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

Dan Tufiș, Verginica Barbu Mititelu, Elena Irimia, Ștefan Daniel Dumitrescu, Tiberiu Boroș

Research Institute for Artificial Intelligence "Mihai Drăgănescu" (RACAI), Romanian Academy, Bucharest

Horia Nicolai Teodorescu, Dan Cristea,

Andrei Scutelnicu, Cecilia Bolea,

Alex Moruz, Laura Pistol

Institute for Computer Science (ICS), Iași

Context

- the META-NET whitepapers series (Trandabăț et al., 2012), which documented the availability and use of language technology for 31 European languages, qualified Romanian as “fragmentary supported” by language technologies
- 2012: RACAI in Bucharest finalized the Romanian Balanced Corpus (**ROMBAC**) (Ion et. al, 2012) containing 44,117,360 tokens covering five domains (News, Medical, Legal, Biographic and Fiction).
- Since 2014: ICS in Iasi joined RACAI in the concern for creating a bigger corpus, in a priority project of the Romanian Academy: **The Reference Corpus of Contemporary Romanian Language.**

Objectives

- medium to large corpus (more than 500 million word forms)

- IPR cleared
  - ✓ political, legislative and administrative content is excepted by IPR law
  - ✓ Other types of content: either written accept from IPR owners or use tiny fragments (no more than 10,000 characters)

- representative for the language stage: all functional styles will be represented (scientific, official, publicistic and imaginative)

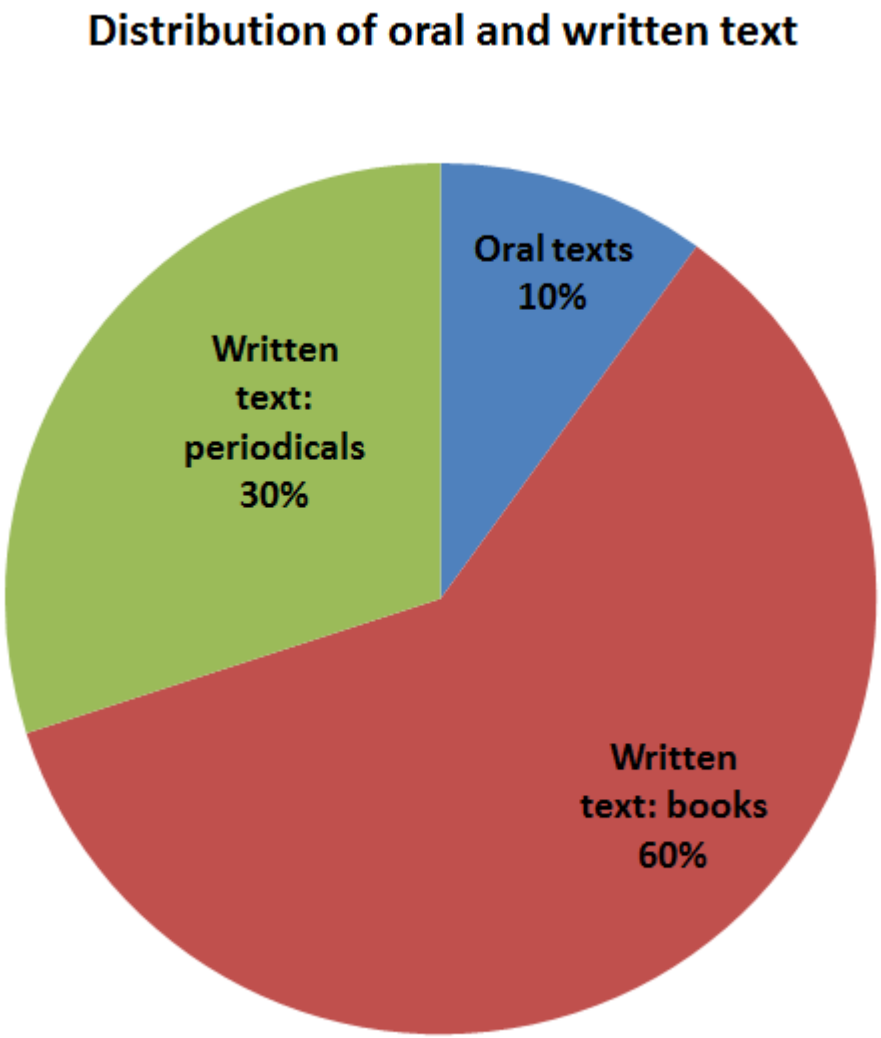
  Obs.: the colloquial style is not a major concern for us, but it will definitely be included, due to its use in imaginative writing;

