Research Institute for Artificial Intelligence ”Mihai Drăgănescu” Romanian Academy

Language-Centered AI

Dan Tufiș, Radu Ion, Verginica Barbu-Mititelu, Vasile Păiș, Ștefan Trăușan-Matu,
Eduard Franți, Andrei Brătan

General Information

- The ICIA Institute was founded, at the proposal of Academician Mihai Drăgănescu in 2002, on the structure of the Center for Advanced Research in Automatic Learning, Natural Language Processing and Conceptual Modeling existing since 1994 in the structure of the Romanian Academy. Since then, the institute has excelled in the fields of Natural Language Processing and Machine Learning, becoming the flagship Romanian institution in the automatic processing of documents in the Romanian language.
- In absolutely all annual evaluations, starting from 1994 until now, the Center and then the Institute received the qualification "excellence”;
- It was, and still is, the reference partner in EC projects aimed at eliminating language barriers in the EU: ICIA is the "Technological Anchor" for Romania in ELRC (European Language Resource Coordination), Competence Center for Romania in the large European Language Grid projects and in European Language Equality.

General Information (cntd)

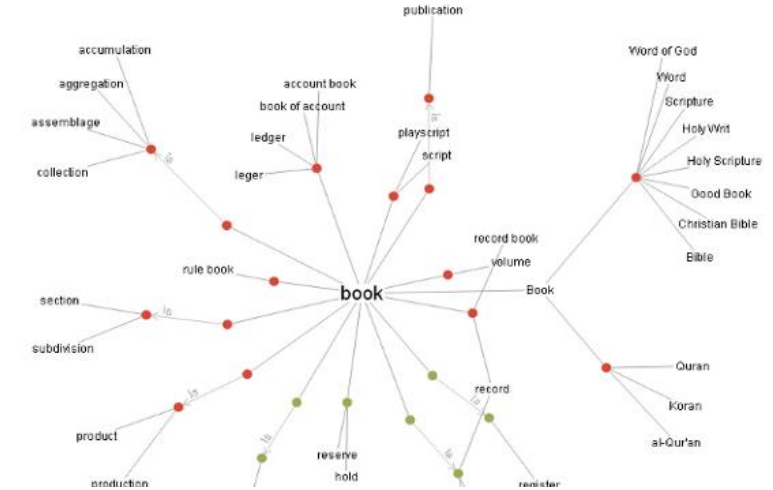
- ICIA is the depository of the most important resources and tools for the automatic processing of the Romanian language and the main contributor for the Romanian language to the European platforms META-SHARE, ELRC-SHARE, ELG, ELE and LDS. We organized, with support from EC, 7 national workshops for these platforms and a next one is planned for 2024.
- The tools for automatic processing of the Romanian language won most of the international competitions in which ICIA researchers participated (NAACL, 2003 – Edmonton, Canada, ACL, 2005 – Ann Arbor, USA, CLEF 2007, 2008, 2010, etc.).
- ICIA is an organizing institution for doctoral studies.
- Co-organizer of high-impact international scientific events (EUROLAN – 16 editions, CONSILR – 17 editions, SPED – 11 editions).
- An impressive list of publications: over 1100 books, articles in journals or in the volumes of international conferences (https://www.racai.ro/publications/).

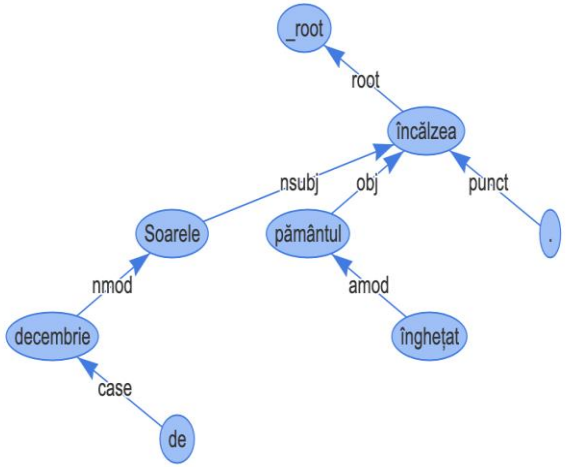
Major resources development (1)

- General corpora with standardized processing:
  - 1998- The Romanian translation of the novel 1984 by G. Orwell and aligning it with the original.
  - 2006 – JRC-Acquis-Ro aligned with the English version
  - 2012 – ROMBAC – the first balanced contemporary language corpus, containing texts in 5 approximately equally represented fields: journalistic, medical and pharmaceutical, legal, philological and fiction
  - 2013-2020 – CoRoLa – the representative corpus for the contemporary Romanian language (4 large domains, over 50 subdomains, over 1.2 billion lexical items) also includes oral texts, metadata for all included documents
- Specialized corpora:
  - RRT, LegalNERo, SiMoNERo, BioRo, MARCELL, PARSEME-RO, CURLICAT, Microblogging, ROBINTASC, USPDATRo
- Dictionaries and Lexicons:
  - WebDex – Implementation of DEX 1996, forerunner of DEXONLINE
  - tbl.wordform – over 1.2 million entries of type inflected form|lema|MSD
  - ROLEX – the most extensive validated phonological lexicon available for the Romanian language (330,866 entries)

