


# Resurse lingvistice dezvoltate în cadrul ICIA

13 iulie 2023, Verginica Barbu Mititelu

# Corpusurile - preocupare mai veche la ICIA

- 1998: varianta românească a 1984 de G. Orwell, adnotată morfologic
- 2006: JRC-Acquis-Ro - 
  - componenta românească a corpusului paralel folosit în TA
- 2012: ROMBAC (36M): 5 domenii aproximativ egal reprezentate: jurnalistic, medical și farmaceutic, juridic, filologic și beletristic



# Corpusuri create mai recent

- Corpusul reprezentativ pentru LRC: **CoRoLa**
- Corpusuri specializate:

**RRT**

**SiMoNEro**

**PARSEME-Ro**

**LegalNEro**

**MARCELL**

**CURLICAT**

**USPDATRO**

**ROBINTASC**

**Microblogging**

# CoRoLa

**Corpus of Contemporary Romanian Language**

- 2013 - anul de început
- Proiect prioritar al Academiei Române
- Parteneriat:
  - Pentru dezvoltare
  - Pentru colectarea datelor
  - Pentru prelucrarea datelor
  - Pentru exploatare

# CoRoLa - conținut

- Texte scrise - texte orale cu echivalent scris
- Texte din România și din publicații ale diasporei
- Texte din ultimii 20 de ani
- Stiluri funcționale: beletristic, jurnalistic, memorialistic, juridic, administrativ, științific, postări pe bloguri
- Domenii: artă & cultură, natură, știință, societate
- Subdomenii: peste 50
- Textele orale: spontane sau citite; înregistrări profesioniste sau de amator
- Toate textele sunt însoțite de metadate

# Bănci de arbori sintactici

- Romanian Reference Treebank (RRT)

- corpus general

RRT

218K

ⒻⒻ



- adnotat morfosintactic conform principiilor din Universal Dependencies 

- cuprins în lansările bianuale ale UD

- SiMoNERo

SiMoNERo

146K

ⒻⒻ



- corpus medical

- adnotat morfosintactic conform principiilor 

- adnotări cu entități medicale: ANAT, DISO, CHEM și PROC

# Corpusuri adnotate cu expresii/entități

- [PARSEME-Ro](#)

- Corpus jurnalistic (1M), adnotat manual cu 4 tipuri de expresii verbale (VID, LVC, IRV, IAV)

- Adnotat morfo-sintactic cu [UDPipe 2](#)

- Parte din campaniile de evaluare organizate

P A R S E M E

- Distribuit ca parte a corpusului PARSEME



- [LegalNERo](#)

- Corpus juridic (265K), adnotat manual cu 5 clase de NEs (PER, LOC, ORG, TIME, LEGAL) (Kappa 0.89)

- Adnotat morfo-sintactic cu [UDPipe](#)

- Disponibil pe

The Zenodo logo, which consists of the word "zenodo" in a white, lowercase, sans-serif font, centered within a blue rectangular background.

# Corpusuri ca resurse pentru eTranslation



- Corpus juridic - 7 limbi (bg, hr, hu, pl, ro, sk, sl) - ro: 412M
- Adnotat automat morfo-sintactic (TTL, NLP-Cube), cu NEs și cu descriptori EuroVoc și termeni IATE
- Clusterizare automată a documentelor similare în limbi diferite, aliniată la domeniile principale din EuroVoc



- Corpus multidomeniu (cultură, economie, educație, natură, sănătate, politică, știință) - 7 limbi (bg, hr, hu, pl, ro, sk, sl) - ro: 92M
- Adnotat automat morfo-sintactic (TTL, NLP-Cube), cu NEs și cu descriptori EuroVoc și termeni IATE, cu terminologie îmbogățită
- Anonimizat



# Corpusuri de voce

- **USPDATRO**

- Colecție (~6h) de clipuri video cu tipuri de voce absente sau slab reprezentate în seturile de date existente: femei, copii, adolescenți, bătrâni (○ 19-29, M)
- Descărcate de pe diverse platforme (YouTube, Vimeo, SoundCloud), cu licențe CC (Attr)
- Transcrise și aliniat manual

- **ROBINTASC**



- Corpus (6h25') înregistrat, cu voci F și M, cu întrebări citite, specifice interacțiunii cu un robot agent de vânzări într-un departament de calculatoare; vocabular specializat, pronunții variate, în special pentru cuvinte străine
- Utilizat pentru îmbunătățirea performanțelor unui sistem de RAV, pentru o microlume clar delimitată

# Microblogging

## (Twitter)

- Utilizarea Twitter API pentru colectarea mesajelor, folosind cuvinte-cheie. Eliminarea duplicatelor generate prin redistribuire.
- Anonimizarea
- Adnotare manuală cu 9 tipuri de NE: ORG, PER, LOC, TIME, LEGAL, ANAT, CHEM, DISO, MED\_DEVICE. Manualele de adnotare existente au necesitat adaptarea la specificul microbloggingului
- Codificarea manuală a limbii mesajelor, cu interes pentru code-switching
- Evaluarea manuală a:
  - emoției transmise (+/-/o)
  - prezenței elementelor specifice Hate speech
  - folosirii elementelor specifice microbloggingului ('LOL' etc.)

