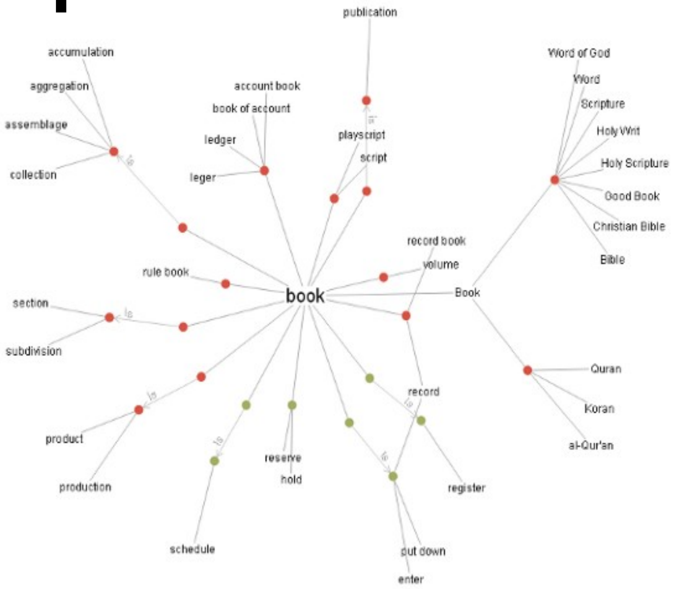
Innovation in language resources development

Water flows, rocks remain.
(Romanian proverb)

Major resources development



Romanian WordNet

– lexical ontology

– 60,000 synsets

– aligned with Princeton WordNet for English and thus with other wordnets aligned to it: e.g. Hindi, Assamese, Bengali, Boro, etc.

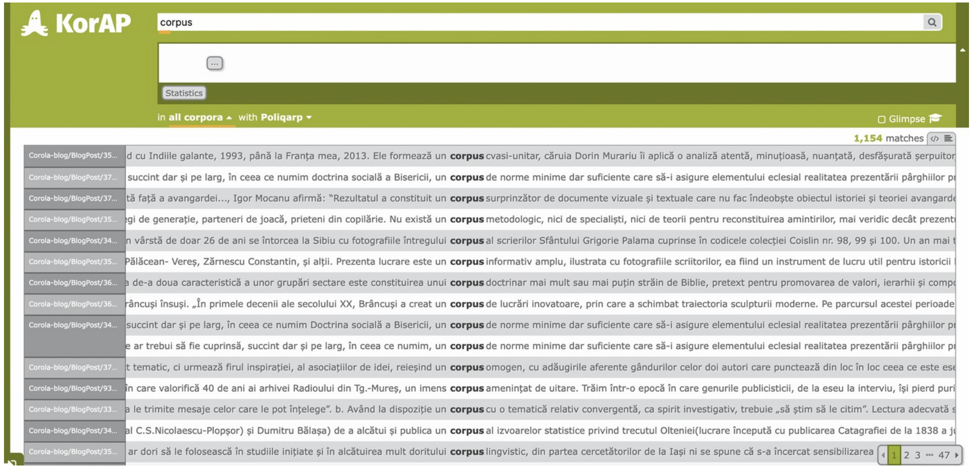
Corpora: CoRoLa

- Priority project of the Romanian Academy
- Partners:
  - Institute of Computer Science
  - University of Bucharest, Institute for the German Language
  - Texts providers
- 1.2 mill. words
- Bimodal: written, spoken
- diverse wrt styles, domains, subdomains
- several annotation levels
- metadata
- exploitation: KorAP (manual), OCQP (oral component), word embeddings (for downstream applications)

