[14] http://www.clul.ul.pt/en/resources/183-reference-corpus-of-contemporary-portuguese-crpc

[15] http://metashare.ibl.bas.bg/repository/browse/bulgarian-national-corpus/817c127064aa11e281b65cf3fcb88b705d83aefb9d21409dbaf029c8dddcce00/

[16] http://www.hnk.ffzg.hr/cnc.htm

[17] In order to convince the legal owners of language data to donate their texts and speech recordings, awareness seminars have been organized in Bucharest – by our colleagues at RACAI, and in Iaşi, at ICS, by ourselves. We have explained the reasons for having a big corpus and demoed examples of running corpora. This endeavour produced unexpected positive results; in sum, more than 20 entities have declared their willingness to offer language data to be included in CoRoLa.

[18] Usually students from the Faculties of Letters and of Computer Science of the "Alexandru Ioan Cuza" University of Iaşi and from the Faculty of Bioengineering of the "Gr. T. Popa" University of Medicine and Pharmacy of Iaşi. This happened with the help of the "Alexandru Ioan Cuza" University Alumni Foundation, as a host organization of volunteer activities within the CoRoLa Project.

BDD-A30401 © 2019 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 19:21:48 UTC)

4. PRELIMINARY DATA PROCESSING

To include any new document in the corpus, original versions of text needed to be cleaned and processed, so that only the parts relevant for indexing were kept. This meant the following cleaning actions (Cristea *et al*. 2017): removal of book titles, authors, ISSN, publishers, etc. (everything from the front page of a book), of headers, footers (i.e. book title, chapter, author, page number), of contents and captions of tables and figures, of any span of a text written in an international language other than Romanian (abstract, keywords, translated poems, etc.), of table of contents, bibliographies and reference lists, of citations, footnotes, endnotes, and formulas.

4.1. Cleaning the texts

Textual data had to be cleaned to make the data compatible with further processing phases (Tufiş 1999; Barbu Mititelu *et al*. 2017) carried on over an XML format of the raw text (sentence segmentation, tokenization, lemmatization, POS-tagging, and, in the future, syntactic, semantic or discourse level annotation). Any garbage in the textual data inherited from page layouts artefacts (as are, for instance, non-standard character codes or inner-words hyphens) would introduce errors in the process of XML annotation.

Generally, the process of cleaning primary textual data can be seen as made up of two steps: an initial cleaning (of the raw text) and a more elaborate cleaning (which makes use of text processing technology over the XML marked correspondent of the original text file) (Tufiş *et al*. 2019, in this volume). We will refer in this paper only to the first level.

A significant amount of the primary documents received from owners were PDF files. Extracting useful information from this format is a lot more difficult than extracting it from DOC files. This is particularly true with respect to the raw text, but also regarding the automatic identification of certain parts of metadata (see the next subsection).

We have used Apache PDFBox[19] to extract text from PDF files. This software package extracts the textual information ignoring graphical information such as images, figures and tables, but does not differentiate between the main text content of a file and its headers, footnotes, page numbers, etc. So, with the help of the CoDaP frontend, a number of document elements usually present on the document guard page (title of the book, author, ISSN, publishing house), on headers and footers (book title, chapter, author, page number), together with tables of contents, figures and table legends, sentences written in other languages than Romanian, references, footnotes, endnotes, formulas and appendices were manually removed. The following elements remained in the text: autobiography or biography of the author, chronology, introduction, abstract, preface, citations, citations in bibliographic references in the formats: (author/authors, year) or [order of the bibliographic list], and full explicit references. Also, some bullet characters, like "♦, •, □", not also "-", were replaced with "*" (Bibiri *et al.* 2015). The cleaning process in the project development was a great deal simplified as compared to the one described by Bibiri *et al.* (2015). If in the first phases of the elaboration of the project, we have used XML tags to mark citations, footnotes, formulas, etc., the difficulty of introducing these marks in the textual output and their apparent incompatibility with the displaying KorAP frontend

[19] From Apache Foundation https://www.apache.org/, a freely available tool.

BDD-A30401 © 2019 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 19:21:48 UTC)

(Cristea *et al*. 2019, in this volume) made us give them up and mainly eliminate the sentences where such elements would occur.