# Modele lingvistice

<https://relate.racai.ro>

The screenshot shows a web browser window with the URL <https://relate.racai.ro/index.php?path=lr/lt/models>. The page title is "Romanian Portal of Language Technologies".

**RELATE**

- TEPROLIN Service >
- CoRoLa >
- RoWordNet >
- Machine Translation >
- Speech >
- EUROVOC Classification
- CURLICAT Anonymization
- Named Entity Recognition
- Punctuation Restoration
- Social Media >
- Question Answering >
- Resources and Models ▾
  - Language Models
- Language Resources

**Contextualized embeddings**

- RoBERT: There are two models available [bert-base-romanian-cased-v1](#) and [bert-base-romanian-uncased-v1](#). A GitHub repo with useful scripts is available [here](#). Related paper is *Dumitrescu Stefan, Andrei-Marius Avram, and Sampo Pyysalo. "The birth of Romanian BERT." In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings, pp. 4324-4328, 2020.*
- Romanian DistilBERT: Constructed based on the [bert-base-romanian-cased-v1](#) model, the model is available on HuggingFace as [distilbert-base-romanian-cased](#). A GitHub repo is available [here](#).

**Word Embeddings from the CoRoLa project**

- All word embeddings from the CoRoLa project can be downloaded and used interactively [here](#).
- The recommended model, according to a number of experiments, can be downloaded directly from [here](#).

**BioMedical Word Embeddings**

These word embeddings were trained on the BioRo corpus ( *Mitrofan, Maria and Tufiş, Dan. BioRo: The Biomedical Corpus for the Romanian Language. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018). pp. 1192-1196, 2018* ). The models have dimension 300 and were trained using FastText. Words occurring only 1,5 or 20 times were considered.

- [bioro.300.1.5.vec.gz](#).
- [bioro.300.5.5.vec.gz](#).
- [bioro.300.20.5.vec.gz](#).

Copyright © RACAI. All rights reserved.

# Lexicoane

- RoLEX - cel mai extins lexicon fonologic validat disponibil pentru limba română (330.866 de intrări):
  - forma lema eticheta\_MSD silabe accent transcriere\_fonetica
- tbl.wordform:
  - forma lema eticheta\_MSD

# Wordnetul românesc

- Rețea lexico-semantică
  - Noduri: mulțimi de sinonime
  - Arce: relații semantice între cuvintele din noduri
- ~60.000 noduri, > 85.000 cuvinte, >76.000 lume unice
- Aliniat cu Princeton WordNet

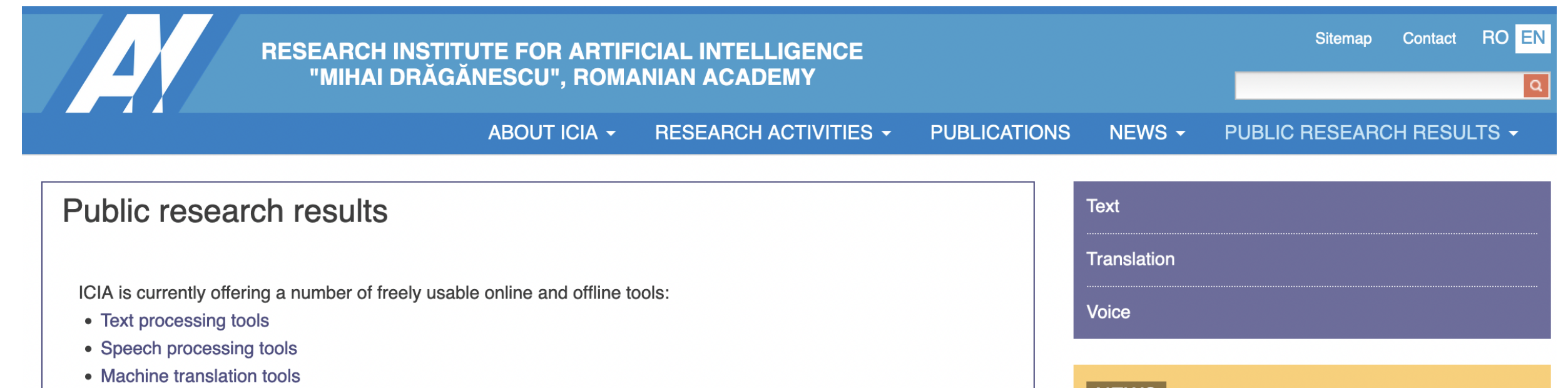
# Standardizare

- Linked Data
- Nexus Linguarum COST Action
- F(indable)A(ccessible)I(nteroperable)R(eusable)