- time span covered by the project: 1945-present;

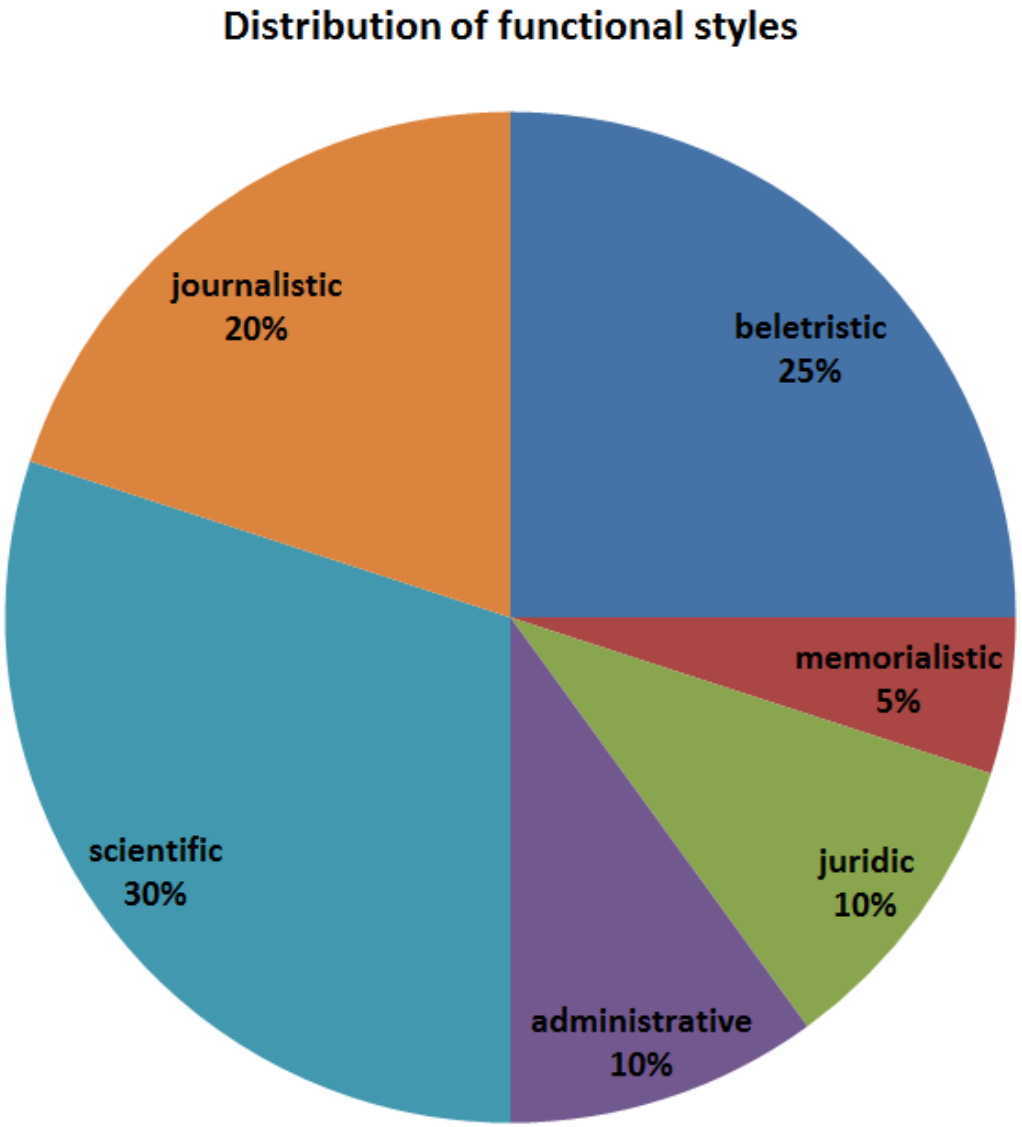
Objectives

- all textual data will be morpho-lexically processed (tokenized, POS-tagged and lemmatized)
- will include a syntactically annotated sub-corpus (treebank: 10,000 sentences, dependency grammar formalism)
- an oral component (300 hours of transcribed recorded speech)
- consider only texts written with correct diacritics (otherwise, the linguistic annotation will be highly incorrect)
- particular attention paid to data documentation, i.e. associating it with standardized metadata.