Major resources development(2)

- Ro-Wordnet lexical ontology
  - 60,000 synsets, over 85,000 words aligned with Princeton Wordnet, provides navigation in all other ontologies aligned with Princeton Wordnet
- Parse tree banks:
  - General
    - RRT (Romanian Reference Treebank)
  - Specialized
    - SiMoNERo medical field
- All language resources have different standardized representations, created in recent years (e.g. Link Data Format) in European projects (Nexus Linguarum)



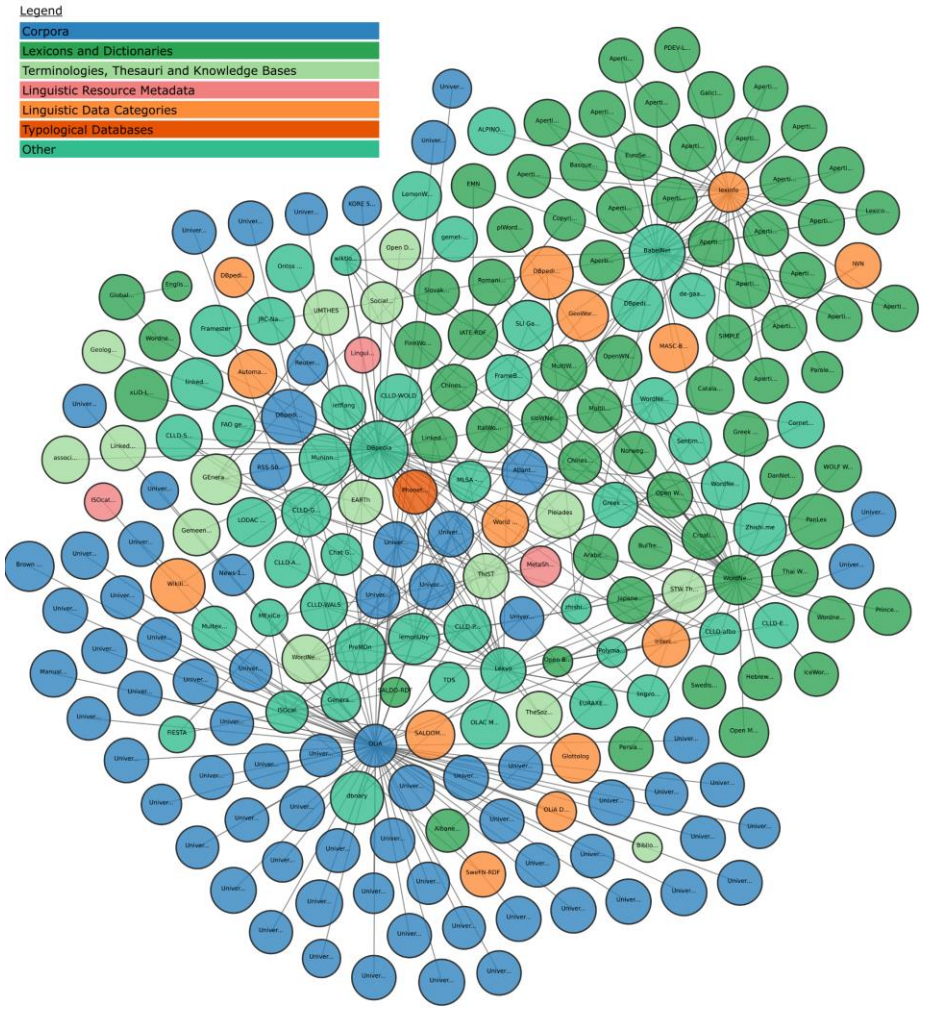


RoLLOD 2020-2023

ROMANIAN LANGUAGE RESOURCES CONVERTED TO *LINKED DATA* SPECIFICATIONS

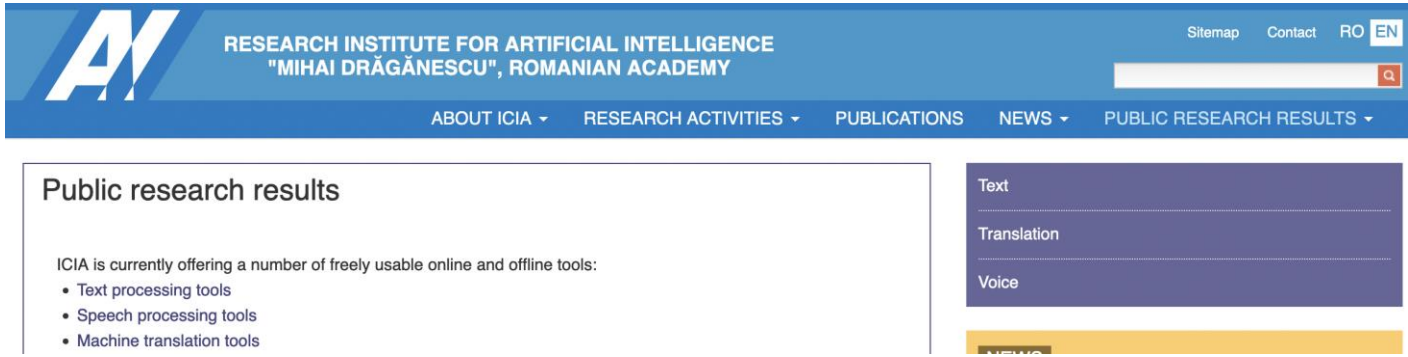
Standardization

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



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

Access



- 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

UniDive

- - an interdisciplinary scientific network devoted to universality, diversity and idiosyncrasy in language technology
- - main objective: reconcile language diversity with rapid progress in language technology
- - both inter- and intra-language diversity, i.e. a diversity understood both in terms of the differences among the existing languages and of the variety of linguistic phenomena exhibited within a language

Aims:
To prepare language researchers for what is coming;
To facilitate longer term dialogue between linguists and technology developers.

