EUROPEAN LANGUAGE DATA SPACE

The importance of language data for the development of LT solutions - future steps

Verginica Barbu Mititelu (Romanian Academy Research Institute for Artificial Intelligence, Romania)
vergi@racai.ro

22-10-2024 LDS Country Workshop, Cluj-Napoca, Romania
https://language-data-space.ec.europa.eu

RACAI4RO

Aims of the presentation

- to explain the importance of language data nowadays

Aims of the presentation

- to explain the importance of language data nowadays

- to paint the picture of the datasets for the Romanian language

Aims of the presentation

- to explain the importance of language data nowadays
- to paint the picture of the datasets for the Romanian language
- to raise awareness about the importance of key data characteristics for the future of LTs

Aims of the presentation

- to explain the importance of language data nowadays
- to paint the picture of the datasets for the Romanian language
- to raise awareness about the importance of key data characteristics for the future of LTs
- to contextualize the above aspects in the European Language Data Space initiative

AI represents one of the greatest opportunities for global societal and economic progress

- automation, optimization, and advanced data analysis
- faster, cheaper, and more accurate
- new skills, new jobs, new industries
- personalized solutions (education, health, entertainment…)
- remote solutions for local problems in isolated places lacking expertise/history
- …

AI has become the focus for a large number of technologies and applications

- due to its ability to learn, adapt, and improve autonomously

- it is ubiquitous

- embedded in the digital tools we use on a daily basis
  - especially assistants - able to understand human speech and text and respond to them

Importance of data for AI

- fuel/lifeblood
- learn, reason and make decisions based on fueled data
- level of “intelligence”, performance accuracy, results’ relevance are directly impacted by data’s
  - quantity
  - quality
  - diversity
- continuous learning form data =>
  - personalised solutions/recommendations/etc.
  - need for:
    - keeping data current,
    - correcting existing data,
    - augmenting data => improved, enhanced results

Language Technologies, Artificial Intelligence and (Large) Language Models

- Language Technologies are directly impacted by AI
- Well-known techniques such as Machine Learning and Deep Learning are subsets of Artificial Intelligence
- Deep learning applies to many language technologies, such as:
  - Acoustic Speech Recognition (ASR)
  - Text to Speech Systems (TTS)
  - Dialog Systems / Chat Bots
  - Question-Answering Systems (QA)
  - Named Entity Recognition (NER)
  - Relation Extraction (RE)
  - Text Summarisation
  - Sentiment Analysis

Importance of language data for Language Technology (I)

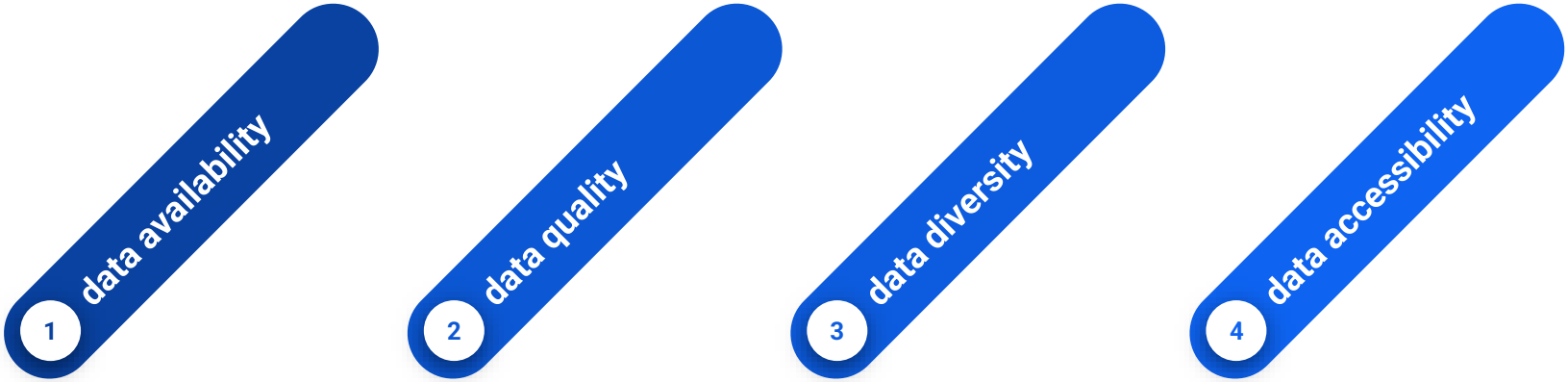
- language - at the heart of communication
- language: source of measurable, analyzable data that drives business, technology, and research insights
- ensures people’s interaction via devices
- is augmented by the use of these devices
- multilingual environment
- globally available

©AI

Importance of language data for Language Technology (II)

- drives advances in language technologies

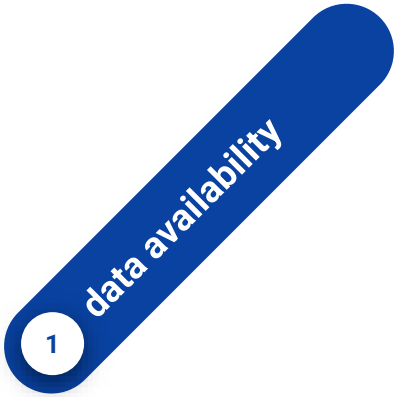
- decisive factors of progress and competitiveness:



Importance of language data for Language Technology (II)

- drives advances in language technologies

- decisive factors of progress and competitiveness:



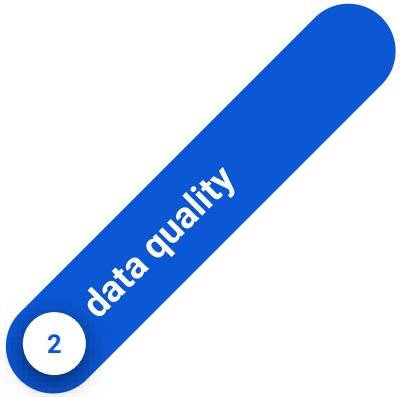
data exists and is ready to use

Importance of language data for Language Technology (II)

- drives advances in language technologies

- decisive factors of progress and competitiveness:

- garbage in, garbage out
- reduce hallucinations, biases



- context-rich data > results accuracy
- domain-specific effectiveness

Importance of language data for Language Technology (II)

- drives advances in language technologies

- decisive factors of progress and competitiveness:

  - accuracy
  - equity
  - global relevance

Importance of language data for Language Technology (II)

- drives advances in language technologies

- decisive factors of progress and competitiveness:

  - data is findable, reusable, understandable, leverageable
  - democratisation of data

RACAI4RO

- corpora
  - written text and speech
  - general and domain specific
  - standard and dialectal
  - mono-, bi-modal and multi-modal
- wordnet
- lexicons
  - general and task-specific
- annotation levels:
  - morphology
  - syntax
  - semantics
  - pragmatics

CoRoLa - the reference Corpus of the Contemporary Romanian Language

- developed as a priority project of the Romanian Academy (2014-2019)
- by:
  - Research Institute for Artificial Intelligence (Bucharest) and
  - Institute of Computer Science (Romanian Academy, the Iași Branch)
- in collaboration with:
  - the philology/linguistics institutes of the Romanian Academy
  - computer science/philology faculties of several Romanian universities
- benefiting from a Humboldt project in partnership with:
  - University of Bucharest and
  - Institute for the German Language (IDS) from Mannheim, Germany

CoRoLa - content

- written texts
  - all functional styles: legal, scientific, journalistic, imaginative, administrative, memoirs, blog posts
  - 4 major domains: society, nature, science, arts & culture
  - 63 subdomains subsumed to the domains: e.g.: society (15), nature (4), science (31), arts & culture (13)
- speech
  - mainly professional recordings from studios
  - read text and relatively spontaneous speech
  - always with the written counterpart

CoRoLa - data clearance and sanitation

- clearance wrt Intellectual Property Rights
- cleaning:
  - removal of meta-textual information:
    - title
    - contents
    - foot-/end-notes
    - figures
    - tables
    - captions
    - etc.
  - recovery of paragraph limits
  - recovery of words hyphenated at the end of the row
  - consistency in the type of diacritics used
  - removal of longer strings of texts in another language

CoRoLa - metadata

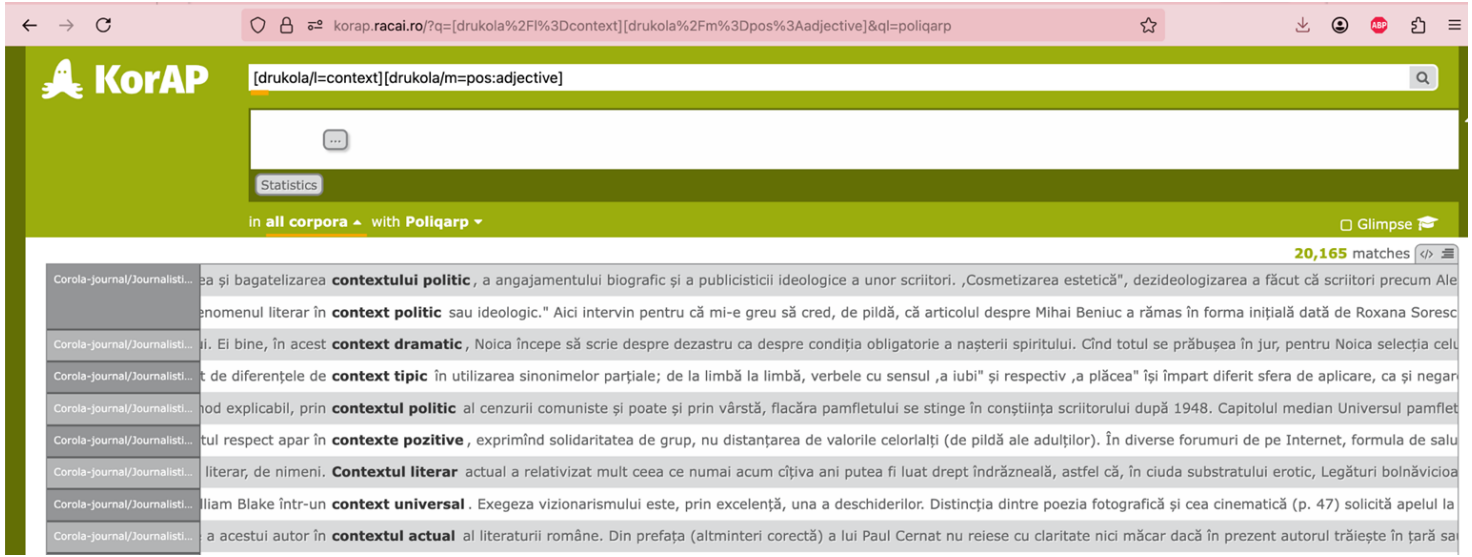
- CMDI format (with some adjustments)

- for each file in the corpus

- manually or automatically (with preliminary manual mapping of categories between the ones in the text source and those decided upon in the corpus)

CoRoLa - access to data (I)

for the written texts:

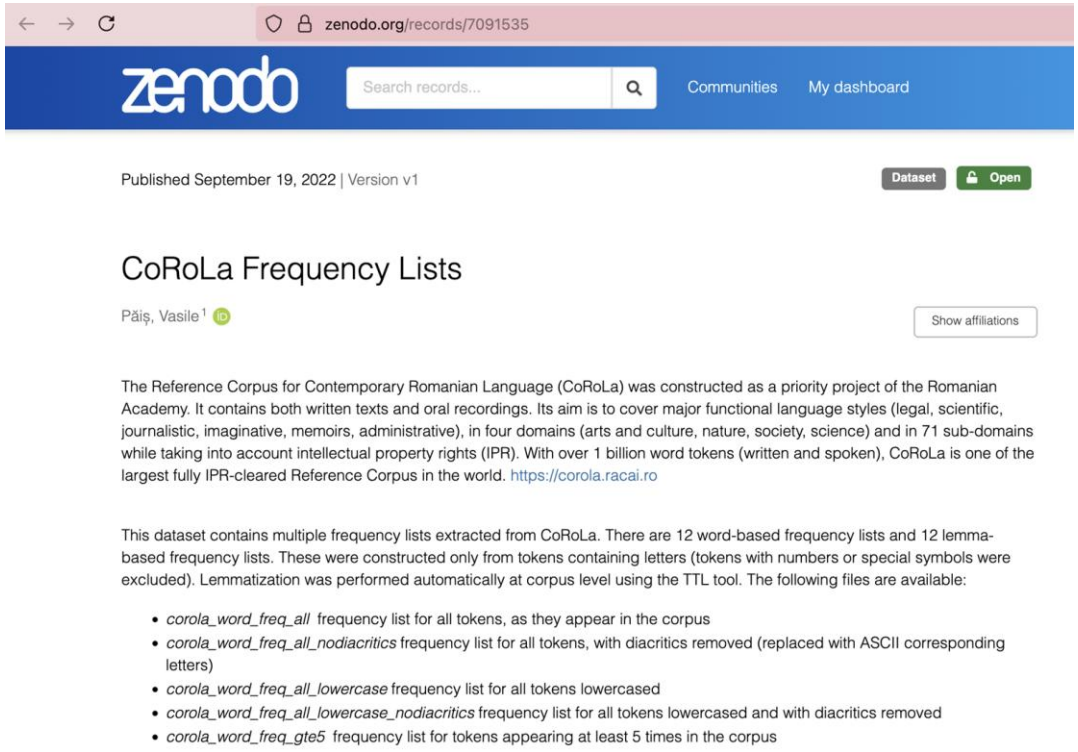


CoRoLa - access to data (II)

for the oral texts:

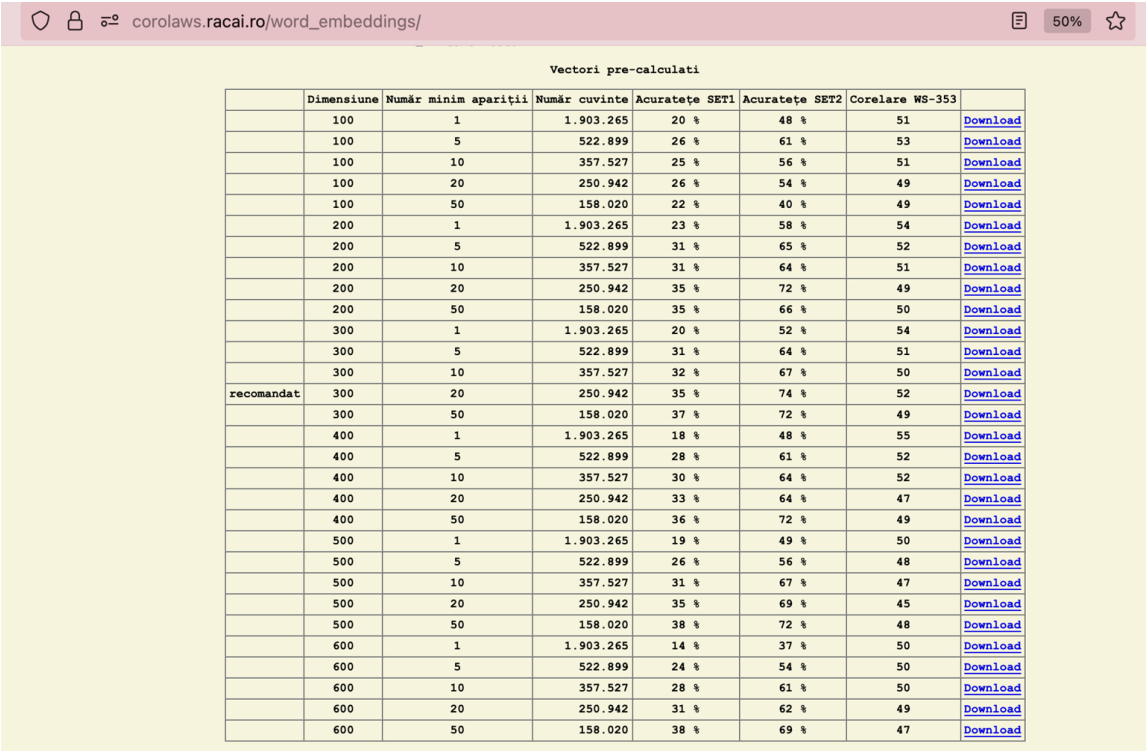
CoRoLa - access to data (III)

frequency lists:



CoRoLa - access to data (IV)

embeddings:



Vectori pre-calculati

	Dimensiune	Număr minim apariții	Număr cuvinte	Acurateţe SET1	Acurateţe SET2	Corelare WS-353	
	100	1	1.903.265	20 %	48 %	51	Download
	100	5	522.899	26 %	61 %	53	Download
	100	10	357.527	25 %	56 %	51	Download
	100	20	250.942	26 %	54 %	49	Download
	100	50	158.020	22 %	40 %	49	Download
	200	1	1.903.265	23 %	58 %	54	Download
	200	5	522.899	31 %	65 %	52	Download
	200	10	357.527	31 %	64 %	51	Download
	200	20	250.942	35 %	72 %	49	Download
	200	50	158.020	35 %	66 %	50	Download
	300	1	1.903.265	20 %	52 %	54	Download
	300	5	522.899	31 %	64 %	51	Download
	300	10	357.527	32 %	67 %	50	Download
recomandat	300	20	250.942	35 %	74 %	52	Download
	300	50	158.020	37 %	72 %	49	Download
	400	1	1.903.265	18 %	48 %	55	Download
	400	5	522.899	28 %	61 %	52	Download
	400	10	357.527	30 %	64 %	52	Download
	400	20	250.942	33 %	64 %	47	Download
	400	50	158.020	36 %	72 %	49	Download
	500	1	1.903.265	19 %	49 %	50	Download
	500	5	522.899	26 %	56 %	48	Download
	500	10	357.527	31 %	67 %	47	Download
	500	20	250.942	35 %	69 %	45	Download
	500	50	158.020	38 %	72 %	48	Download
	600	1	1.903.265	14 %	37 %	50	Download
	600	5	522.899	24 %	54 %	50	Download
	600	10	357.527	28 %	61 %	50	Download
	600	20	250.942	31 %	62 %	49	Download
	600	50	158.020	38 %	69 %	47	Download

CoRoLa - continuous enrichment (I)

enrichment of the oral component with new recordings

2017-2020
Funded by UEFISCDI,
Romanian Ministry of Ed.
and Research

CoRoLa - continuous enrichment (II)

DIGITAL-2024-AI-06-LANGUAGE-01
Alliance for Language Technologies

LLMs4EU

alt-edic

Funded by the European Union

- Development of LLMs in European languages
- Evaluation of LLMs in European languages
- Development of specific use-cases

This is an opportunity:

- to create new Romanian language resources
- compute statistics on Romanian resources
- aid in the evaluation of LLMs using Romanian resources (text and multimodal)

CoRoLa - continuous enrichment (II)

DIGITAL-2024-AI-06-LANGUAGE-01
Alliance for Language Technologies

LLMs4EU

alt-edic

Funded by
the European Union

CoRoLa - continuous enrichment (III)

Automatic detection of AI-generated texts from Moldova and Romania (ADAMo)

- One of its objectives: **Enriching CoRoLa with texts originating from the Republic of Moldova**

CoRoLa - continuous enrichment (IV)

Defending against deep fake news with large language and image models - **DeepNewsDef**

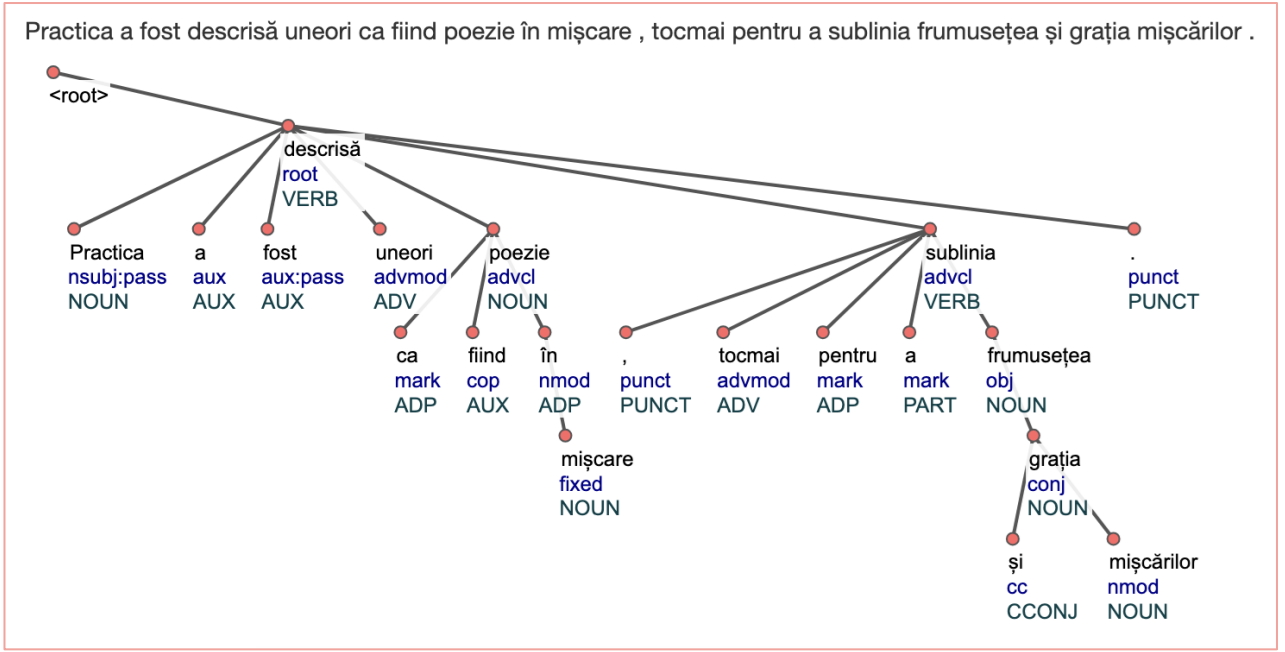
- will enrich CoRoLa with data from the News domain in the two countries
- will employ existing data for generated text detection

https://www.racai.ro/p/deepnewsdef/

Treebanks - Universal Dependencies compliant

- morphological description
- syntactic relations between words in a sentence
- labelling of syntactic relations