Note that in a PDF file each line ends with a newline character (\t), which makes each separate line be considered a new paragraph. We simply eliminated the newlines, which caused the information regarding paragraphs to be also dropped. Because in the text extracts presented by the corpus frontend, paragraphs are not being displayed, their identification is not obligatory, and we had no reason to pay special attention to this aspect.

Much more important, however, was to correctly identify and remove inner words hyphens occurring for alignment reasons at the end of lines, yet not those that separate auxiliaries or clitics from verb forms. Moruz and Scutelnicu (2014) describe a set of heuristics to achieve this distinction.

Although the extraction process provides a usable text, the quality of the extracted text is dependent on the encoding of the original file. Problems occur especially with the encoding of Romanian diacritics and of special symbols, like bullets. For instance, in certain PDF files, diacritics are not represented in UTF-8 format but are instead replaced with ASCII characters painted with specific fonts to represent the correct glyph. Because of this, in the extracted text, diacritics will be represented as non-literal ASCII symbols. Here is an example of a text fragment displaying incorrect diacritics:

> *Deci ...două femei fără bărbaþi, speriate de viitorul sumbru care se prevedea la orizontul roºu ºi trei copii (ba chiar patru, dacã o punem la socotealã ºi pe veriºoara mea Rodica Popescu) puºi numai pe joacã ºi nãz- bâtii, în ciuda sãrãciei de care nu eram conºtienþi - cam acesta era mediul în care trãiam ºi creºteam mari zi cu zi.* (EN after correction: *So ... two women without men, scared of the gloomy future that was predicted on the red horizon, and three children (or even four, if we count my cousin Rodica Popescu), only predisposed to play and larks in spite of the poverty of which we were not aware of - that was the environment in which we lived and grew up day by day.*)

A unique translation table does not work since different documents could use different diacritics encryptions. However, replacements are unique within a document, in the sense that once an equivalence is found (association between a glyph and a symbol), it will be used throughout the entire text. Based on this observation, an algorithm for character standardisation has been written and tested, showing less than 3% error rate at the level of recognised words. The algorithm uses a glossary against which the codes of replaced words are looked for.

4.2. Filling-in Metadata

The manual fill-in of document metadata is a time-consuming process; using an automatic alternative is most of the time not feasible. In some cases, however, and for some of the metadata fields, automatic assistance or even extraction was possible. Since many of the primary documents we have used have been provided by publishing houses as electronic formats of printed books, they also had an attached CIP-description (*Catalogarea Înaintea Publicării*, *Cataloguing Before Publishing*), a mandatory field for all published books.

A typical CIP-description for one of the documents included in CoRoLa is given below:

BDD-A30401 © 2019 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 19:21:48 UTC)

DIACONU-POPOVICI, RĂZVAN. Impactul jocurilor on-line asupra tinerilor / Răzvan Diaconu-Popovici. - Iaşi : Ştef, 2015.
ISBN 978-606-575-461-4
004.738.5
159.922.8

A CIP-description is very helpful in automatically determining part of the metadata. As seen in the example above, the metadata extracted are (1) the author(s), in the case of authored books always given in capital letters; (2) the title, followed by "/" and the author name(s) given in a normal script (for edited volumes, this appears as the first field); (3) other information such as translator, preface, etc. This always ends with a ":" character followed by the name of the publishing house and the year of publishing; (4) the ISBN; (5) the Universal Decimal Classification (UDC) of the document.

We have built a parser (IIT 2015) to extract this information from the CIP-description. The UDC was particularly useful, as it helped to classify the document according to a standard bibliographic and library classification system that covers all fields of human knowledge. This code was automatically mapped to CoRoLa domain and subdomain categories. For that, we have put up a correspondence table, which was first automatically filled in using UDC descriptions and then manually linked to the Wikipedia domain information, representing the basis of the above-mentioned classification. The alignments are as follows:

Domain (*Arts&Culture*) **– Subdomains/UDC codes** (*Literature/82*, *Art History/930.85*, *Folklore/39*, *Film/791*, *Architecture*/72, *Sculpture/730*, *Painting & Drawing/74/75*, *Theatre/792*, *82-2*, *Dance/793.3*).