The main tools for processing the Romanian language (1)

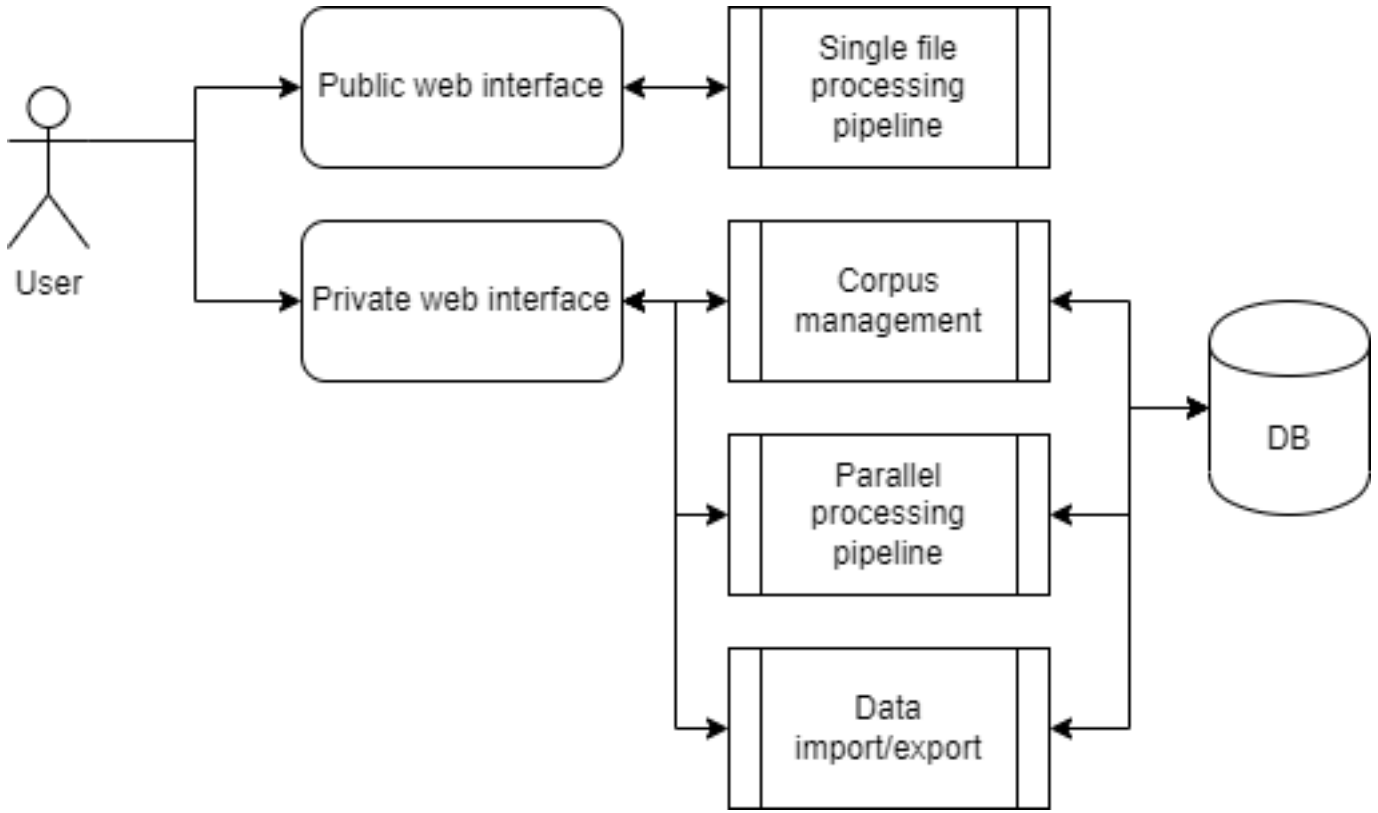
- Pre-trained LLMs (Large Language Models) for the Romanian language BERT-like (Bidirectional Encoder Representation Transformers) :
  -RoBERT: Two models bert-base-romanian-cased-v1 and bert-base-romanian-uncased-v1.
  -Romanian DistilBERT: Constructed based on the bert-base-romanian-cased-v1 model, it is available on HuggingFace as distilbert-base-romanian-cased.
  - A Lite Romanian BERT: ALR-BERT (https://huggingface.co/datasets/dragosnicolae555/RoITD)
  - CoRoLA-based small LLM (https://github.com/racai-ai/ro-corola-bert-small
- Vector representations (word embeddings) generated from: CoRoLa, Bioro
- The Chatbot for the Doctoral School of the Romanian Academy (SCOSAAR) - an application of the outcome of the European project Enrich4All (https://www.enrich4all.eu/).