- Romanian Reference Treebank (RoRefTrees – **RRT**) - general language
- core of a treebank for the Aromanian dialect (**ArT**)



Semantically annotated corpora

- multiword expressions: **PARSEME-Ro** v. 2.0
  - all parts of speech
- named entities
  - medical: **SiMoNERo**:
    - ANAT (body parts)
    - CHEM (Chemicals and Drugs)
    - DISO (disorders)
    - PROC (procedures)
  - legal: **LegalNERo**:
    - LEGAL (legal references)
    - PER (person names)
    - LOC (locations)
    - ORG (organization names)
    - TIME (time)

Pragmatically annotated corpus

- RRT
- discourse relations expressed by conjunctional locutions
- compliant with Penn Discourse Treebank v. 3.0

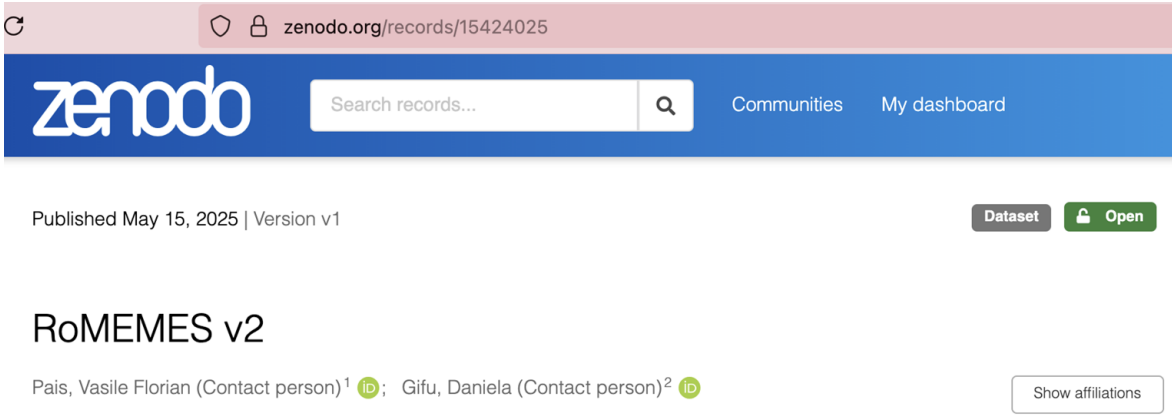
Sense	Total No.	No.	Connectives (count)
TEMPORAL:SYNCHRONOUS	40	29	în timp ce
		9	pe măsură ce
		2	de îndată ce
TEMPORAL:ASYNCHRONOUS:PRECEDENCE	49	35	înainte de
		9	înainte ca
		5	până când
TEMPORAL:ASYNCHRONOUS:SUCCESSION	54	48	după ce
		1	pentru ca să
		5	de îndată ce
TEMPORAL:ASYNCHRONOUS:SUCCESSION— CONTINGENCY:CAUSE:REASON	38	38	după ce
CONTINGENCY:CAUSE:REASON	39	35	pentru că
		4	pe măsură ce
CONTINGENCY:CAUSE:RESULT	31	2	pentru că
		7	astfel încât
		7	așa încât
		7	așa că
		8	astfel că
CONTINGENCY.CAUSE.NEGRESULT	1	1	pentru ca să
CONTINGENCY:CAUSE+BELIEF:REASON+BELIEF	10	10	pentru că
CONTINGENCY:CAUSE+SPEECHACT:REASON+SPEECHACT	4	4	pentru că
CONTINGENCY:CAUSE+SPEECHACT:RESULT+SPEECHACT	3	3	pentru că
CONTINGENCY:CONDITION:ARG1-AS-COND	1	1	pentru ca să
CONTINGENCY:CONDITION:ARG2-AS-COND	3	3	pe măsură ce
CONTINGENCY:PURPOSE:ARG2-AS-GOAL	54	30	pentru ca să
		20	astfel încât
		2	așa încât
		2	așa că
COMPARISON:CONCESSION:ARG1-AS-DENIER	21	13	chiar dacă
		8	cu toate că
COMPARISON:CONCESSION:ARG2-AS-DENIER	2	2	în timp ce
COMPARISON:CONTRAST	18	18	în timp ce
COMPARISON:SIMILARITY	1	1	în timp ce
EXPANSION:MANNER:ARG2-AS-MANNER	9	9	fără să

Underrepresented Speech Corpus

- manually created
- focus on speech types that are underrepresented in other speech datasets
- Sources: open data available on multimedia platforms under a Creative Commons license
- manually transcribed and aligned at segment level
- text and audio files + text annotations (lemmatization, part of speech tags, dependency parsing) in CoNLL-U Plus format.