Domain (*Society*) **– Subdomains/UDC codes** (*Politics/32*, *Law/34*, *Administration/ 351/354*, *Economy/33*, *Army*/*355/359*, *Health/613/614*, *Sport/796/799*, *Family/316.36*, 82-2, *Education/37*, *Social Movements/323.1*, *Tourism/338.48, 796.5*, *Religion/2*, *Entertainment/79*).

Domain (*Nature*) **– Subdomains/UDC codes** (*Environment/502/504*, *Universe/52*, *Natural Resources/55*).

Domain (*Exact/Formal Science*) **– Subdomains/UDC codes** (*Mathematics/51*, *Informatics/004*, *Logics/16*, *Film/791*, *Standards/006.3/.8*).

Domain (*Applied Sciences*) **– Subdomains/UDC codes** (*Medicine/61*, *Archeology/902*, *Engineering/62*, *Architecture*/72, *Techniques/Technology/62*, *Aeronautics/629.7*, *Agronomy/ 631/635*, *Metrology/006.91, Criminalistics/343.9*, *Constructions/69*, *Military Science/ 355/359*, *Pharmacology/615*, *Ecology/663.2)*.

Domain (*Social Sciences*) **– Subdomains/UDC codes** (*Geography/91*, *Economy/33*, *History/93/94*, *Psychology/159.9*, *Sociology*/*316*, *Ethnology/39*, *Anthropology/572*, *Religious Studies and Theology/2-1*, *Juridical Sciences/340, Linguistics/81*, *Political Sciences/32*, *Philosophy/10/14, 17*, *Philology/80*).

Domain (*Natural Science*) **– Subdomains/UDC codes** (*Biology/57*, *Physics/53*, *Astronomy/52*, *Chemistry/54*).

CoRoLa's present level of technology does not allow a completely automatic process of acquiring the metadata of the articles in a document. Therefore, manual processing was used for inputting metadata fields which cannot be automatically discovered in the CIP data. Equally manually performed was the splitting of articles in magazines and edited volumes, facilitated by an interface, as will be seen below. It is also worth noting

BDD-A30401 © 2019 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 19:21:48 UTC)

that, since documents containing collections of articles and documents made up of a single piece are being treated differently, these have been manually selected, prior to processing, into two distinct categories.

5. TOWARDS AN INTEGRATED CORPUS DEVELOPMENT TECHNOLOGY

Metadata, as standardized information blocks, represent general and specific information about authentic/primary texts that are processed with specific tools. Metadata can be used in the process of retrieval, allowing the user to restrict the search query. Metadata are "the key to ensuring that resources will survive and continue to be accessible in the future"[20].

For each processed file, a CMDI[21] metadata format was associated, following the one proposed in CLARIN (*Common Language Resources and Technology Infrastructure*)[22] and further used on the META-SHARE (*Multilingual Europe Technology Alliance*) platform[23]. 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[24].

The corpus covers 71 sub-domains grouped into 4 domains, described in Bibiri *et al*. (2015) and Tufiș *et al*. (2016). In Figure 1 we present 2 different types of common metadata, according to the text category (novel and edited/coordinated volume).

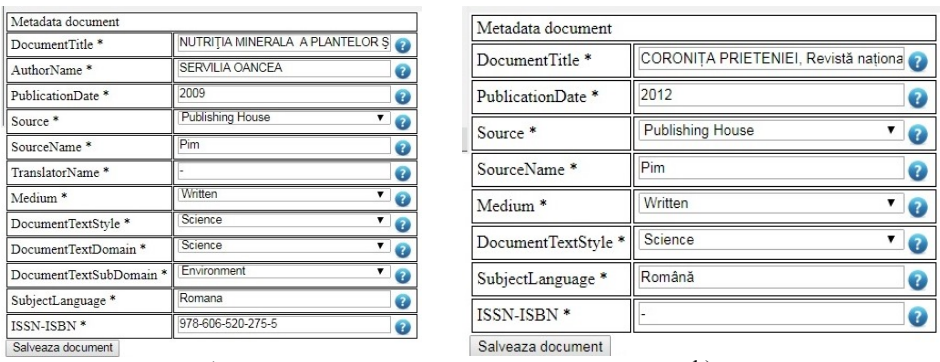


Fig. 1. Example of metadata for a) a novel, b) an editing book.

