An Insight into CoRoLa
The Reference Corpus of Written and Spoken Contemporary Romanian

Dan Tufiș, Dan Cristea

Romanian Academy

Data collections

- **Archives** – a repository of readable electronic texts not linked in any coordinated way

- **Electronic Text Library -** a collection of electronic texts in standardised format with certain conventions relating to content etc, but without rigorous selectional constraints.

- **Corpus -** a subset of an ETL, built according to explicit design criteria for a specific purpose: a corpus is not a collection of texts which are deemed 'interesting' or 'useful' of themselves; the texts in the corpus are interesting and useful for the study of language

Sue Atkins, Jeremy Clear, Nicholas Ostler – Corpus Design Criteria, 1991

A Life-time Enterprise

- One of the most important decisions of a NLP community is building a reference corpus for the language in case.
- It is a scientifically exciting, multidisciplinary project and it has a major cultural dimension.
- In an IPR strictly regulated society, gathering large quantities of text and speech data, representative for a language is not an easy task.
- It has to be maintained over an indefinite period of time

What is COROLA?

- A priority project of the Romanian Academy (2014-2017)

- The Reference Corpus for Contemporary Romanian Language:
  Contemporary: since (~)1999
  Reference: covering all literary language registers and styles
  <= electronic format available at text providers
  NO: scanning, OCR-ization

- One of the largest Reference Corpus in the world which is and will be fully IPR-cleared (we have signed agreements with some of the most important publishing houses, journals and broadcasting agencies for using texts and voice recordings)

The making-up of COROLA

Who develops it?

Prezidium of the Romanian Academy commissioned in 2014 this project to the Research Institute for Artificial Intelligence "Mihai Drăgănescu" in Bucharest (ICIA) and Institute for Computer Science in Iași (IIT)

The project is backed up by experts and students from University "A.I. Cuza" of Iași, Linguistic Institute "Al. Philippide" of Iași, University "Politehnica" of Bucharest, University of Bucharest, Technical University of Cluj-Napoca and University of Craiova.

The project is expected to deliver the first operational version by the end of 2017.

The Objectives

- Large volumes of textual data (> 500,000,000 words: ~5,000 novel books) and speech data (~ 300 hours of recordings) deeply processed (morpho-lexically, syntactically and partial semantically) with standard documentation
- Covering all all functional styles of the literary language: scientific, official, journalistic and imaginative.
- Covering 5 large domains (arts &culture, society, science, nature and others) which are further refined into 71 sub-domains.

Foreseen applications

- Linguistic studies (lexicography, terminology, syntax, semantics); educational instrument for students
- Language modeling for automatic processing of Romanian language
- Translation modeling
- Language learning
- Intelligent indexing and multi-criterial retrieval (text &speech)
- Semantic classification of large volumes of data (text & speech)
- Knowledge extraction from data (text & speech)
- Automatic summarization of documents
- Question answering in Romanian language (v. Watson – Jeopardy!)
- Automatic speech recognition and synthesis
- Machine Translation (text & speech)

COROLA corpus

What are the level of linguistic annotations:

Currently:

Phoneme, syllable, word, POS tagging, syntactic chunking, dependency parsing (tree-bank prototype)

Foreseen:

sense-tagging (based on Ro-Wordnet sense inventory);

discourse mark-up (NE linking, anaphora resolution)

Tools for language data processing:

Currently:

TTL, Bermuda, RARE, MIRA, Grammar Studio, MaltParser, yEd, Sphinx, HTK, and many others.

The current status (June 2017)

distribution of textual data

Domain		Style	
arts&culture	**27,697,861**	journalistic	**77,277,228**
society	**119,150,171**	science	**184,761,720**
others	**571,986,834**	imaginative	**51,617,302**
science	**160,309,410**	others	**2,100,318**
nature	**1,831,275**	memoirs	**26,135,623**
		administrative	**11,564,015**
		law	**527,519,345**
TOTAL	**880,975,551**	TOTAL	**880,975,551**

distribution of speech data

Corpus	Type	Source	Time length (h:m:s)
RASC	many speakers	(read) RoWikipedia	14:22:02
RSS-ToBI	single speaker	(read) news&fairy tales	03:44:00
RADOR	many speakers	read news& interviews	106:52:33
Radio Iaşi	many speakers	read news& interviews	under development
Audio-books (not IPR cleared)	single/multiple speaker	read stories (~200h)	**134:57:24**

SWARA

COROLA corpus

We concluded Agreements with IPR holders for language data to be included into the corpus:

- 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 (books)
- România literară, Muzica, Actualitatea muzicală, Destine literare, DCNEWS, PRESSONLINE.RO, the school magazine of Unirea National College from Focșani (journals, news)
- Bloggers: Simona Tache, Dragoș Bucurenci, Irina Șubredu and Teodora Forăscu.
- Oral texts (read news, live transmissions and live interviews) are provided by Rador, the press agency of Radio Romania from Bucharest and by Radio Iași

What's in a corpus?