Star of the Carpathians

Other corpora

- Oral: ROBINTASC
- NER: LegalNERo, SiMoNERo, MARCELL



Corola-blog/BlogPost/35...	d cu Indiile galante, 1993, până la Franța mea, 2013. Ele formează un	**corpus** cvasi-unitar, căruia Dorin Murariu îi aplică o analiză atentă, minuțioasă, nuanțată, desfășurată șerpuitor
Corola-blog/BlogPost/37...	succint dar și pe larg, în ceea ce numim doctrina socială a Bisericii, un	**corpus** de norme minime dar suficiente care să-i asigure elementului eclesial realitatea prezentării pârghiilor pr
Corola-blog/BlogPost/37...	tă față a avangardei..., Igor Mocanu afirmă: "Rezultatul a constituit un	**corpus** surprinzător de documente vizuale și textuale care nu fac îndeobște obiectul istoriei și teoriei avangarde
Corola-blog/BlogPost/35...	egi de generație, parteneri de joacă, prieteni din copilărie. Nu există un	**corpus** metodologic, nici de specialiști, nici de teorii pentru reconstituirea amintirilor, mai veridic decât prezent
Corola-blog/BlogPost/34...	n vârstă de doar 26 de ani se întorcea la Sibiu cu fotografiile întregului	**corpus** al scrierilor Sfântului Grigorie Palama cuprinse în codicele colecției Coislin nr. 98, 99 și 100. Un an mai t
Corola-blog/BlogPost/35...	Pălăcean- Vereș, Zărnescu Constantin, și alții. Prezenta lucrare este un	**corpus** informativ amplu, ilustrata cu fotografiile scriitorilor, ea fiind un instrument de lucru util pentru istoricii
Corola-blog/BlogPost/36...	a de-a doua caracteristică a unor grupări sectare este constituirea unui	**corpus** doctrinar mai mult sau mai puțin străin de Biblie, pretext pentru promovarea de valori, ierarhii și comp
Corola-blog/BlogPost/36...	râncuși însuși. „În primele decenii ale secolului XX, Brâncuși a creat un	**corpus** de lucrări inovatoare, prin care a schimbat traiectoria sculpturii moderne. Pe parcursul acestei perioade
Corola-blog/BlogPost/34...	succint dar și pe larg, în ceea ce numim Doctrina socială a Bisericii, un	**corpus** de norme minime dar suficiente care să-i asigure elementului eclesial realitatea prezentării pârghiilor pr
	e ar trebui să fie cuprinsă, succint dar și pe larg, în ceea ce numim, un	**corpus** de norme minime dar suficiente care să-i asigure elementului eclesial realitatea prezentării pârghiilor pr
Corola-blog/BlogPost/37...	t tematic, ci urmează firul inspirației, al asociațiilor de idei, reieșind un	**corpus** omogen, cu adăugirile aferente gândurilor celor doi autori care punctează din loc în loc ceea ce este ese
Corola-blog/BlogPost/93...	în care valorifică 40 de ani ai arhivei Radioului din Tg.-Mureș, un imens	**corpus** amenințat de uitare. Trăim într-o epocă în care genurile publicisticii, de la eseu la interviu, își pierd puri
Corola-blog/BlogPost/33...	a le trimite mesaje celor care le pot înțelege". b. Având la dispoziție un	**corpus** cu o tematică relativ convergentă, ca spirit investigativ, trebuie „să știm să le citim". Lectura adecvată s
Corola-blog/BlogPost/34...	al C.S.Nicolaescu-Plopşor) și Dumitru Bălașa) de a alcătui și publica un	**corpus** al izvoarelor statistice privind trecutul Olteniei(lucrare începută cu publicarea Catagrafiei de la 1838 a ju
Corola-blog/BlogPost/35...	ar dori să le folosească în studiile inițiate și în alcătuirea mult doritului	**corpus** lingvistic, din partea cercetătorilor de la Iași ni se spune că s-a încercat sensibilizarea