[20] http://www.niso.org/publications/press/UnderstandingMetadata.pdf
[21] http://www.niso.org/publications/press/UnderstandingMetadata.pdf
[22] https://www.clarin.eu/content/component-metadata
[23] http://www.meta-net.eu/meta-share/index_html
[24] 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).

BDD-A30401 © 2019 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 19:21:48 UTC)

5.1. The CoDaP Platform

As already shown, most of the files received by the owners of textual data are in the PDF format. Their conversion into editable TXT files is done via a web application – the CoDaP Platform (***Co****RoLa* ***Da****ta Cleaning and Metadata* ***P****latform*)[25], developed in the Institute of Computer Science of Iași (Moruz and Scutelnicu 2014). The Portal offers information on the activity of acquisition of CoRoLa (description of the project, groups involved, list of text and audio records providers, working meetings, an annotation manual for text files and audio recordings, etc.) and allows working access to authenticated users for the following types of activities:

- conversion of input PDF files into TXT ones;
- acquisition and editing of specific metadata for all document types: novels, magazines, edited books;
- processing of a primary document to extract its clean raw content text.

The user finds the interface under the *PLATFORMĂ>Aplicația grupului de procesare a textului* (*PLATFORM>Text Processing Group Application*) menu. After the authentication with username/password, one can start working by pushing the *Incepe lucrul* (*Start working*) button. When exiting, it is recommended that working sessions be explicitly closed, by clicking the *Delogare* (*Unsubscribe*) button, finally logging out from the web application.

The user receives a working file, which has been randomly selected from the database of PDF files received from the providers, for which no metadata has been entered yet. Once the document is allocated to a user, it will remain linked to his account until it is finished.

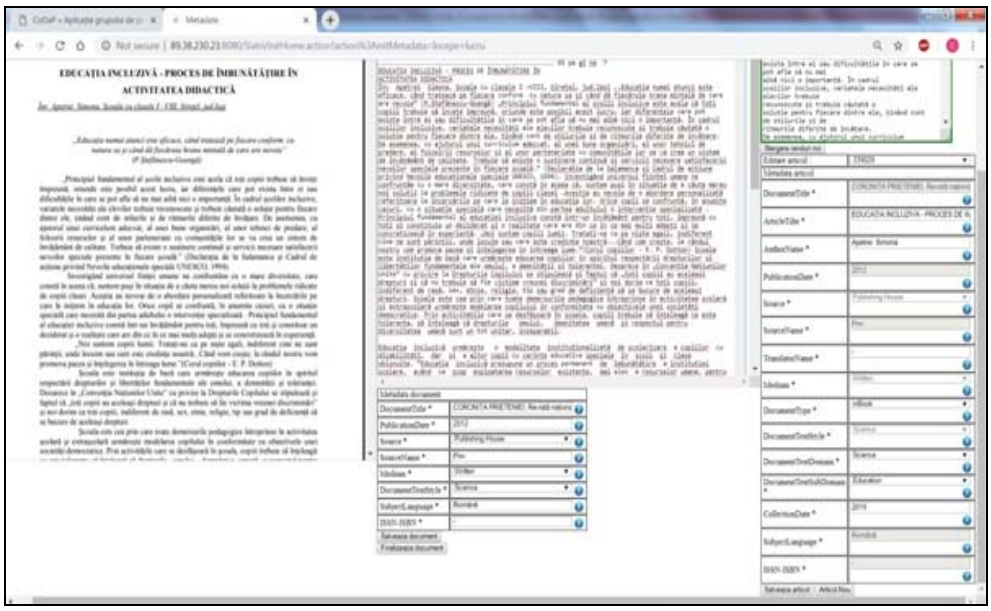


Fig. 2. CoDaP working window.