A corpus, besides proper language data, contains additional information on the properties of texts (written or spoken) that are included. This is achieved by means of annotation.

The annotation is a principal feature of the corpus, distinguishing it from collections of texts. Representing this type of information is a matter of standardization for lots of good reasons (identification, dissemination, aggregation, interoperability, etc.)

A corpus, usually, includes two types of annotation:

a)metatextual (information about the text) -metadata,

b)linguistic- phonetic, prosodic, morphological, phrasal, syntactic, semantic, pragmatic (not necessarily all)

Annotations

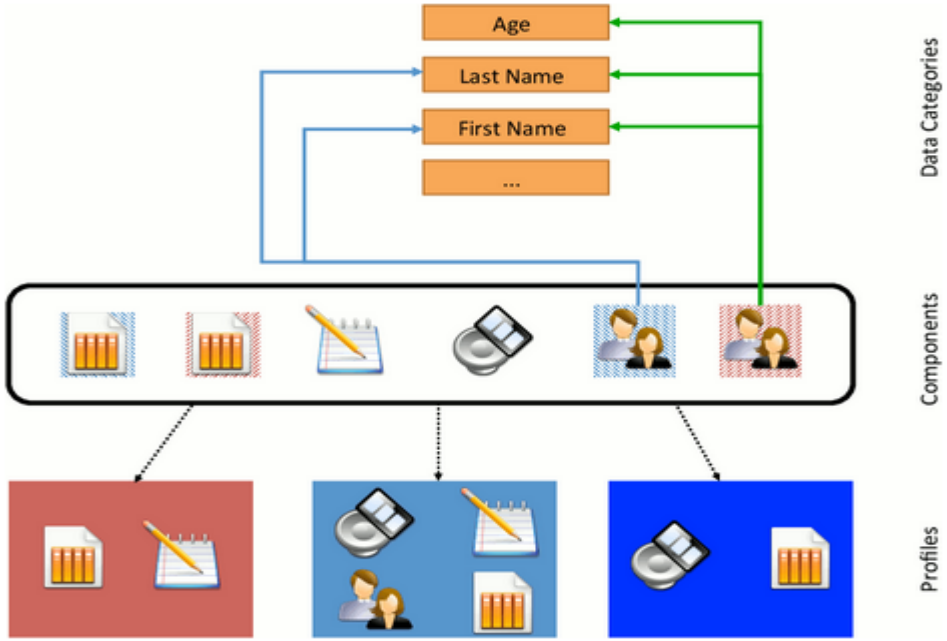
- Inline – traditional annotations (LOB, Pen-treebank, ROMBAC, etc) most of them based on XML schemas. Meta-information resides into the text-files. Adequate for non-overlapping layers and single viewpoints (tokenization, POS tagging, chunking).
- Stand-off – annotations are stored separately from the primary data leaving the primary data untouched (DeReKo).
- Hybrid annotations - ex. TCF uses inline annotation for the tokenization and POS tagging layer and stand-off for the syntactic layer.

Metadata schema (I)

- Metadata are essential for indexing the corpus and they facilitate the search process for end users.
- Metadata for language resources and tools exists in a multitude of formats: we opted to use the **CMDI (Component MetaData Infrastructure)** metadata format.
- CMDI offers ready-made sets of metadata elements (components) for various types of resources; they can be edited, modified, or combined into personalized metadata schemas - profiles
- The CMDI model has close ties to the ISOcat data category registry.

Metadata schema (II)

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



When adding an element to a CMDI component, the metadata modeler has to add a link to a **Concept Registry** (based on the **ISOcat data category registry**) where very detailed definitions are available. This link provides a persistent and unique identification of the intended semantics.

Metadata schema (III)

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 license, the development status, the projects and cooperation agreements that support the creation etc.
- ✓ **specific information (document level):** the document/article title, collection and publication date, document type, document (literary) style, document domain/sub-domain, the author, the source, annotation details (tools, level of annotation, validation of annotation, etc.), the number of words.

Linguistic annotations

Currently:

Phoneme, syllable, word, POS tagging, syntactic chunking, dependency parsing;

Foreseen:

sense-tagging (based on Ro-Wordnet sense inventory);

discourse mark-up (NE linking, anaphora resolution)

Tools for language data processing:

TTL, Bermuda, RARE, MIRA, MaltParser, TensorFlow, yEd, Sphinx, HTK, and many others.