Foreseen structure



Foreseen structure



Data collection

Large amounts of data + Quality of the data + Copyright agreements

→ establish collaborations with publishing houses and editorial offices

Publishing Houses	Humanitas, Polirom, Romanian Academy Publishing House, Bucharest University Press, "Editura Economică", ADENIUM Publishing House, DOXOLOGIA Publishing House, the European Institute Publishing House, GAMA Publishing House, PIM Publishing House
Magazines/Newspapers	România literară, Muzica, Actualitatea muzicală, Destine literare, DCNEWS, PRESSONLINE.RO, the school magazine of Unirea National College from Focșani, SC INFOIASI SRL, Candela de Montreal
Bloggers	Simona Tache, Dragoș Bucurenci, Irina Șubredu and Teodora Forăscu
Writers	Corneliu Leu and Liviu Petcu
Broadcasting agencies	Rador (the press agency of Radio Romania) and Radio Iaşi (local broadcasting agency)

Agreements signed until March 2015

Data cleaning

Challenge in corpus creation: to have texts in a clean format, easy to process and annotate.

- collaborators dispatch a textual resource usually in unprotected pdf files, rarely in doc files → convert it into an adequate format for our pre-processing tools, UTF-8 encoded and saved as plain text documents.

- we automated a part of the process (Moruz and Scutelnicu, 2014):
  - ✓ the text is automatically retrieved from the pdf files,
  - ✓ paragraph limits are recovered,
  - ✓ column marking, newlines and hyphens at the end of the lines are erased

- a lot of manual work remains to be done:
  - ✓ separating articles from periodicals in different files,
  - ✓ removal of headers, footers, page numbers, figures, tables,
  - ✓ dealing with foot- or end-notes,
  - ✓ dealing with text fragments in foreign languages,
  - ✓ dealing with excerpts from other authors, etc.

Data accessibility

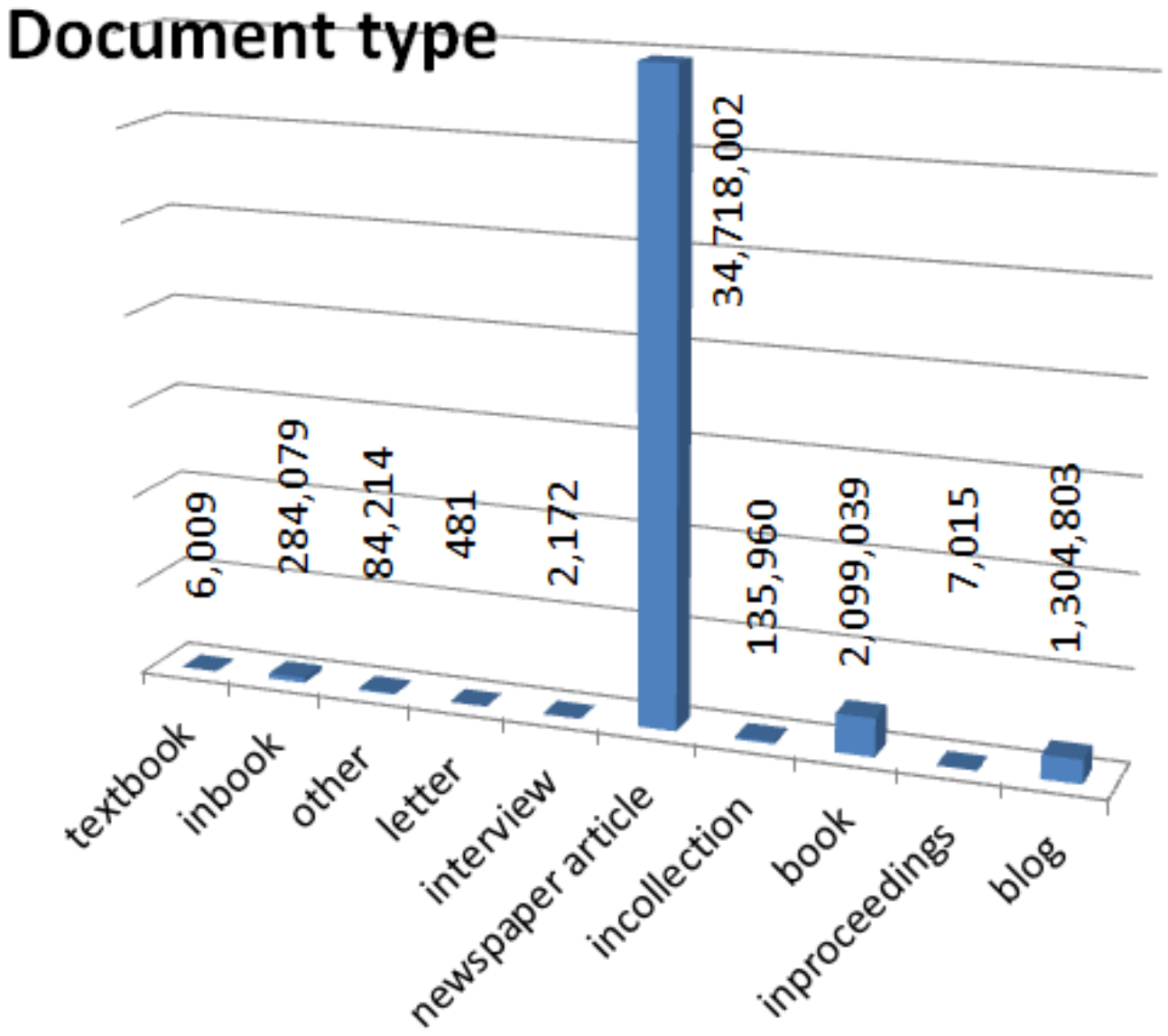
- Corpus indexing and searching done with **IMS Open Corpus Workbench** (**CWB**, http://cwb.sourceforge.net/): an open source medium, allows complex searching with multiple criteria and support for regular expressions:

  - ✓ to choose the (sub)corpus/(sub)corpora with which to work (choose from among the domains and subdomains, but also from the available authors)
  - ✓ to find out words frequencies in a (specified) (sub)corpus;
  - ✓ to search for a lemma or a word form;
  - ✓ to search for more words (either consequent or permitting intervening words);
  - ✓ to find words co-occurrences and collocations (within a window of a pre-established size);
  - ✓ to find lexicalization of specified morphological or/and syntactic structures;

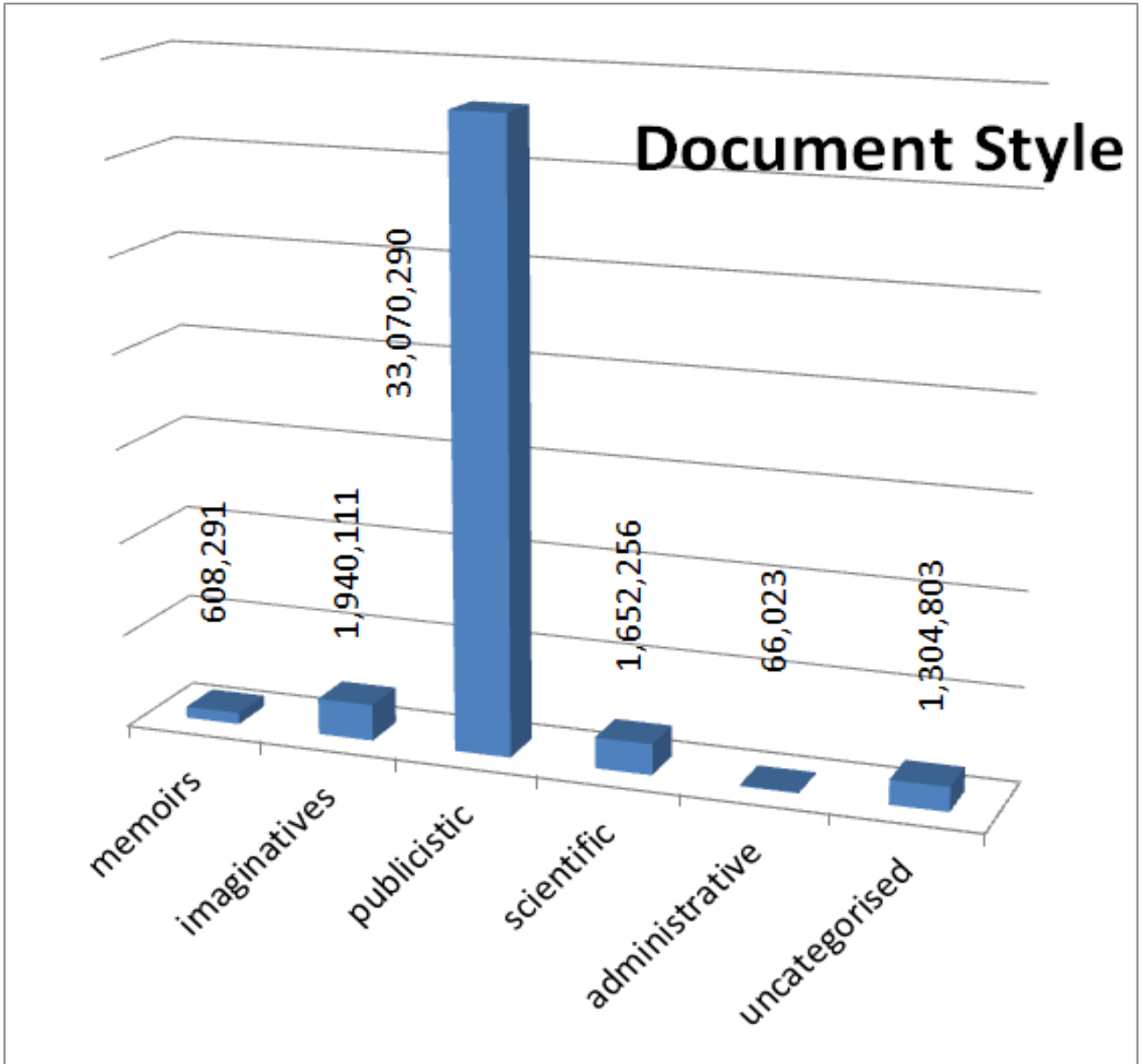
- CWB tested on the ROMBAC corpus and coupled with our processing chain, whose annotated output format is compatible with CWB;