The main tools for processing the Romanian language (2)

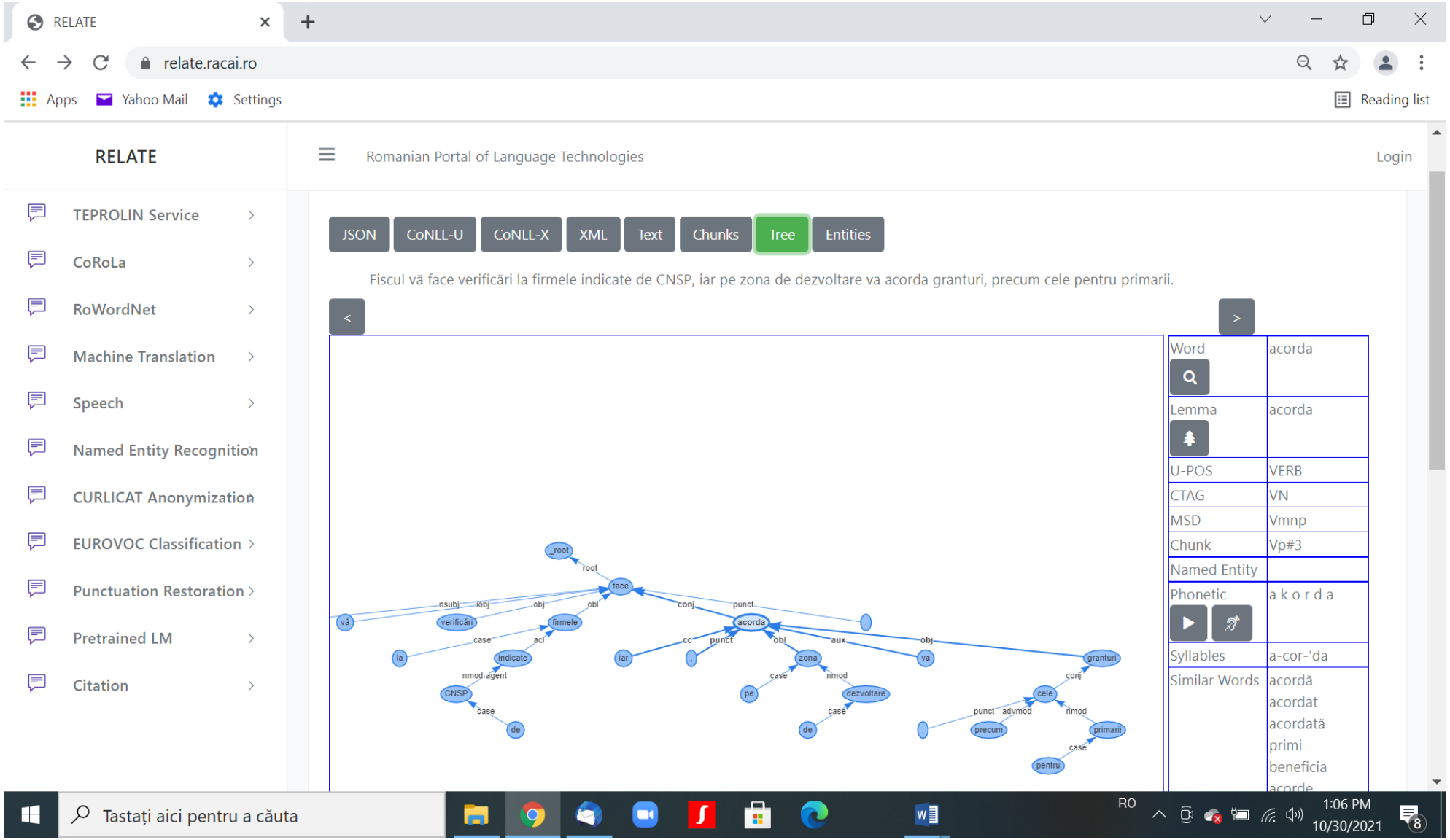
- RELATE – the portal of resources and processing tools for the Romanian language (www.relate.ro):
  - TEPROLIN – a fully configurable flow of primary processing of a text
  - RODNA (Romanian Deep Neural Network Architectures) is a Python 3/TensorFlow /Keras project and includes high-performance, essential modules specifically targeted at Romanian text processing (sentence splitter, tokenizer, morphology analyzer, POS tagger, dependency parser)
  - Interfaces to CoRoLa, RoWordNet, to the translation system (Ro-En-Ro) developed for Romania's presidency of the Council of Europe
  - RO-EN and EN-RO voice translation, various voice signal processing modules (ASR, TTS)
  - Classifier of documents according to EUROVOC classification