Morpho-lexical processing: example

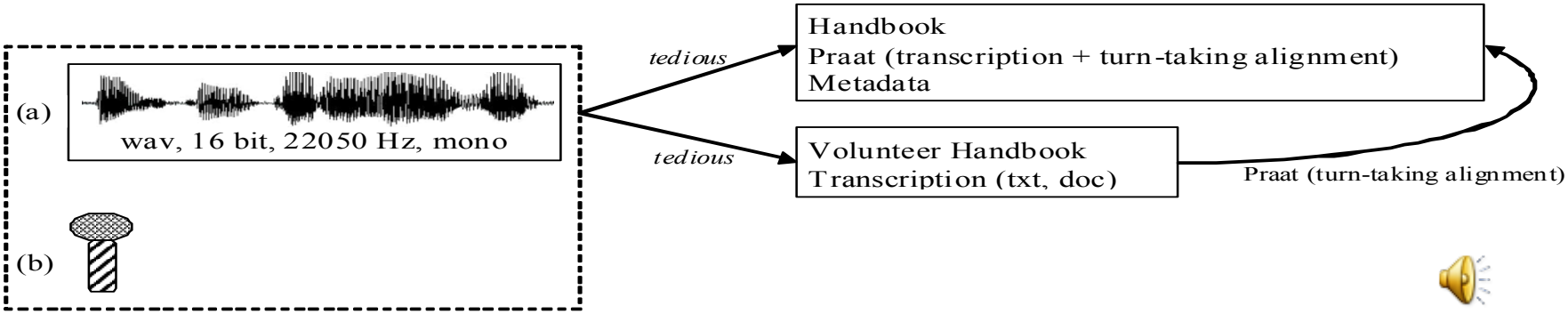
```
- <S id="5" offset="637">
    <W Case="direct" Definiteness="yes" Gender="feminine" LEMMA="scumpire" MSD="Ncfsry" Number="singular" POS="NOUN" Type="common" id="5.1" offset="0">Scumpirea</W>
    <W Case="oblique" Definiteness="yes" Gender="feminine" LEMMA="ou" MSD="Ncfpoy" Number="plural" POS="NOUN" Type="common" id="5.2" offset="10">ouălor</W>
    <W Case="direct" LEMMA="sine" MSD="Px3--r" POS="PRONOUN" Person="third" Type="reflexive" id="5.3" offset="17">se</W>
    <W EXTRA="tranzitiv" LEMMA="datora" MSD="Vmip3p" Mood="indicative" Number="plural" POS="VERB" Person="third" Tense="present" Type="predicative" id="5.4" offset="20">datorează</W>
    <W LEMMA="în" MSD="Sp" POS="ADPOSITION" id="5.5" offset="30">în</W>
    <W Case="direct" Definiteness="yes" EXTRA="NotInDict" Gender="masculine" LEMMA="pricipal" MSD="Ncmsry" Number="singular" POS="NOUN" Type="common" id="5.6" offset="33">pricipal</W>
    <W Case="oblique" Definiteness="yes" Gender="feminine" LEMMA="criză" MSD="Ncfsoy" Number="singular" POS="NOUN" Type="common" id="5.7" offset="42">crizei</W>
    <W LEMMA="de~pe" MSD="Sp" POS="ADPOSITION" id="5.8" offset="49">de pe</W>
    <W Case="direct" Definiteness="no" Gender="feminine" LEMMA="piață" MSD="Ncfsrn" Number="singular" POS="NOUN" Type="common" id="5.9" offset="55">piață</W>
    <W LEMMA="pentru~că" MSD="Cs" POS="CONJUNCTION" id="5.10" offset="61">pentru că</W>
    <W LEMMA="nu" MSD="Qz" POS="PARTICLE" Type="negative" id="5.11" offset="71">nu</W>
    <W EXTRA="tranzitiv" LEMMA="avea" MSD="Vmip1p" Mood="indicative" Number="plural" POS="VERB" Person="first" Tense="present" Type="predicative" id="5.12" offset="74">avem</W>
    <W Case="direct" Definiteness="no" Gender="feminine" LEMMA="ou" MSD="Ncfprn" Number="plural" POS="NOUN" Type="common" id="5.13" offset="79">ouă</W>
    <W LEMMA="de" MSD="Sp" POS="ADPOSITION" id="5.14" offset="83">de</W>
    <W Case="direct" Definiteness="yes" Gender="feminine" LEMMA="calitate" MSD="Ncfsry" Number="singular" POS="NOUN" Type="common" id="5.15" offset="86">calitatea</W>
    <W Form="roman" LEMMA="i" MSD="Mo-s-r" Number="singular" POS="NUMERAL" Type="ordinal" id="5.16" offset="96">I</W>
    <W LEMMA="," MSD="COMMA" id="5.17" offset="97">,</W>
    <W Case="direct" Definiteness="no" EXTRA="ParticipleLemma:obține(tranzitiv)" Gender="feminine" LEMMA="obținut" MSD="Afpfprn" Number="plural" POS="ADJECTIVE" id="5.18" offset="99">obținute</W>
    <W LEMMA="de~la" MSD="Sp" POS="ADPOSITION" id="5.19" offset="108">de la</W>
    <W Case="direct" Definiteness="yes" Gender="feminine" LEMMA="găină" MSD="Ncfpry" Number="plural" POS="NOUN" Type="common" id="5.20" offset="114">găinile</W>
    <W Case="direct" Definiteness="no" EXTRA="NotInDict" Gender="feminine" LEMMA="nestresate" MSD="Afpfprn" Number="plural" POS="ADJECTIVE" id="5.21" offset="122">nestresate</W>
    <W LEMMA="." MSD="PERIOD" id="5.22" offset="132">.</W>
  </S>
</POS_Output>
```