[25] http://89.38.230.23/

BDD-A30401 © 2019 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 19:21:48 UTC)

As shown in Figure 2, the working windows is split in three sections: the left section displays the original image of the PDF file, the middle window shows the whole text resulted from the boilerplate removal process and the metadata of a general/common nature (here there are no more figures and tables, and diacritics are converted to the UTF-8 codes) and the right window displays the text of a specific article in a book including a collection of articles, as is an edited book, or a magazine and its specific metadata.

General/common metadata refers to that part of the metadata fields which are common to the whole document (as, for instance, author of a book, editor of a multi-article volume, year of publication, publishing house, etc.). The fields of a general nature (referring to the document as a whole) are listed in Fig. 3 in the corresponding XML format.

```
<root>
  <Metadata>
       <DocumentTitle>Nutriţia minerală a plantelor şi
   implicaţiile ecotoxicologice</DocumentTitle>
       <AuthorName>Servilia Oancea</AuthorName>
       <PublicationDate>2009</PublicationDate>
       <Source>Publishing House</Source>
       <SourceName>Pim</SourceName>
       <TranslatorName>-</TranslatorName>
       <Medium>Written</Medium>
       <DocumentType>Book</DocumentType>
       <DocumentTextStyle>Science</DocumentTextStyle>
       <DocumentTextDomain>Science</DocumentTextDomain>
       <DocumentTextSubDomain>Environment</DocumentTextSubDomain>
       <Collection Date>2019</Collection Date>
       <SubjectLanguage>Română</SubjectLanguage>
       <ISSN-ISBN>978-606-520-275-5</ISSN-ISBN>
  </Metadata>
</root>
```

Fig. 3. Example of the metadata XML schema of a CoRoLa document

The text corresponding to a specific article from a magazine or an edited book must be copied by the user from the middle to the right window, in direct pair with its specific metadata filled in in the right window, below that text.

When the user pushes the *Salvează articol* (*Save article*) button, two files are saved (both getting the same name – a number – but different extensions), corresponding to the information on the right side of the CoDaP working window: the XML metadata file and the article itself, in .txt format. The user should iterate editing/correcting operations *Editează articol* (*Edit article*) – *Salvează articol* for each article of the original edited book or magazine. The article can be edited even after it has been saved, as long as the work on the entire volume was not declared yet completed. In the case of a novel, only one article (the novel's content) should be inserted, edited and saved in the right window.

BDD-A30401 © 2019 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 19:21:48 UTC)

After all the items in the original document have been entered, the user proceeds to the *Finalizează document* (*Finalize document*). After this action, it is impossible to return to a previously edited file. The work continues by opening another PDF document and following the steps mentioned above. It is worth mentioning that the Platform supports interactive sessions of up to 30 users working simultaneously.

5.2 Tools created for CoRoLa – Crawlers

The role of the Web for the corpus construction is becoming increasingly significant, but to automatically extract texts from specific web pages web crawlers are needed. A Web crawler is a program or automated script that browses the WWW in a methodical, automated manner and performs content extraction. Web crawlers are especially suitable for specialised corpora construction.

In order to automatically obtain textual resources for CoRoLa corpus, custom web crawlers were created for different web pages. A common practice for web crawling is using Python programming language and Beautiful Soup, a free open-source library written in Python for parsing HTML pages. Beautiful Soup provides methods and Python idioms for navigating, searching and modifying a parse tree. Currently available as Beautiful Soup 4 and compatible with both Python 2.7 and Python 3, Beautiful Soup creates a parse tree from parsed HTML and XML documents (including documents with non-closed tags or other malformed markups). One of the crawlers created for CoRoLa has two main functionalities: first, the portal is searched for all relevant documents. This is achieved by accessing the portal, generating a search request for all articles (pages containing relevant text) and then parsing the resulted http pages by fetching all links. Then, an attempt is made to find the base http page where the text resides. The crawler at this stage uses the BeautifulSoup library to get all the <*a*> tags and then uses the get method to extract the exact links. A list of URLs is generated as a result of this stage. In the second stage, based on the list of URLs generated in the first step, the crawler accesses each of these pages and attempts to find the text between the body tags. This is done by iterating through each of the relevant tags (e.g. span) and building a concatenation of the text found in these tags. This is achieved by considering the fact that various parts of text can be found throughout the downstream hierarchy of children tags.