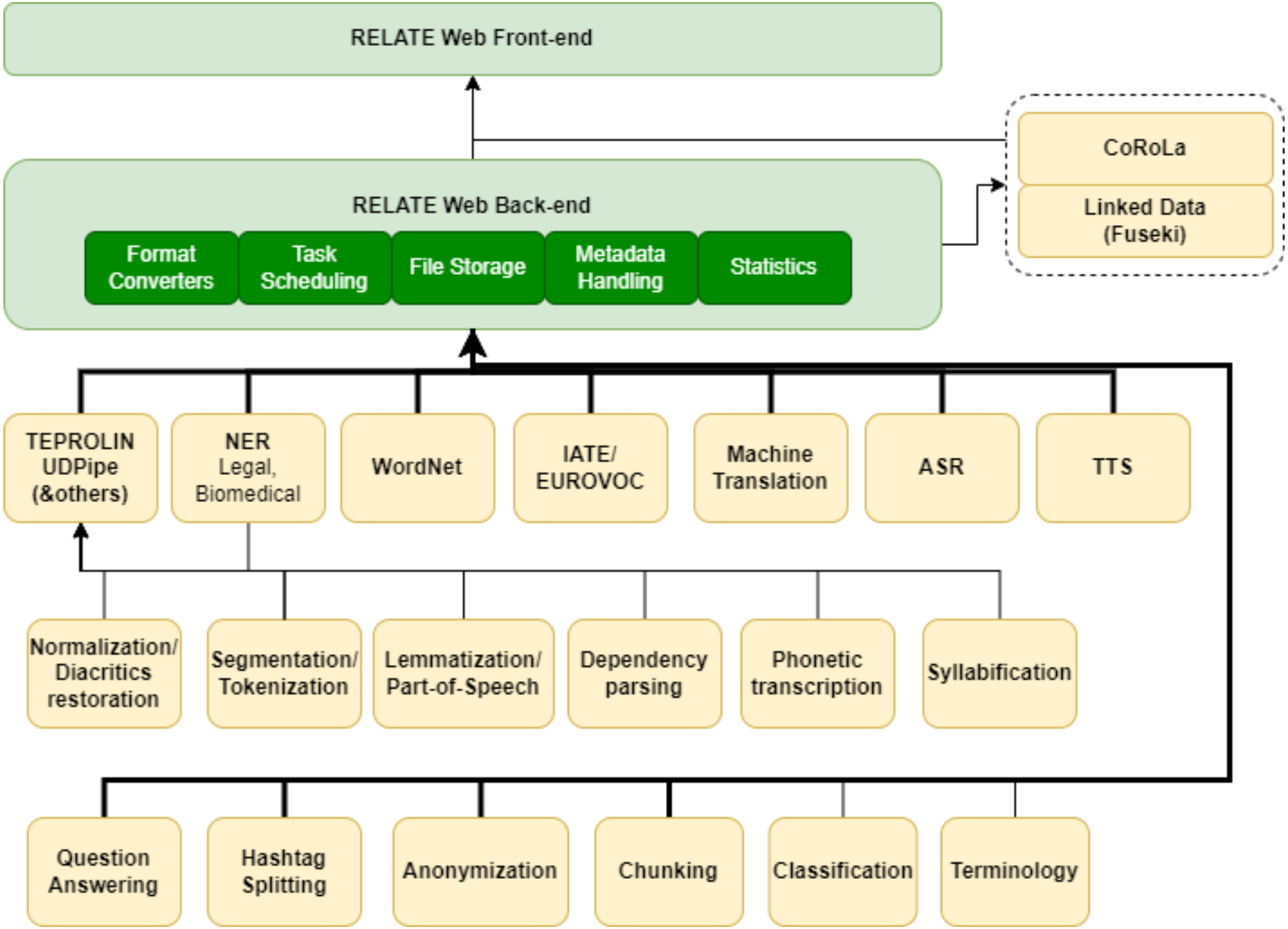
The main tools for processing the Romanian language (3)

- RELATE – the portal of resources and processing tools for the Romanian language (www.relate.ro):
  - The anonymization module built in the CURLICAT project, the recognition of named entities that can be subject to anonymization, the punctuation restoration module, the question-answer module.
  - A more powerful anonymization system (SAROJ) used to anonymize the content of the Romanian jurisprudence database (a project funded by the Council of Europe for the benefit of the Romanian Superior Council of Magistracy
  - The portal offers access to various tools and corpora created at ICIA but also in other European research groups (Resources and Models/Repository)
  - it follows the ELG implementation philosophy:
    - Web services, REST APIs, dockers
    - The services may be distributed over different physical servers/nodes
    - The services may be consumed directly from partners