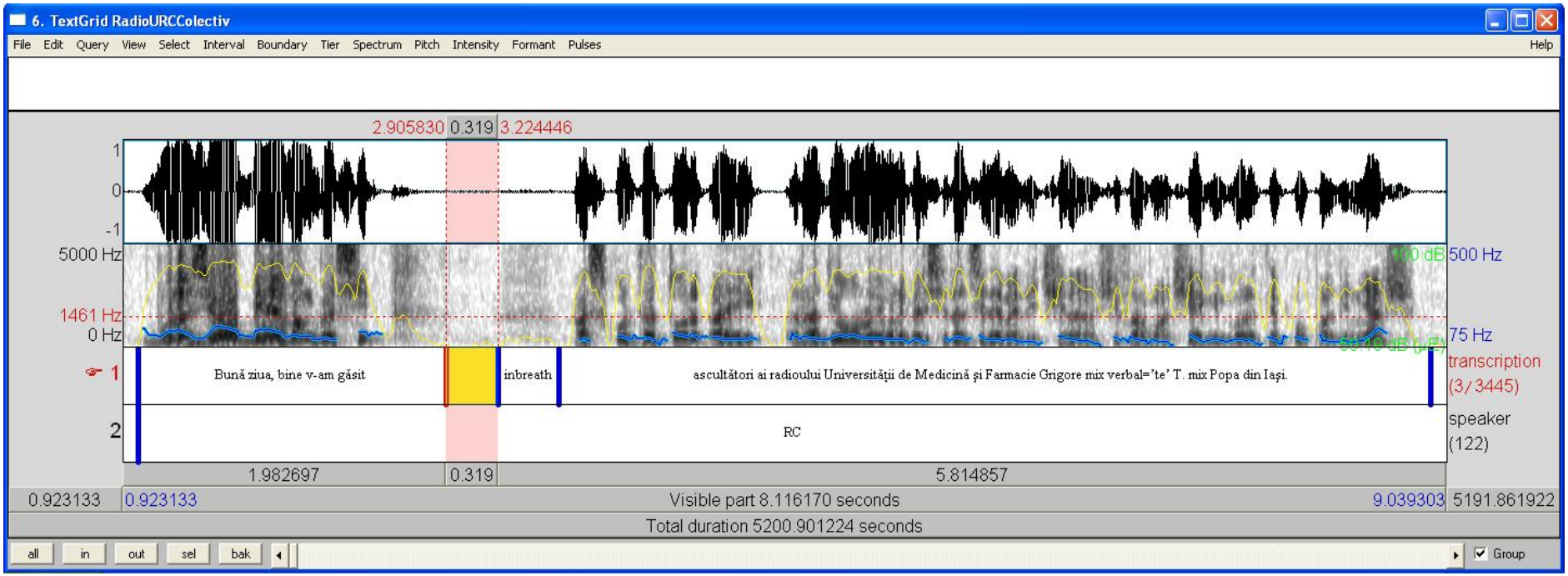
Oral texts automatic processing - annotation levels -

- Accompanied by their written counterpart
- Alignment: oral sentence – written sentence
  - Lemmatization
  - Tokenization
  - Part-of-speech tagging
  - Syllabification
  - Some allophones