A drawback to this approach is that texts fetched by crawling are not IPR secured, therefore we have adopted a rather cautious attitude, accessing only blog type pages and asking the authors' permission to use their texts.

6. CONCLUSION AND FUTURE WORK

The work to develop a corpus of a language cannot be considered closed at the end of any supporting project since language is in continuous evolution. As such, there is a permanent concern to update the existing corpus. Up to a point, this endeavour will still allow its future developers to name it "contemporary", but extended in time, the corpus will gain more and more features of a diachronic one, reflecting the evolution of language. The diachronic characteristic of the Romanian corpus interests to such a large extent the

BDD-A30401 © 2019 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 19:21:48 UTC)

linguists, that it should become a predominant preoccupation in the years to come, this meaning its temporal extension at both ends, in the future but equally in the past.

Currently, we are also investigating ways in which CoRoLa could be linked to other existing language resources for Romanian (lexical resources in particular). To this extent we have taken under consideration linking CoRoLa to the Romanian WordNet (Tufiş *et al*. 2015) and to eDTLR – the electronic version of the Romanian Thesaurus Dictionary (Cristea *et al*. 2007). By adopting methods inspired from Linguistic Linked Open Data (Chiarcos *et al.* 2013), supplementary thesaurus information can be retrieved to the interested user, by associating them to a CoRoLa ordinary search (Moruz *et al*. 2018). At the moment, the main disadvantage of the process is that it is very time-consuming. Our tests for linked retrievals performed on a section of CoRoLa containing only one hundred thousand documents showed that the complex extraction for a particular query is not feasible in real-time. In a further development we will pre-compute lexical correspondences for each document in CoRoLa, such that they are available beforehand.

ACKNOWLEDGMENTS

We address special and warm thanks to Ruxandra Cosma, from the Department of Germanic Languages, University of Bucharest, for her continuous support during the elaboration of the Corpus, and for being at the heart of the launch of the DRuKoLA project and the collaboration with the Leibniz-Institut für Deutsche Sprache (IDS), Mannheim. We thank our colleagues from the Leibniz-Institut für Deutsche Sprache, who, beneficiaries of huge experience in corpus creation, have given us constant, friendly and extremely specialised technical support in all phases of corpus creation. We thank Dan Cristea for ideas and support while leading our team. The Iaşi team also included Daniela Gîfu, Cecilia Bolea and Mihaela Onofrei – responsible for acquisition of primary data, solving of IPR issues, managing the work of volunteers to acquire metadata and eliminate errors in the textual data, themselves work intensively working with the CoDaP interface, Alex Moruz – who has written and tested most of the pdf2txt convertor and cleaning software, Andrei Scutelnicu – the developer of the web technology (the CoDaP frontend & backend) and responsible for the installation and administration of the Iaşi server, and Laura Pistol – curator of a great part of the Romanian speech data. We thank our Romanian data providers, owners of digital written and spoken data, who, understanding the importance of having a Romanian corpus of language extract largely open to the public, generously offered their data to the benefit of the Corpus. We thank all our volunteers involved in cleaning textual data, in the acquisition of the corresponding metadata and in doing voice recordings (students from the "Alexandru Ioan Cuza" University of Iaşi, the University "Politehnica" of Bucharest, the University of Bucharest, the Technical University of Cluj-Napoca and the University of Craiova). We are grateful to the Alexander von Humboldt-Foundation for supporting the DRuKoLa project, thus sponsoring the acquisition of the computer infrastructure on which CoRoLa resides (two servers make up the distributed architecture, installed and running in Iaşi, at ARFI-IIT, and in Bucharest, at RACAI), a number of international scientific visits in the triangle Bucharest-Iaşi-Mannheim, as well as two project workshops.