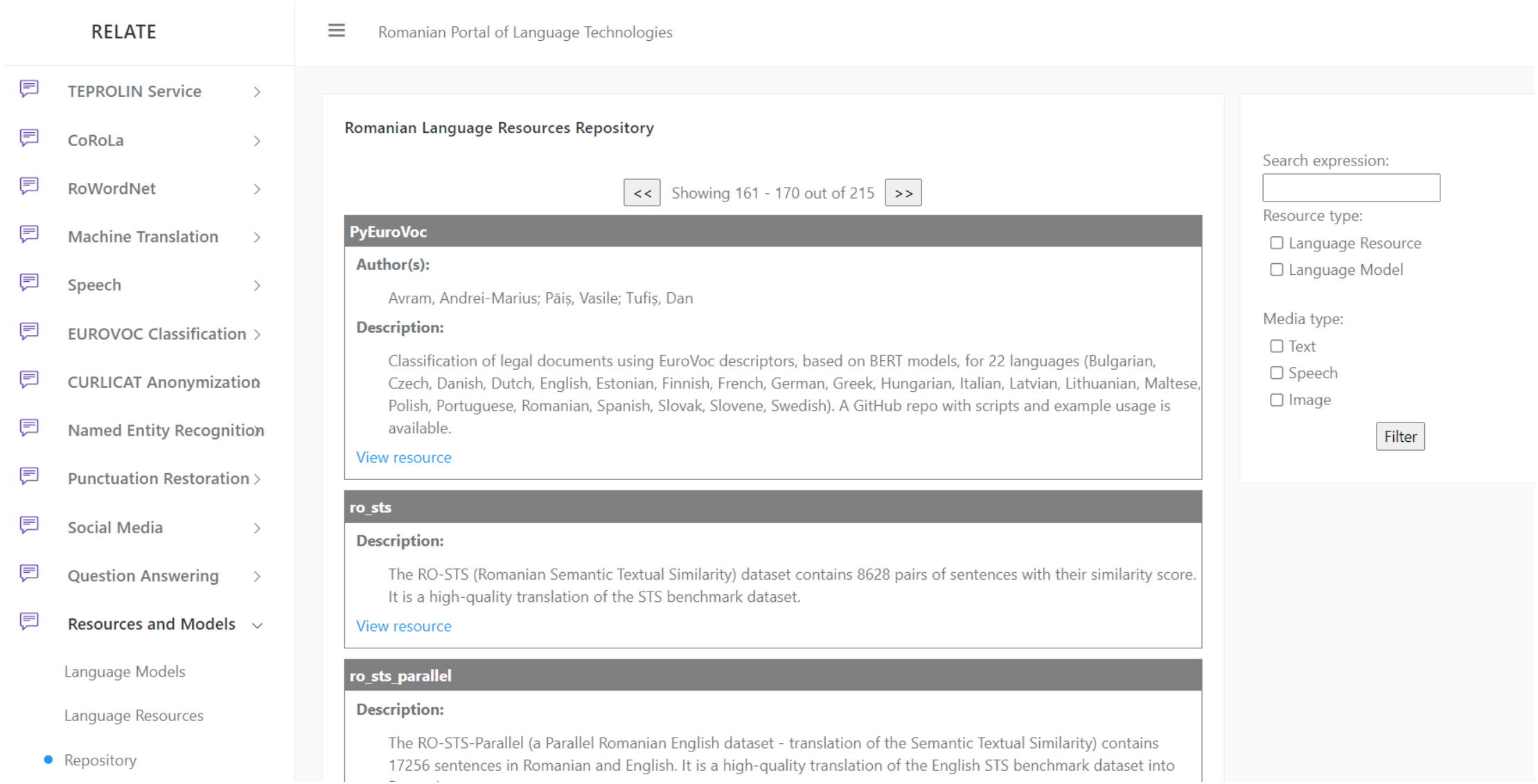

User
Public web interface
Private web interface
Single file processing pipeline
Corpus management
Parallel processing pipeline
Data import/export
DB