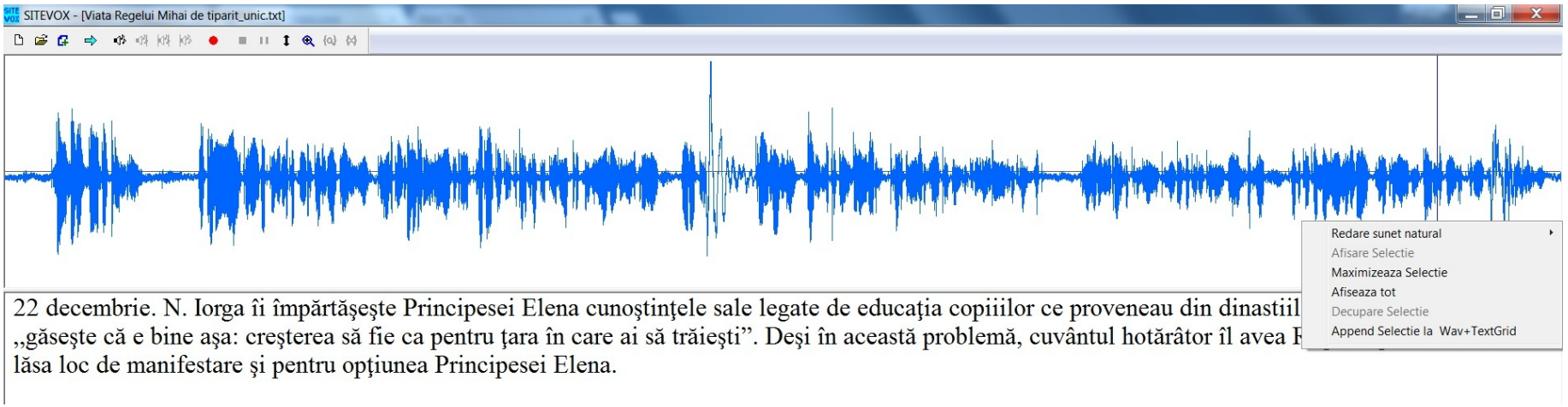
In-house recording

- Application developed at ITI-Iași (dr. V. Apopei) coupled with Praat (transcription and turn-taking alignment).





In-house sound recordings



Processing speech data

0.00 0.63 [silence]
0.63 1.38 destinderea
1.38 1.48 [silence]
1.48 1.92 r3ce1te
1.92 2.45 aburul
2.45 2.76 [silence]
2.76 3.04 astfel
3.04 3.17 [illegible]
3.17 3.46 poate
3.46 3.78 ap3rea
3.78 3.82 [silence]
3.82 4.27 condensar
4.27 4.50 unei
4.50 4.83 p3r2i
4.83 5.10 din
5.10 5.42 abur
5.42 5.79 [silence]

Word-time alignment

Phoneme-time alignment

0 6300000 sil sil6300000 7000000
d destinderea7000000 7500000
e7500000 8400000
s8400000 9300000
t9300000 9900000
i9900000 10800000
n10800000 11200000
d11200000 11800000
e11800000 12100000
r12100000 12600000
e@[illegible]300000 13800000
[illegible]13800000 14800000
sp14800000 15500000
r r3ce1te15500000 16200000
@16200000 17200000
ch17200000 17500000
e17500000 18000000
. . .

Corpus Management Platforms

- Corpus management:
  - acquisition of raw data (text, speech)
  - cleaning
  - metadata
  - maintaining
  - access