BDD-A30401 © 2019 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 19:21:48 UTC)

The work of building a corpus of the Romanian language is a never-ending process because it should be kept updated with texts newly published, while processing technologies should be maintained and renewed. In this respect, the platform of text and speech processing, that is now under development due to a grant of the Romanian Ministry of Research and Innovation, CCCDI – UEFISCDI, project number PN-III-P1-1.2-PCCDI-2017-0818 / 73PCCDI (ReTeRom) within PNCDI III, has had a significant effect on the modernisation of the CoRoLa technology. Also, the work done in the README project "Interactive and Innovative application for evaluating the readability of texts in Romanian Language and for improving users' writing styles", contract no. 114/15.09.2017, MySMIS 2014 code 119286, contributed to the selection process of textual primary data.

REFERENCES

Armstrong, E. M., A. Ferguson, 2010, "Language, meaning, context and functional communication", *Aphasiology*, 24, 4, 480–496.

Barbu Mititelu, V., E. Irimia, D. Tufiş, 2014, "CoRoLa – The Reference Corpus of Contemporary Romanian Language", in: N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (eds), *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC 2014), Reykjavik, Iceland, 1235–1236.

Barbu Mititelu, V., D. Cristea, R. Cosma, 2017, "Corpus of Contemporary Romanian. Architecture, Annotation Levels and Analysis Tools", in: H. Bogdan Oprea, A.-V. Grigore, R. Zafiu (eds), *Lingvistică românească, lingvistică romanică. Actele celui de-al XVI-lea Colocviu Internațional al Departamentului de Lingvistică,* Bucharest, November 25-26, 2016, București, Editura Universității din București, 13–20.

Barbu Mititelu, V., D. Tufiş, E. Irimia, 2018, "The Reference Corpus of the Contemporary Romanian Language (CoRoLa)", in: N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis, T. Tokunaga (eds.), *Proceedings of the Eleventh International Conference on Language Resources and Evaluation* (LREC 2018), Japan, 1178–1185.

Bibiri, A. D, C. Bolea, L. A. Scutelnicu, A. Moruz, D. Cristea, 2015, "Metadata of a Huge Corpus of Contemporary Romanian Data and Organization of the Work", in C. Bădică, Y. Manolopoulos (eds), *Proceedings of the 7th Balkan Conference in Informatics*, Craiova, Romania, September 2-4, 2015, ACM, New York.

Chiarcos, C., P. Cimiano, T. Declerck, J. Mc Crae, 2013, "Linguistic Linked Open Data (LLOD). Introduction and Overview", in C. Chiarcos, P. Cimiano, T. Declerck, J. P. McCrae (eds), *Proceedings of LDL 2013*, Pisa, Italy, I-XI.

Cristea, D., M. Răschip, C. Forăscu, G. Haja, C. Florescu, B. Aldea, E. Dănilă, 2007, "The Digital Form of the Thesaurus Dictionary of the Romanian Language", in: H. Teodorescu, C. Burileanu (eds), *Proceedings of SPeD-2007 (Speech Technology and Human - Computer Dialogue)*, May 10-12, 2007, Iași, Editura Academiei, 193–204.

Cristea, D., D. Gîfu, A. Moruz, M. Onofrei, L. Pistol, A. Scutelnicu, C. Bolea, 2017, "An Insight into the Corpus of Contemporary Romanian", *Memoirs of the Scientific Sections / Memoriile Secțiilor Științifice*, *Series IV, Tome XL*, Editura Academiei, 67–84.

Cristea, D., N. Diewald, G. Haja, C. Mărănduc, V. Barbu Mititelu, M. Onofrei, 2019, "How to Find a Shining Needle in the Haystack. Querying Corola: Solutions and Perspectives", in this volume.

Dash, N.S., 2005, Corpus Linguistics and Language Technology, New Delhi, Mittal Publications.

Ide, N., C. Macleod, 2001, "The American National Corpus: A Standardized Resource of American English", in: P. Rayson, A. Wilson, T. McEnery, A. Hardie and S. Khoja (eds), *Proceedings of Corpus Linguistics 2001*, Lancaster UK, 274–280.

BDD-A30401 © 2019 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 19:21:48 UTC)

Janus, D. and Przepiórkowski, A., 2007a, "Poliqarp1.0: Some technical aspects of a linguistic search engine for large corpora", in: J. Wali´nski, K. Kredens and S. Go´zd´z-Roszkowski (eds), *The Proceedings of Practical Applications in Language and Computers* PALC 2005, Frankfurt am Main, Peter Lang.