RELATE Interface

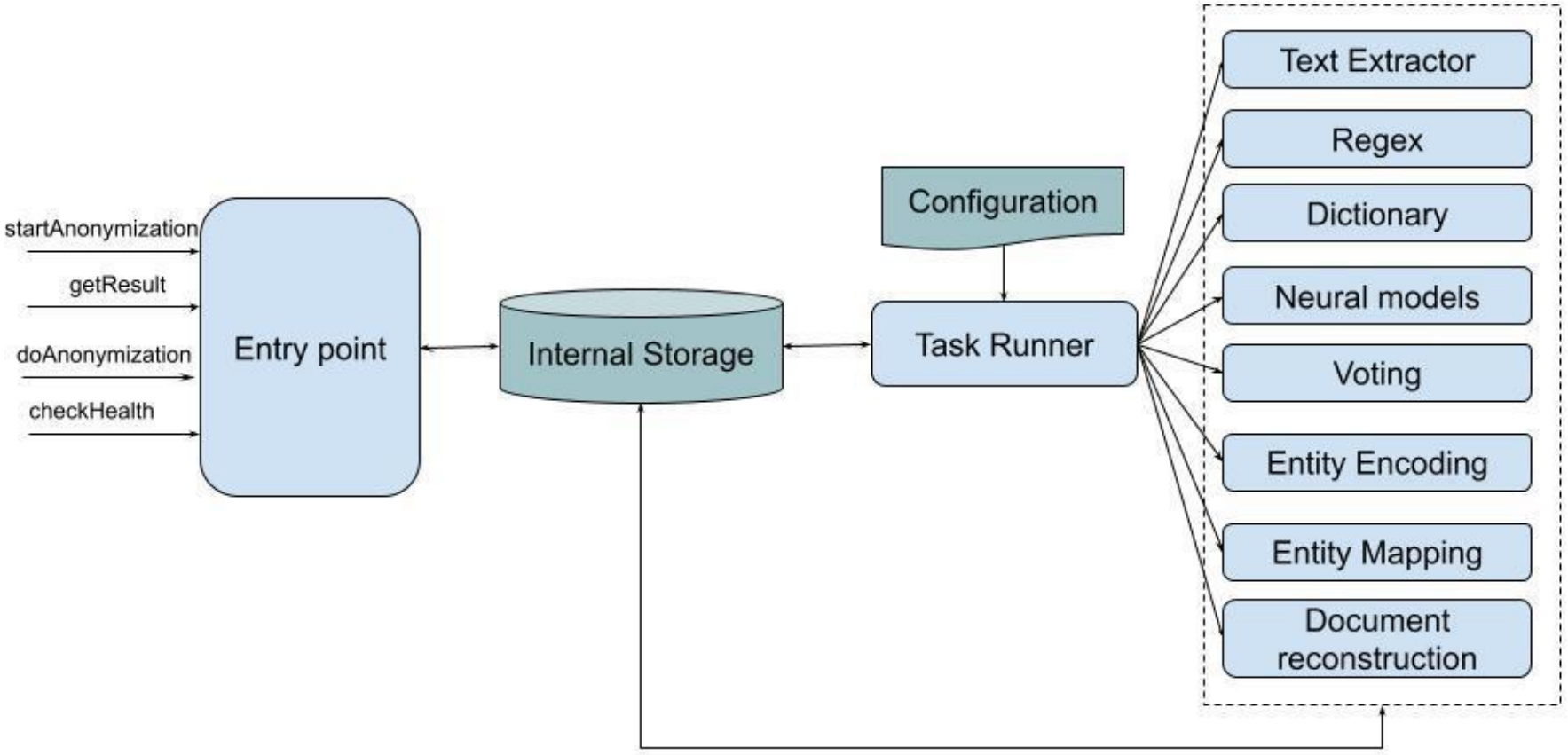



RELATE Web Front-end
RELATE Web Back-end
Format Converters
Task Scheduling
File Storage
Metadata Handling
Statistics
CoRoLa
Linked Data (Fuseki)
TEPROLIN UDPipe (&others)
NER Legal, Biomedical
WordNet
IATE/ EUROVOC
Machine Translation
ASR
TTS
Normalization/ Diacritics restoration
Segmentation/ Tokenization
Lemmatization/ Part-of-Speech
Dependency parsing
Phonetic transcription
Syllabification
Question Answering
Hashtag Splitting
Anonymization
Chunking
Classification
Terminology