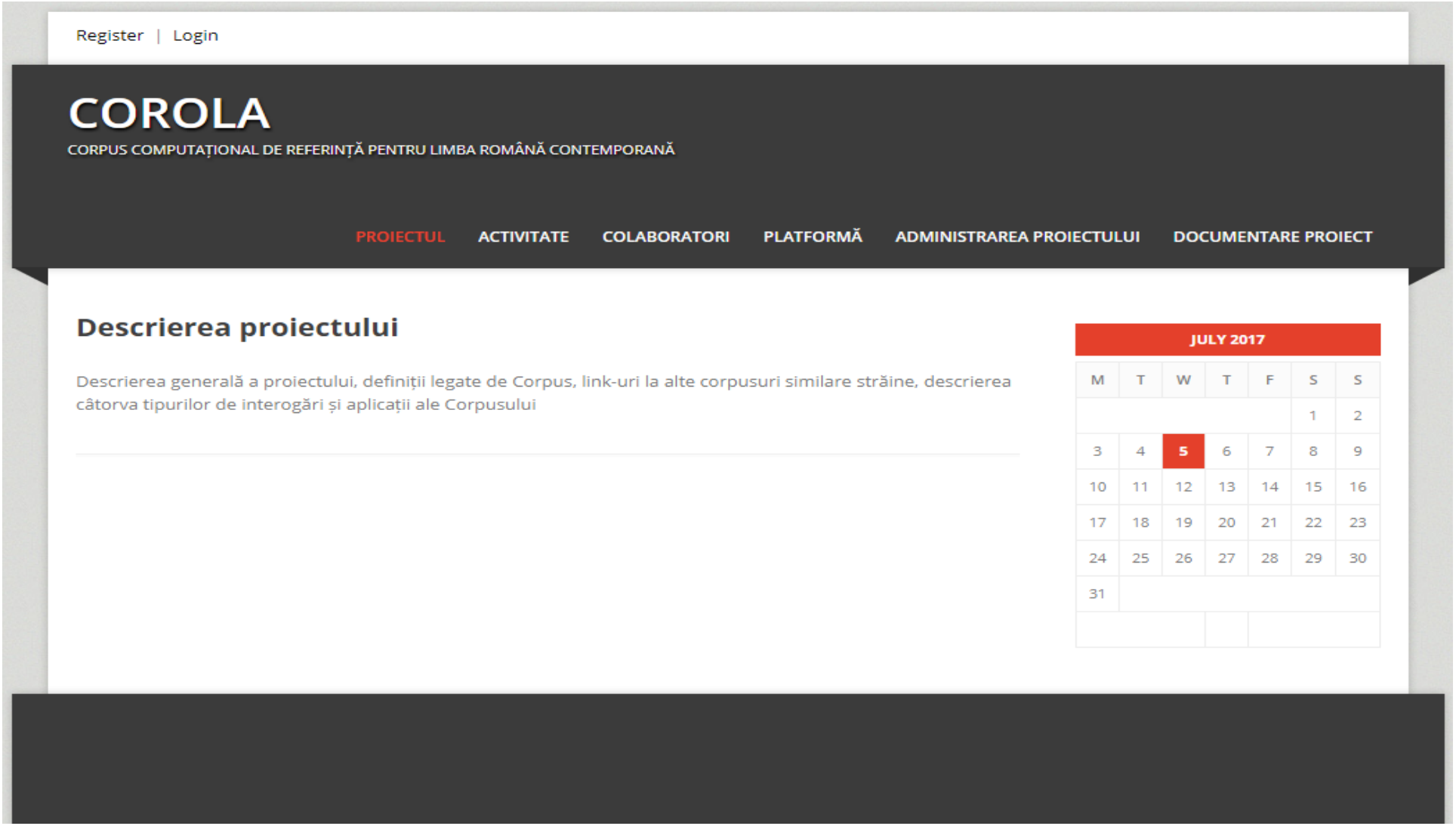
Processing data:
Curator – Provider – Portal

Portalul COROLA

CoDaP:

CoRoLa Data cleaning and metadata Platform

(http://89.38.230.23/)