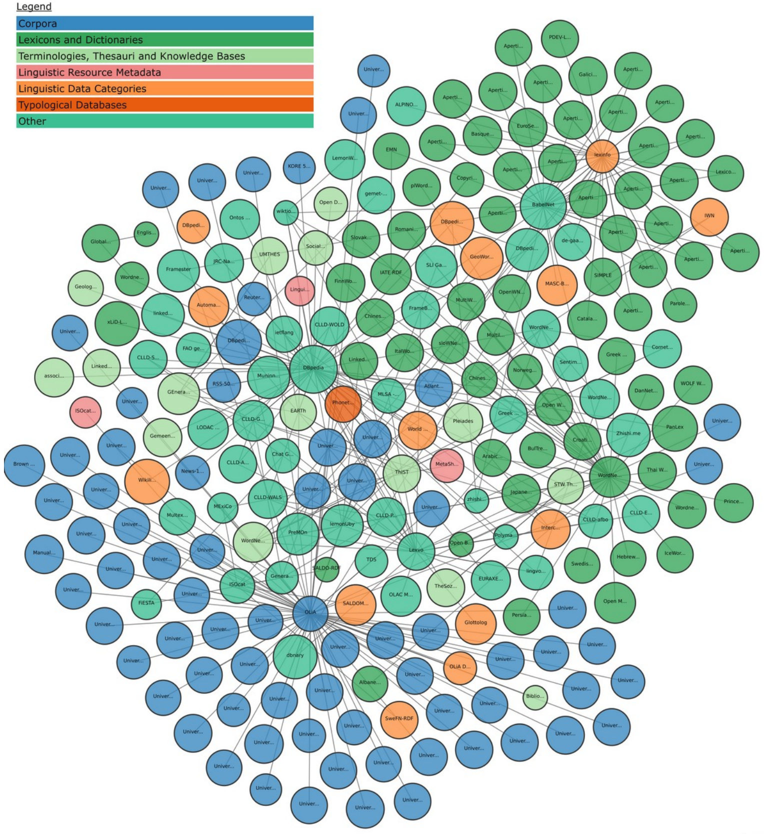
- Multiple annotation layers: lexical, morphologic, syntactic, phonetic
- Exploited in various tasks: anonymisation, machine translation, etc.

RoLLOD 2020-2023

ROMANIAN LANGUAGE RESOURCES CONVERTED TO *LINKED DATA* SPECIFICATIONS

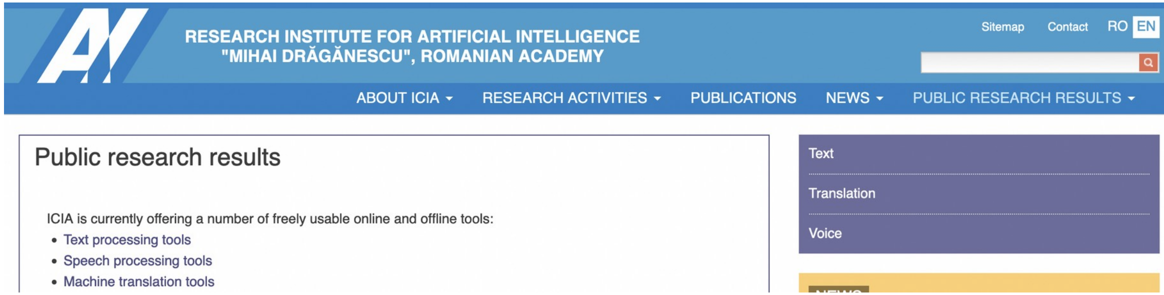
Standardization

- Linked Data format
- Nexus Linguarum COST Action
- Data FAIRness:
  - Findable
  - Accessible
  - Interoperable
  - Reusable


Legend
Corpora
Lexicons and Dictionaries
Terminologies, Thesauri and Knowledge Bases
Linguistic Resource Metadata
Linguistic Data Categories
Typological Databases
Other

The Linguistic Linked Open Data Cloud from lod-cloud.net

Access



RESEARCH INSTITUTE FOR ARTIFICIAL INTELLIGENCE
"MIHAI DRĂGĂNESCU", ROMANIAN ACADEMY

Public research results

ICIA is currently offering a number of freely usable online and offline tools:

- Text processing tools
- Speech processing tools
- Machine translation tools

- Free
- Sometimes limitations: CoRoLa
- Metadata available in major European Language Technologies hubs:
  - META-SHARE
  - European Language Grid
  - Linked Open Data Cloud
- Data dump available

EUROPEAN LANGUAGE GRID

European Language Resource Coordination
Connecting Europe Facility

META-SHARE