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

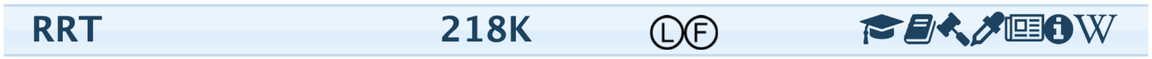
- 2013 - anul de început
- Proiect prioritar al Academiei Române
- Parteneriat:
  - Pentru dezvoltare
  - Pentru colectarea datelor
  - Pentru prelucrarea datelor
  - Pentru exploatare

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



- adnotat morfosintactic conform principiilor din Universal Dependencies
- cuprins în lansările bianuale ale UD
- **SiMoNERo**

- 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
- 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

Corpusuri ca resurse pentru eTranslation

European Language Resource Coordination
Connecting Europe Facility
ELRC-SHARE

MARCELL
Multilingual Resources for CEF.AT in the legal domain

- 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, aliniate la domeniile principale din EuroVoc

CURLICAT
Curated Multilingual Language Resources for CEF.AT

- 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 aliniate manual
- **ROBINTASC**

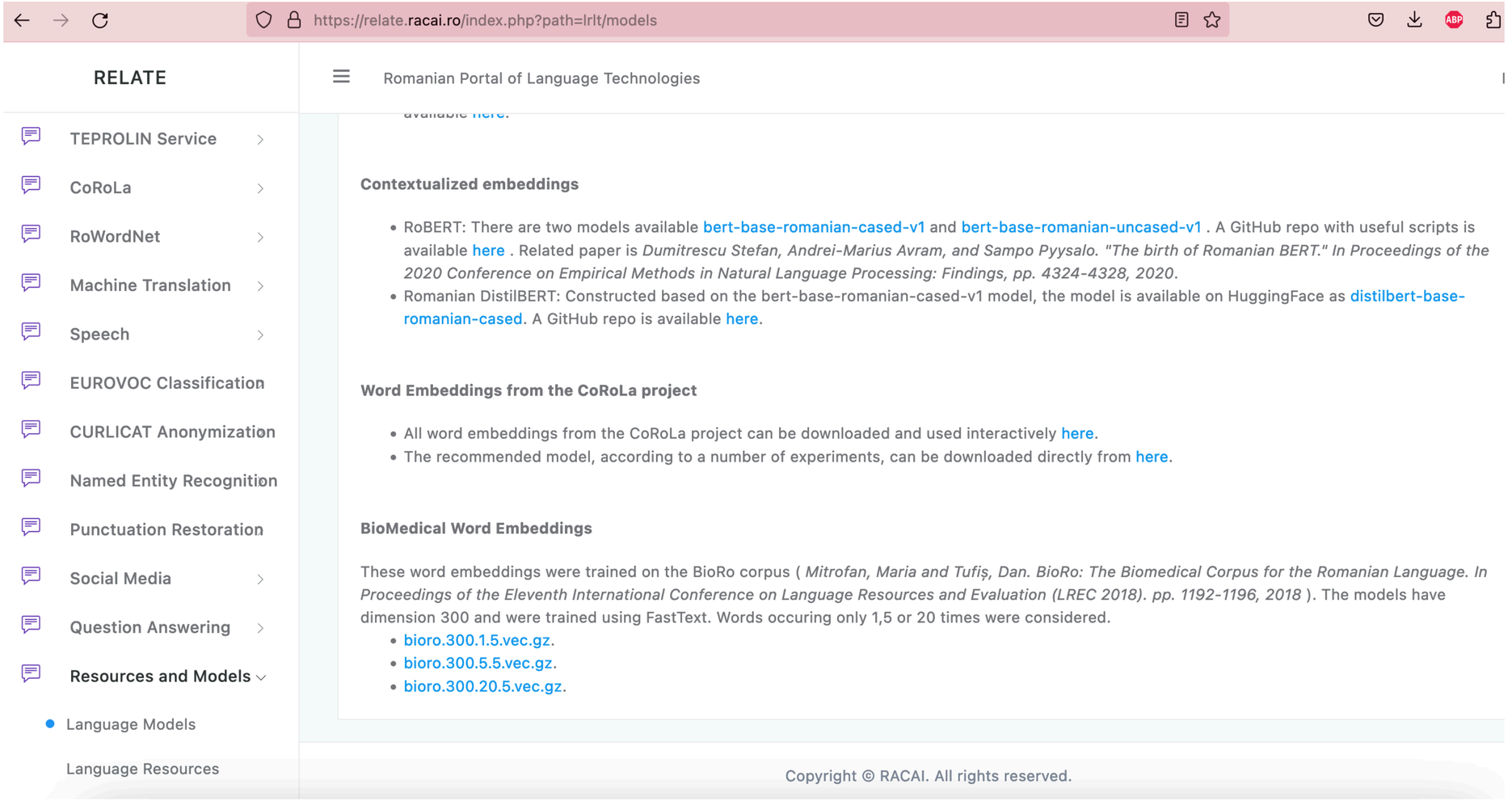
- Corpus (6h25') înregistrat, cu voci F și M, cu întrebări citite, specifice interacțiunii cu un robot agent de vânzări într-un departament de calculatoare; vocabular specializat, pronunții variate, în special pentru cuvinte străine
- Utilizat pentru îmbunătățirea performanțelor unui sistem de RAV, pentru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