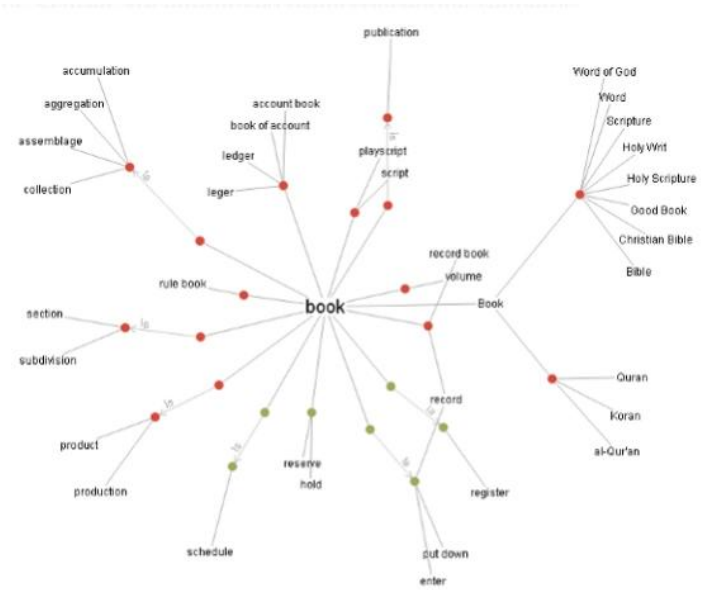
More than a thousand words: RoMEMES

- a dataset of Romanian language memes
- collected from public social media platforms
- manually annotated with:
  - associated text in Romanian language;
  - image complexity;
  - polarity;
  - sentiment;
  - political content



Romanian Wordnet

- lexical knowledge graph
- nodes: synonym sets
- edges: semantic relations
- alignment with Princeton WordNet and to other wordnets