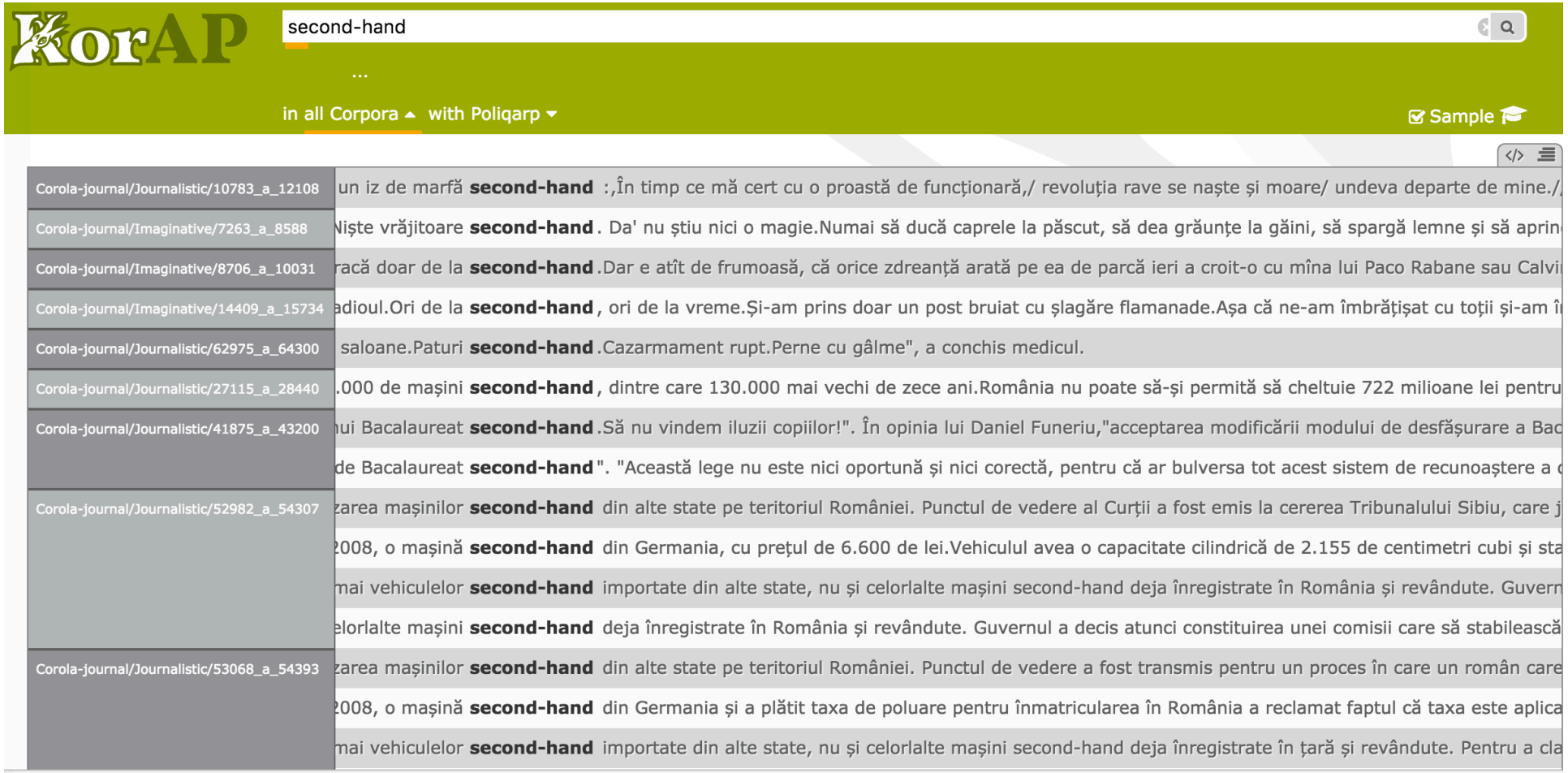


Access

- CoRoLa will be open for querying in two environments
  - The IMS Open Corpus Workbench (CWB), http://cwb.sourceforge.net/
  - The KorAP Query interface (IDS Mannheim)
- Both IMS and KorAP are equipped with specific corpus investigation facilities (counting tokens filtered by user-specified criteria, collocation analysis, concordancing, various statistical test batteries, etc.).
- Downloadable to a certain extent

CoRoLa in IMS Open Corpus Workbench (CWB)

CoRoLa in KorAP



The KorAP Query interface

- Suited for management of large corpora (tens of billions of words)
- Easily adaptable to different annotation styles
- Powerful query language:
  - multiple levels,
  - query criteria: any field in the metadata and any possible combination of these fields
  - user can build his/her own virtual corpus: filtered subcorpora (e.g. ”texts on architecture published between 2000 and 2005”)
- Search results: snippets of a reasonable size for linguistic investigations (1-2 sentences)
- Allows for distributed data (Bucharest, Iași, Mannheim)

KorAP accesses also: DeReKo - *Deutsche Reverenzkorpus*

DeReKo – at IDS, Mannheim

- the world's largest collection of German texts (>25 billion tokens)
- a broad variety of text types with a quantitative focus on newspaper texts and rapidly growing portion of computer mediated communication

DRuKoLA Objectives (I)

1. **Construction and harmonization of comparable corpora** in German and Romanian
2. Development of **criteria for building comparable virtual sub-corpora** from DeReKo and CoRoLa, based on metadata and other possible text properties
3. Exploration of **language-specific peculiarities of the studied languages and equivalences** with respect to different parameters and structures
4. Some **corpus-based comparative case studies** on

   a) markers of modality: *haben/a avea* with *zu*-infinitives and supine,

   b) (abstract) demonstratives in German and Romanian,

   c) investigation of distributional semantic and syntagmatic properties of corresponding forms and structures.

DRuKoLA Objectives (II)

5. Experimentation and enhancement of a **common corpus analysis platform** to share the corpus, technical and research results

6. Building a **crystallization structure to serve other national or reference corpora**, with the long-term goal of pioneering a federated, at least European, reference corpus, where each collection of texts is still physically located at and curated by its responsible institute, but can be dynamically queried and extracted to different comparable corpora

Towards integrating multiple linguistic resources

- Methodology:
  - Common practice: high level language processors are trained on resources that mix the raw linguistic data with expert annotation
  - Then:
    - use CoRoLa as an anchor on which these other linguistic resources are coupled
    - build an environment that allows complex queries, simultaneously accessing resources of different types

Linguistically Linked Open Data – resources that help to build text processors for CoRoLa