Lexicoane

- RoLEX - cel mai extins lexicon fonologic validat disponibil pentru limba română (330.866 de intrări):
    - forma lema eticheta_MSD silabe accent transcriere_fonetică
- 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

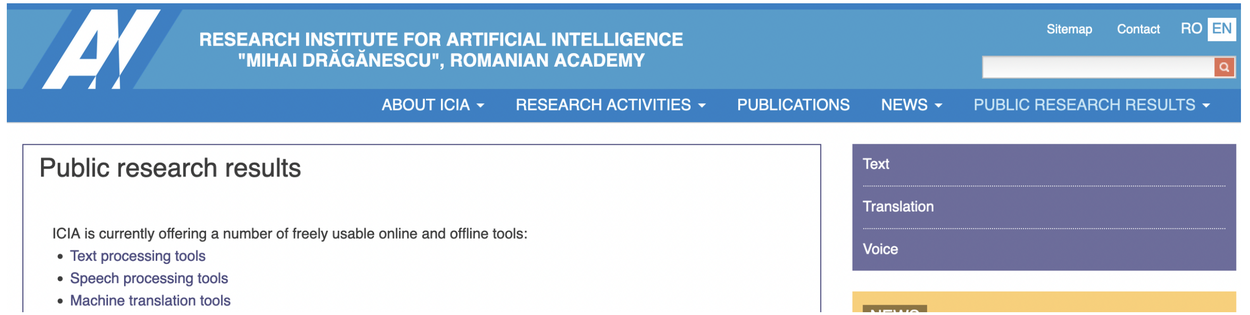
Standardizare

RoLLOD 2020-2023

ROMANIAN LANGUAGE RESOURCES CONVERTED TO *LINKED DATA* SPECIFICATIONS

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

Accesul la resurse

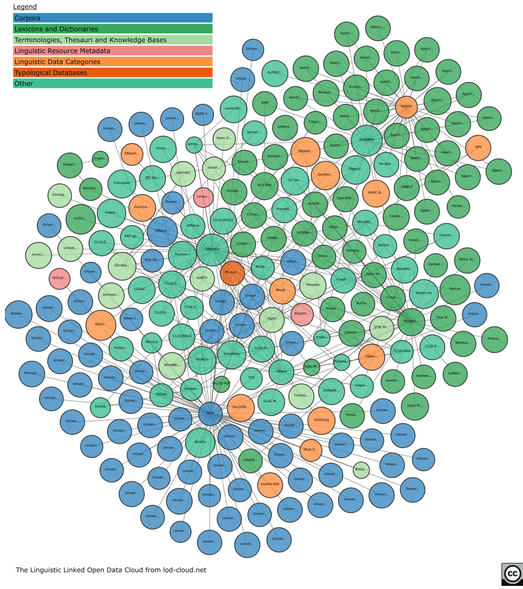


- Liber
- Limitări externe: CoRoLa
- Metadate accesibile pe platformele europene:
  - European Language Grid
  - META-SHARE
  - Linked Open Data Cloud
- Acces direct la date pe site-ul ICIA, GitHub

EUROPEAN LANGUAGE GRID

European Language Resource Coordination
Connecting Europe Facility

META-SHARE



Mulțumesc!