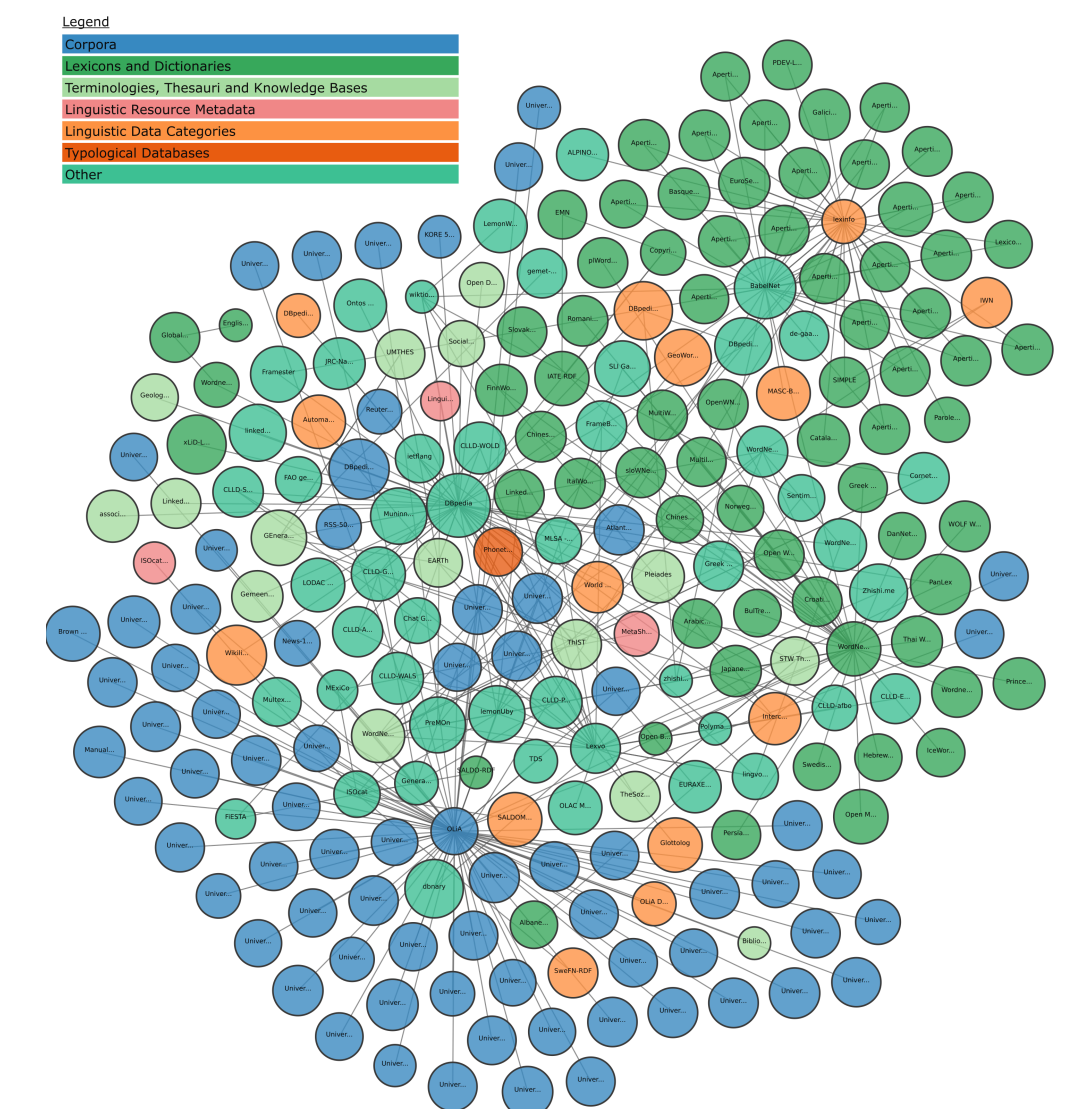
RoLLOD 2020-2023

ROMANIAN LANGUAGE RESOURCES CONVERTED  
TO *LINKED DATA* SPECIFICATIONS

# Accesul la resurse



- Liber
- Limitări externe: CoRoLa
- Metadate accesibile pe platformele europene:
  - European Language Grid
  - META-SHARE
  - Linked Open Data Cloud
- Acces direct la date pe site-ul ICIA, GitHub



**Muṭumesc!**