- Thesaurus dictionary of the Romanian language in electronic form – eDTLR

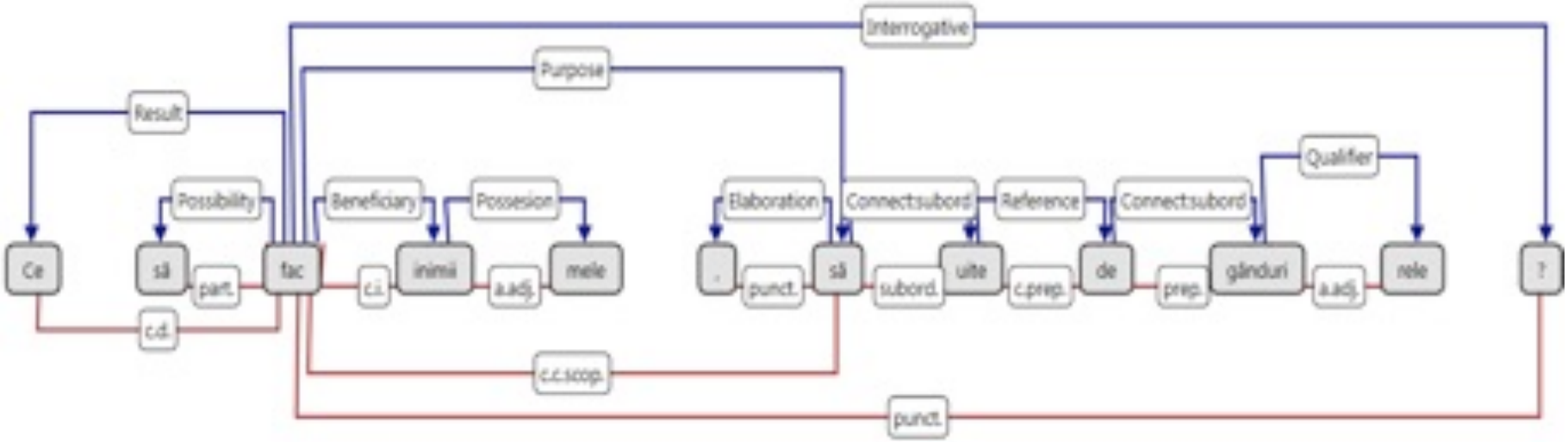
  http://edtlr.info.uaic.ro/

  –train a word sense disambiguation program
- Romanian treebank with +10,000 sentences in 2017 – RoDep Treebank at RACAI and UAIC-FII

  –train a syntactic parser

Linguistically Linked Open Data – resources that help to build text processors for COROLA

- A semantically annotated treebank –
  - correct syntactic roles on semantic ground



Thanks Cătălina Mărănduc, PhD report, Nov. 2016

Linguistically Linked Open Data – resources that help to build text processors for COROLA

- A corpus of semantic relations – QuoVadis

  http://nlptools.info.uaic.ro/Resources.jsp

  –train a program to recognise coreference, affective, kinship, social relations

Continuation of CoRoLa: a diachronic corpus

- Acquisition of textual data
  - including helped by a Cyrillic OCR (CyRo – a project in the second phase of evaluation)
- Infer paradigmatic morphology of old Romanian
  - from eDTLR citations and other sources
- JUST IMAGINE: manuscripts → scanned → OCRed → transcribed → POS-tagged etc. → included in the diachronic corpus

Further prospects

- **EuReCo - The idea of a European Reference Corpus**

Current Situation

  - several national initiatives loosely connected by bilateral contacts
  - co-operation within CLARIN but subordinated to various other goals and funding necessities
  - coordination via EFNIL, so far mostly unrelated to corpora
  - some initiatives maintain their own parallel or comparable corpora

- **Joining forces**
  - particularly desirable for comparable corpora, several national and reference corpora built and maintained anyway
  - creating methodology and techniques for joining them virtually
    - each national centre still responsible for its language
    - each corpus still physically located at its centre

Acknowledgements

- To our colleagues involved in CoRoLa:
  - from AR-RACAI: Verginica Barbu-Mititelu, Tibi Boroș, Ștefan Dumitrescu, Radu Ion, Elena Irimia
  - from AR-IIT: Vasile Apopei, Cecilia Bolea, Daniela Gîfu, Alex Moruz, Mihaela Onofrei, Laura Pistol, Andrei Scutelnicu
- To our collaborators in DRuKoLa:
  - from Univ. Bucharest: Ruxandra Cosma
  - from IDS Mannheim: Nils Diewald, Marc Kupietz, Eliza Margaretha, Andreas Witt

Thank you for your attention