- we plan to switch to the more powerful corpus management platform KorAP (Bański et al., 2014)

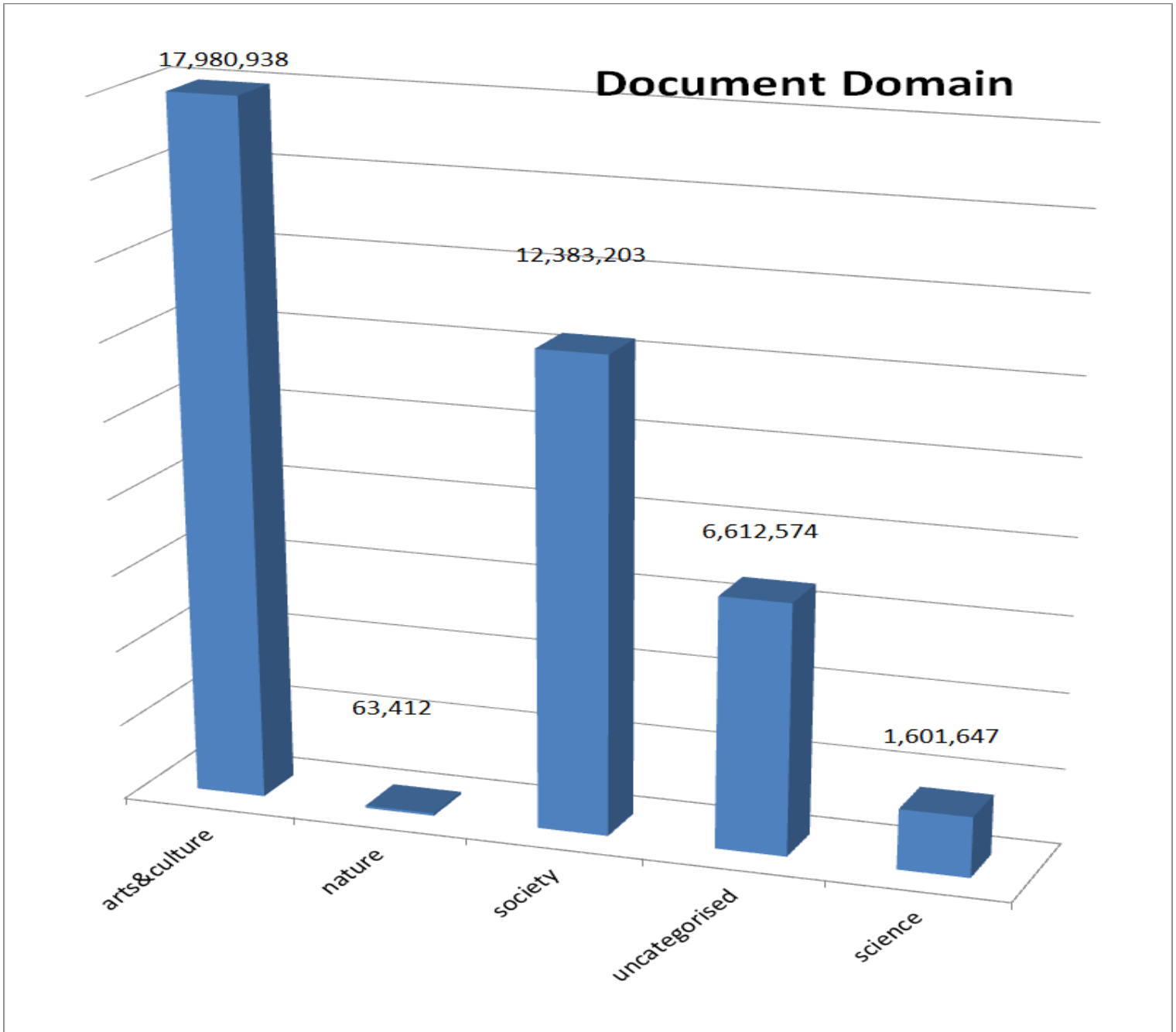
Current statistics: textual data



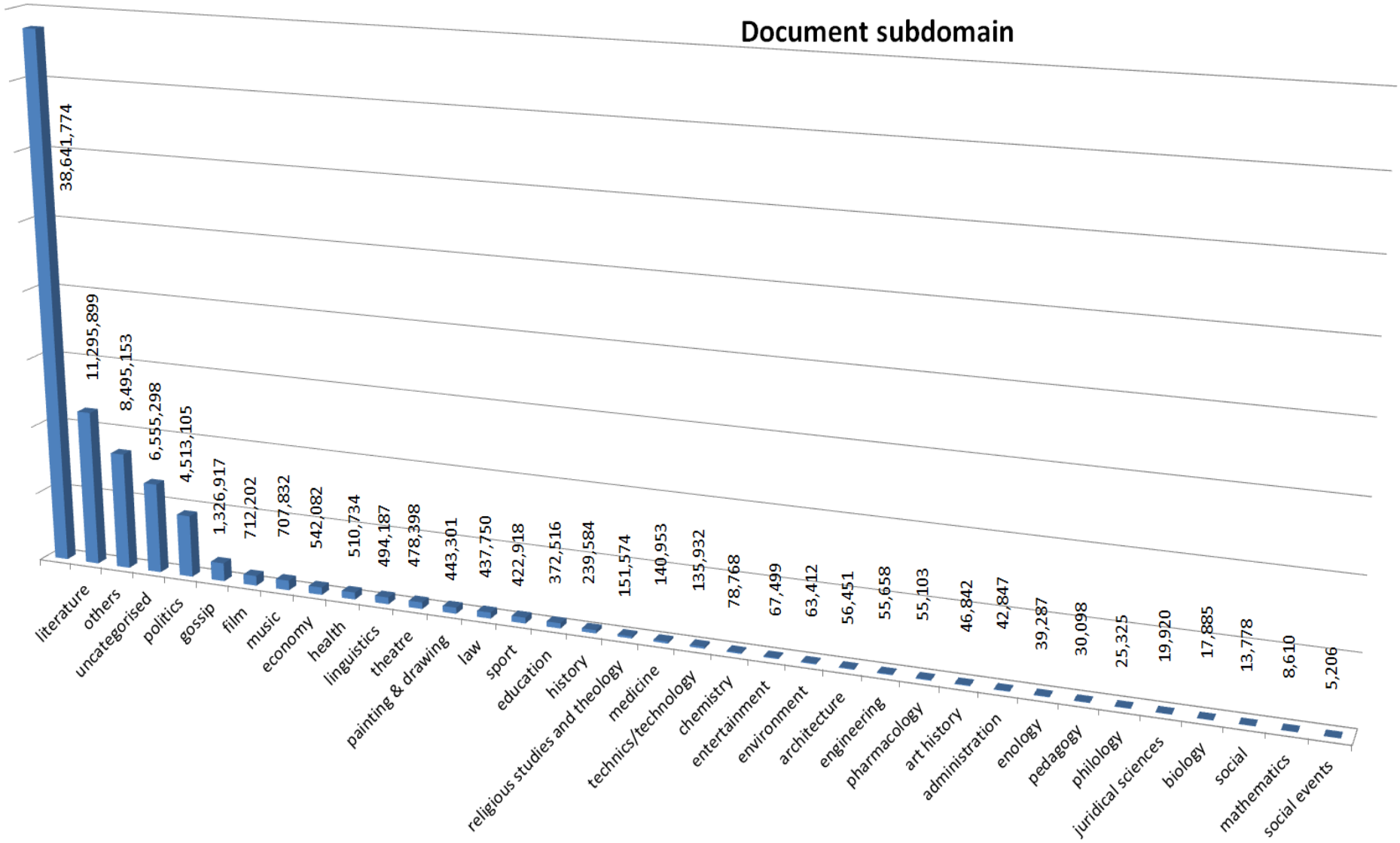
Current statistics: textual data



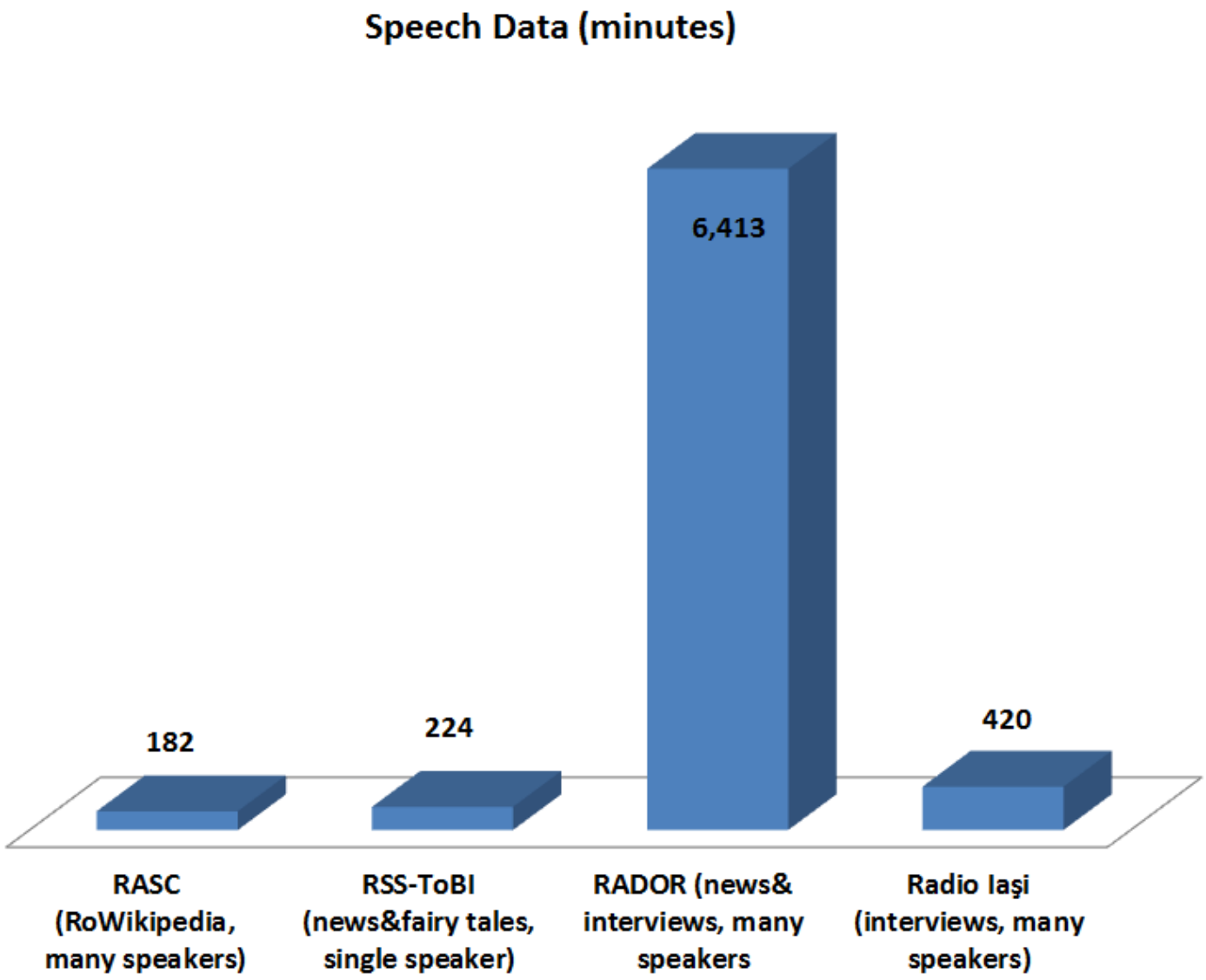
Current statistics: textual data



Current statistics: textual data



Current statistics: speech data



Metadata creation

- essential for the indexing of the corpus and facilitates the searching process for the end users
- **CMDI (Component MetaData Infrastructure)** initiated in CLARIN that proposes a component-based approach: the creator of metadata can combine several metadata components (sets of metadata elements) into a specific scheme, called "profile"; reuse of components and profiles: guiding principle of the CMDI initative.

  http://www.clarin.eu/content/component-metadata

- **Component Registry** (online common metadata repository) in which any user can browse already designed components and profiles and can create, edit, register and store its own.
- **Arbil:** The metadata editor recommended by CLARIN. It may be downloaded and installed locally by the metadata creators; each component and profile made public in the online Component Registry is imediately available locally for loading by Arbil if the machine is connected to the internet;
- We later used the **internal tool** (Moruz and Scutelnicu, 2014)developed for data cleaning that also has a module for metadata editing →dealing with extracting and cleaning the data and creating the metadata in one step.