Lexicons

- general linguistic information: RoLEX
- specialized: RoMWE lexicon

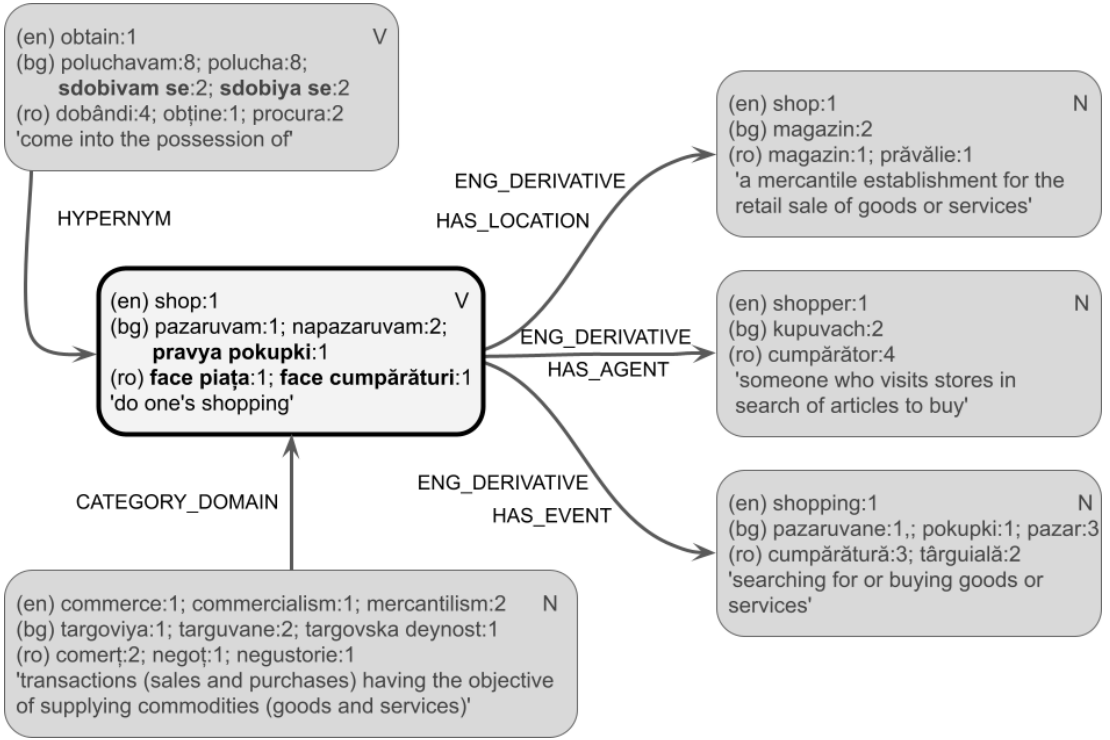
Lexicons

- general linguistic information: RoLEX

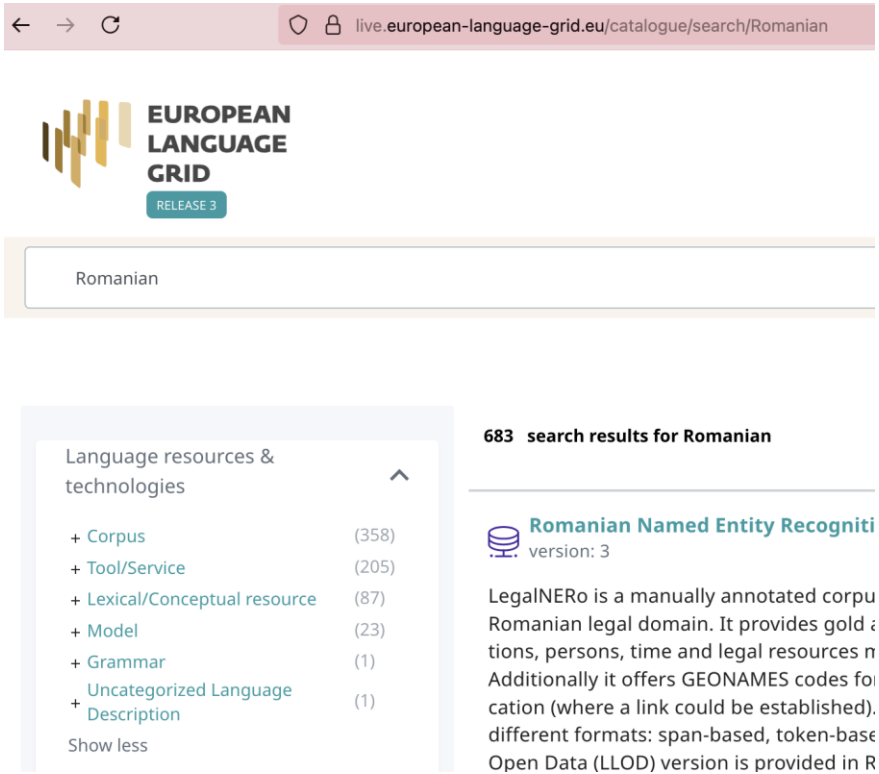
```
298386 turmă » = » Ncfsrn » tur.mă » t'urmă » t u r m @
298387 turnă » turna » Vmis3s » tur.nă » turn'ă » t u r n @
298388 turnăm » turna » Vmip1p » tur.năm » turn'ăm » t u r n @ m
298389 turnări » turnare » Ncfp-n » tur.năr i » turn'ări » t u r n @ r i_0
298390 turnări » turnare » Ncfson » tur.năr i » turn'ări » t u r n @ r i_0
298391 turnării » turnare » Ncfsoy » tur.nă.rii » turn'ării » t u r n @ r i j
298392 turnările » turnare » Ncfpry » tur.nă.ri.le » turn'ările » t u r n @ r i l e
298393 turnărilor » turnare » Ncfpoy » tur.nă.ri.lor » turn'ărilor » t u r n @ r i l o r
298394 turnătoare » turnător » Afpf--n » tur.nă.toa.re » turnăto'are » t u r n @ t o_X a r e
298395 turnătoarea » turnător » Afpfsry » tur.nă.toa.rea » turnăto'area » t u r n @ t o_X a r e_X a
298396 turnătoarei » turnător » Afpfsoy » tur.nă.toa.rei » turnăto'arei » t u r n @ t o_X a r e j
298397 turnătoarele » turnător » Afpfpry » tur.nă.toa.re.le » turnăto'arele » t u r n @ t o_X a r e l e
298398 turnătoarelor » turnător » Afpfpoy » tur.nă.toa.re.lor » turnăto'arelor » t u r n @ t o_X a r e l o r
298399 turnător » = » Afpms-n » tur.nă.tor » turnăt'or » t u r n @ t o r
298400 turnător » = » Ncms-n » tur.nă.tor » turnăt'or » t u r n @ t o r
298401 turnători » turnător » Afpmp-n » tur.nă.tori » turnăt'ori » t u r n @ t o r i_0
298402 turnători » turnător » Ncmp-n » tur.nă.tori » turnăt'ori » t u r n @ t o r i_0
298403 turnătoria » turnătorie » Ncfsry » tur.nă.to.ri.a » turnător'ia » t u r n @ t o r i a
298404 turnătorie » = » Ncfsrn » tur.nă.to.ri.e » turnător'ie » t u r n @ t o r i e
298405 turnătoriei » turnătorie » Ncfsoy » tur.nă.to.ri.ei » turnător'iei » t u r n @ t o r i e j
298406 turnătorii » turnător » Afpmpry » tur.nă.to.rii » turnăt'orii » t u r n @ t o r i j
298407 turnătorii » turnător » Ncmpry » tur.nă.to.rii » turnăt'orii » t u r n @ t o r i j
298408 turnătorii » turnătorie » Ncfp-n » tur.nă.to.rii » turnător'ii » t u r n @ t o r i j
298409 turnătorii » turnătorie » Ncfson » tur.nă.to.rii » turnător'ii » t u r n @ t o r i j
298410 turnătoriile » turnătorie » Ncfpry » tur.nă.to.ri.i.le » turnător'iile » t u r n @ t o r i i l e
298411 turnătoriilor » turnătorie » Ncfpoy » tur.nă.to.ri.i.lor » turnător'iilor » t u r n @ t o r i i l o r
298412 turnătorilor » turnător » Afpmpoy » tur.nă.to.ri.lor » turnăt'orilor » t u r n @ t o r i l o r
298413 turnătorilor » turnător » Ncmpoy » tur.nă.to.ri.lor » turnăt'orilor » t u r n @ t o r i l o r
298414 turnătorul » turnător » Afpmsry » tur.nă.to.rul » turnăt'orul » t u r n @ t o r u l
298415 turnătorul » turnător » Ncmsry » tur.nă.to.rul » turnăt'orul » t u r n @ t o r u l
298416 turnătorului » turnător » Afpmsoy » tur.nă.to.ru.lui » turnăt'orului » t u r n @ t o r u l u j
298417 turnătorului » turnător » Ncmsoy » tur.nă.to.ru.lui » turnăt'orului » t u r n @ t o r u l u j
298418 turn » = » Ncms-n » turn » turn » t u r n
298419 turna » = » Vmii3s » tur.na » turn'a » t u r n a
298420 turna » = » Vmnp » tur.na » turn'a » t u r n a
298421 turnași » turna » Vmis2s » tur.nași » turn'ași » t u r n a S i_0
298422 turnați » turna » Vmii2p » tur.nați » turn'ați » t u r n a ts i_0
298423 turnați » turna » Vmip2p » tur.nați » turn'ați » t u r n a ts i_0
298424 turnați » turna » Vmm-2p » tur.nați » turn'ați » t u r n a ts i_0
298425 turnați » turna » Vmp--pm » tur.nați » turn'ați » t u r n a ts i_0
```