Janus, D., A. Przepiórkowski, 2007b, "Poliqarp: An open source corpus indexer and search engine with syntactic extensions", *Proceedings of the ACL 2007 Demo and Poster Sessions*, Prague, 85–88.

Leech, G., 1992, "Corpora and theories of linguistic performance", in: J. Svartvik (ed), *Directions in Corpus Linguistics*. *Proceedings of Nobel Symposium 82*, August 4–8, 1991, Berlin, New York; Mouton De Gruyter, 105–122.

Moruz, A., A. Scutelnicu, 2014, "An automatic system for improving boilerplate removal for Romanian texts", in: M. Colhon, A. Iftene, V. Barbu Mititelu, D. Cristea, D. Tufiş (eds), *Proceedings of the 10th International Conference Linguistic Resources and Tools for Processing the Romanian Language*, Iaşi, Editura Universităţii "Alexandru Ioan Cuza", 163–170.

Moruz, M. A., A. Scutelnicu, D. Cristea, 2018, "Interlinking and Extending Large Lexical Resources for Romanian", in: V. Păiş, D. Gîfu, D. Trandabăţ, D. Cristea, D. Tufiş (eds), *Proceedings of the 13th International Conference Linguistic Resources and Tools for Processing the Romanian Language,* November 22-23, 2018, Iasi, Romania, 125–132.

Przepiórkowski, A., R.L. Górski, M. Łaziński, P. Pęzik, 2010, "Recent Developments in the National Corpus of Polish", in: N. Calzolari, K. Choukri, B. Maegaard, J. Mariani, J. Odijk, S. Piperidis, M. Rosner, D. Tapias (eds), *Proceedings of the Seventh International Conference on Language Resources and Evaluation* (LREC 2010), May 2010, Valletta, Malta, 994–997.

Sinclair, J., 1991, *Corpus, concordance, collocation,* Oxford, Oxford University Press.

Sinclair, J., 1992, "The automatic analysis of corpora", in: J. Svartvik (ed.), *Directions in Corpus Linguistics, Proceedings of Nobel Symposium 82,* Berlin, Mouton de Gruyter, 379-397.

Tufiş, D., 1999, "Tiered Tagging and Combined Classifiers", in: F. Jelinek, E. Nöth (eds), *Text, Speech and Dialogue, Lecture Notes in Artificial Intelligence 1692*, Springer, 28–34.

Tufiş, D., V. Barbu Mititelu, E. Irimia, Ş. D. Dumitrescu, T. Boroş, H. N. Teodorescu, D., Cristea, A. Scutelnicu, C. Bolea, A. Moruz, L. Pistol, 2015, "CoRoLa Starts Blooming – An update on the Reference Corpus of Contemporary Romanian Language", in: P. Bański, H. Biber, E. Breiteneder, M. Kupietz, H. Lüngen, A. Witt (eds), *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora* (CMLC-3), Mannheim, Institut für Deutsche Sprache, 5–10.

Tufiş, D., V. Barbu Mititelu, E. Irimia, Ş. D. Dumitrescu, T. Boroş, 2016, "The IPR-cleared Corpus of Contemporary Written and Spoken Romanian Language", in: N. Calzolari, K. Choukri, T. Declerck, M. Grobelnik, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, S. Piperidis (eds), *Proceedings of the Tenth International Conference on Language Resources and Evaluation* (LREC 2016)*,* Portorož, Slovenia, 2516–2521.

Tufiş, D., V. Barbu Mititelu, M. Onofrei, M. Mitrofan, E. Irimia, N. Diewald, R. Ion, V. Păiş, 2019, "Little Strokes Fell Great Oaks. Creating CoRoLa, the Reference Corpus for Contemporary Romanian", in this volume.

BDD-A30401 © 2019 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 19:21:48 UTC)

BDD-A30401 © 2019 Editura Academiei
Provided by Diacronia.ro for IP 216.73.217.75 (2026-11-06 19:21:48 UTC)