Metadata creation

Starting from detailed CMDI profiles created in the CLARIN project for annotated text and speech corpora, we have designed profiles tailored to our specific needs:

- ✓ **general information (corpus level):** creators of the corpus, the availability and the licence, the development status, the projects and cooperation agreements that support the creation etc.
- ✓ **specific information (document level):** author of the metadata and of the manual pre-processing work, annotation details (tools, level of annotation, validation of annotation, etc.), the author, source, type and genre of the text, the number of words and other statistics for the document.

- **Manually:** using Arbil and later the in-house tool, for documents sent by our providers;
- **Automatically:** for text files crawled from the web (articles, blogs): preliminary phase of mapping the existent classifications of texts on those sites onto our classification of texts.

Data annotation

The TTL (Ion, 2007) processing chain:

- ✓ **sentence splitting:** it uses regular expressions for the identification of a sentence end;
- ✓ **tokenization:** words are separated from the adjacent punctuation marks; compound words are recognized as a single lexical atom; cliticized words are split as distinct lexical entities;
- ✓ **POS tiered-tagging** (Tufiș, 1999), MULTEXT-East tag set; its accuracy is above 98%;
- ✓ **lemmatization:**
  - Word form + POS tag: recovers corresponding lemma from a large (over 1,200,000 entries) human-validated Romanian word-form lexicon; precision is almost 99%: POS tagging errors
  - unknown words (not tagged as proper names): the lemma is provided by a five-gram letter Markov Model-based guesser, trained on lexicon lemmas with the same POS tag as the token being lemmatized; precision about 83%
- ✓ **chunking:** assigning syntactic phrase labels, guided by a set of regular expression rules defined over POS tags;

Data annotation

- we intend to manually validate the annotation of a limited part of the corpus (2%, i.e. 10 million words)

- for further stages in the corpus development, we intent to add other types of annotations: **syntactic parsing, semantic annotation and discourse analysis**.

- the annotation of the speech data includes: segmentation at sentence level, pauses and non-lexical sounds marking, accent marking, syllabification, plus the grapheme to phoneme alignment and written word/spoken word alignment.

Example of an annotated sentence

Realizarea corpusului de limbă română contemporană este o obligație culturală.
“The realisation of the contemporary Romanian language corpus is a cultural obligation.”

<s id="id_temp.1">

<w lemma="realizare" ana="Ncfsry" chunk="Np#1">Realizarea</w>

<w lemma="corpus" ana="Ncmsoy" chunk="Np#1">corpusului</w>

<w lemma="de" ana="Spsa" chunk="Pp#1">de</w>

<w lemma="limbă" ana="Ncfsrn" chunk="Pp#1,Np#2">limbă</w>

<w lemma="român" ana="Afpfsrn" chunk="Pp#1,Np#2,Ap#1">română</w>

<w lemma="contemporan" ana="Afpfsrn“ chunk="Pp#1,Np#2,Ap#1">contemporană</w>

<w lemma="fi" ana="Vmip3s" chunk="Vp#1">este</w>

<w lemma="un" ana="Tifsr" chunk="Np#3">o</w>

<w lemma="obligație" ana="Ncfsrn" chunk="Np#3">obligație</w>

<w lemma="cultural" ana="Afpfsrn" chunk="Np#3,Ap#2">culturală</w>

<c>.</c>

</s>

Conclusions

- international context of growing interest for creating large language resources;
- joined effort of two Romanian academic institutes, greatly helped by publishing houses and editorial offices, which kindly accepted the inclusion of their texts at no costs;
- although large amount of texts are out there on the web, creating an IPR-clear reference corpus is the greatest challenge of this project and needs vast efforts invested in persuading IPR holders to contribute to such a cultural action;
- the corpus will be freely available online for search for all those interested in the study or processing of the Romanian language;

We express here our gratitude to all CoRoLa volunteers, undergraduate, graduate and Ph.D. students, as well as researchers and university staff in computer science and linguistics, who have generously agreed to help in the process of filling in metadata and cleaning the collection of texts.