Lexicons

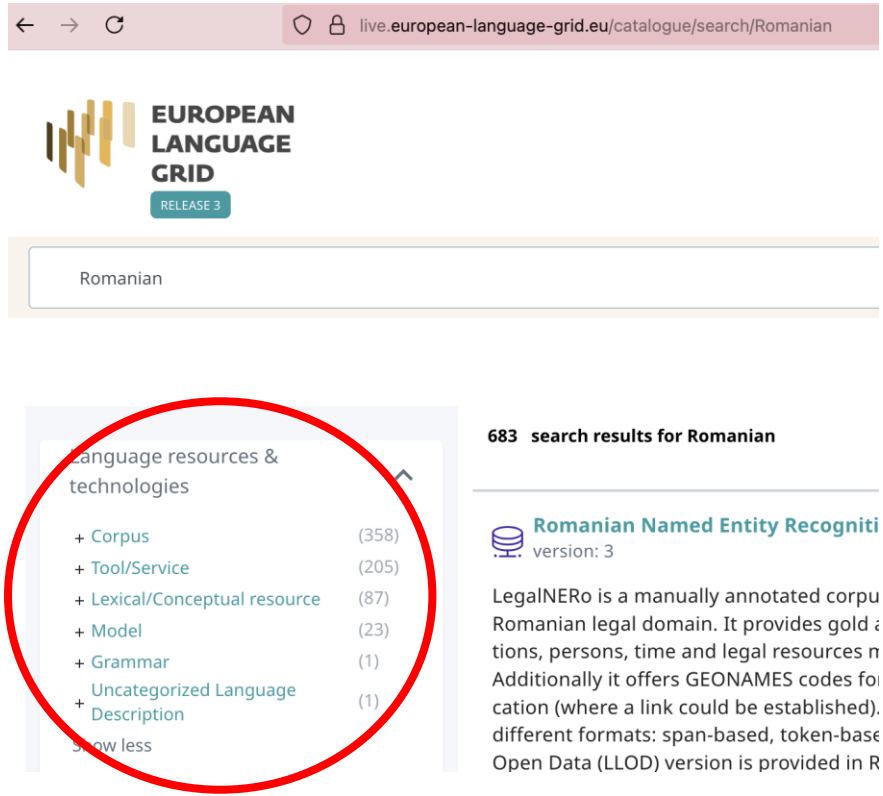
- specialized: RoMWE lexicon



The landscape of language resources for Romanian

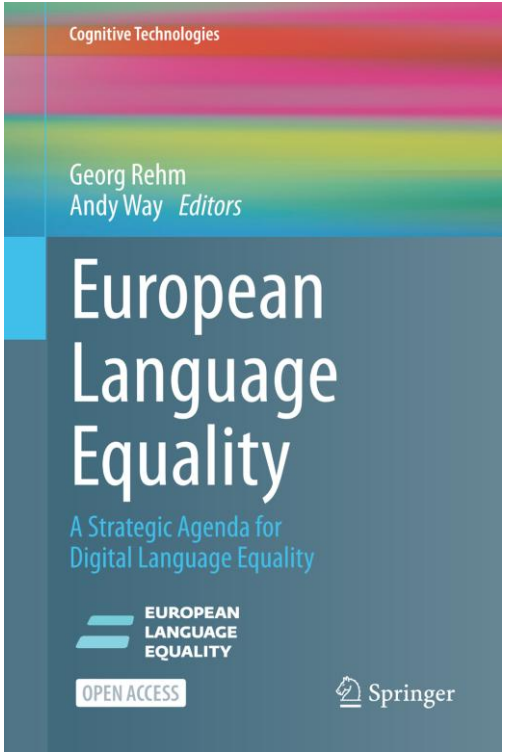


The landscape of language resources for Romanian



Gaps in the resources for Romanian

- more focus is needed for creating **domain-specific resources**
- need for resources rendering **language varieties** and their characteristics



Issues with data

- privacy
- intellectual property
- data governance: set the rules for:
  - collection
  - storage
  - sharing
  - maintaining
  - responsibility
- need for transparency in collection and annotation => trust in data
- sustainable data strategies for future (high-performance) LTs:
  - technological scalability
  - ethical stewardship (responsible, fair, transparent, and respectful of human rights and dignity)
  - linguistic expertise

European strategy for data (I)

- The European strategy for data aims at creating a single market for data that will ensure Europe's global competitiveness and data sovereignty

- New policies are being implemented to ensure that Europe is at an international competitive level, which brings us to work closer with Europe’s pillars: its countries, and our Country Workshop today

- The economic and societal impact of LT, AI and language data are putting pressure on the European industry not to stay behind

- However, this requires extremely large amounts of data for our industry and research institutions to advance

European strategy for data (II)

- It is crucial that private sector data holders and producers have the will to **share their data** and the European strategy for data facilitates this by means of mechanisms allowing to monetise such data: the European data spaces

- “Data is an essential resource for economic growth, competitiveness, innovation, job creation and societal progress in general” (cf. https://digital-strategy.ec.europa.eu/en/policies/strategy-data)

- What next?
  - The European Language Data Space can relieve the procedure of data acquisition and provision through its unique platform of language resources, language models and services
  - Data holders can approach the LDS to see how to make use of it and how to benefit from it (stay for the LDS presentation this morning!).

Common European Language Data Space

Thank you!

Verginica Barbu Mititelu (Romanian Academy Research Institute for Artificial Intelligence, Romania)
vergi@racai.ro

A Common European Language Data Space – funded under contract LC-01936389 with the European Union.

22-10-2024 LDS Country Workshop, Cluj-Napoca, Romania
https://language-data-space.ec.europa.eu