Language resources and pre-trained models

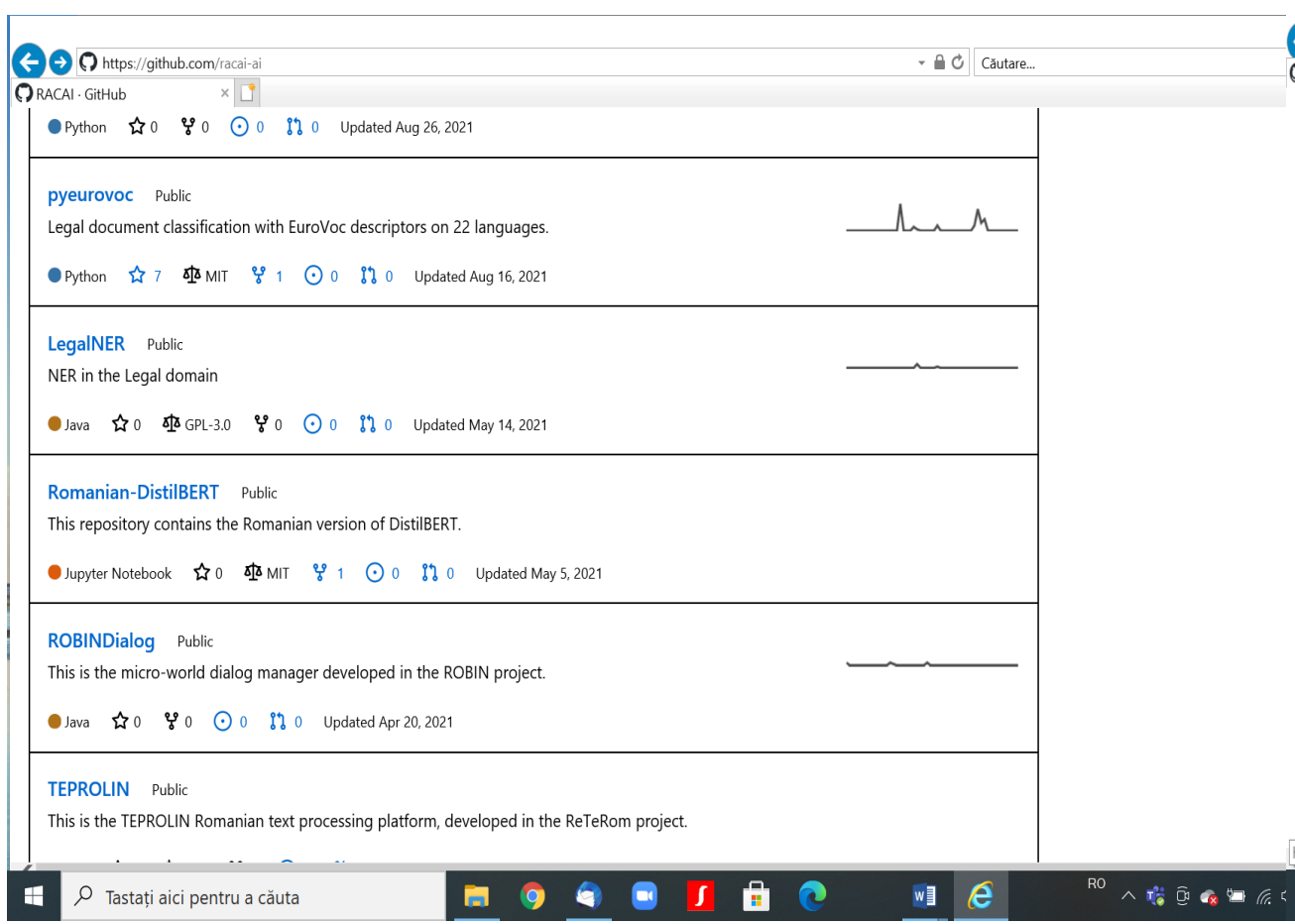


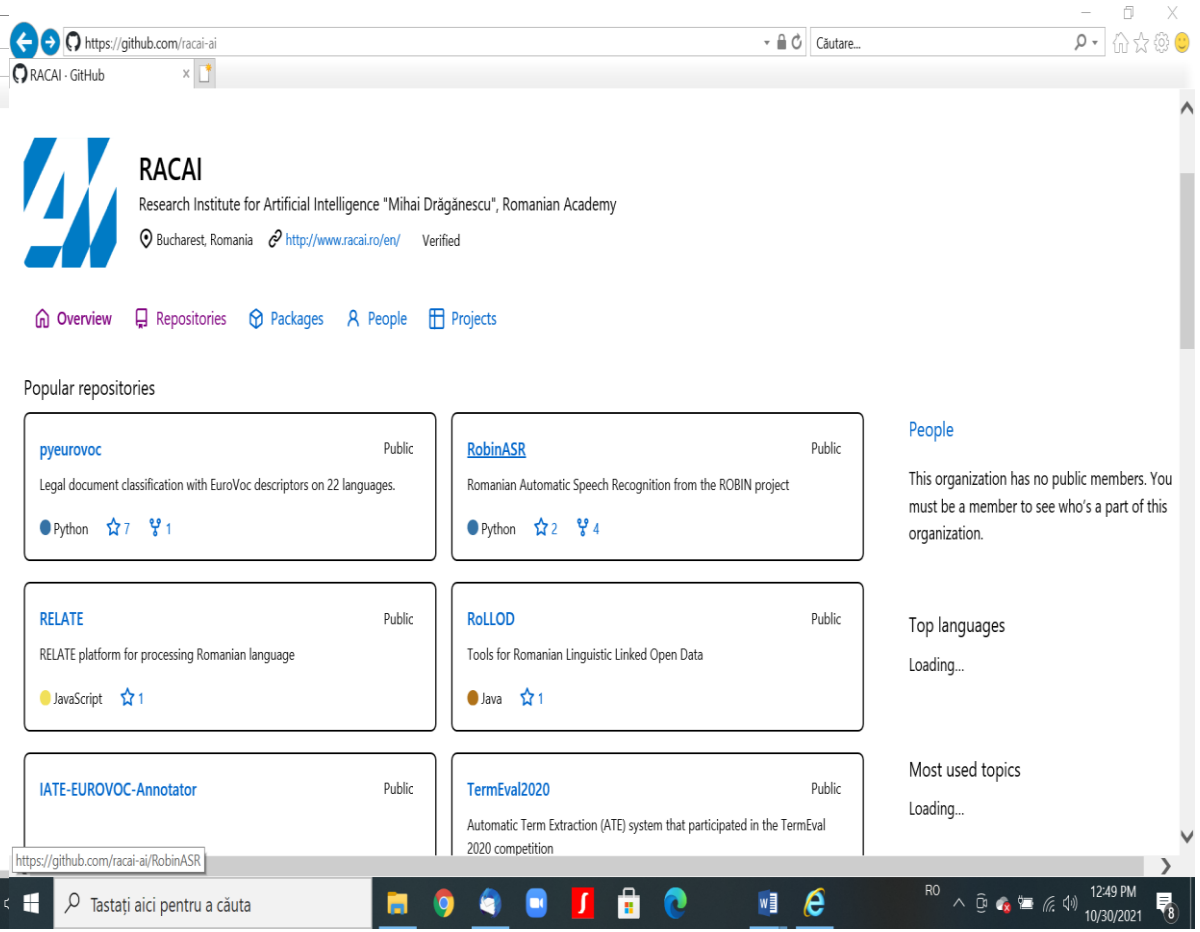
Romanian Jurisprudence Anonimisation



- Modularized architecture
- Integrated in the Romanian jurisprudence portal https://ReJust.ro
- Funded by the Council of Europe
- Beneficiary: Superior Council of Magistracy

ICIA's results "open source"
ICIA github (github.com/racai-ai)





ICIA Research for Human-Centered Artificial Intelligence (HCAI)

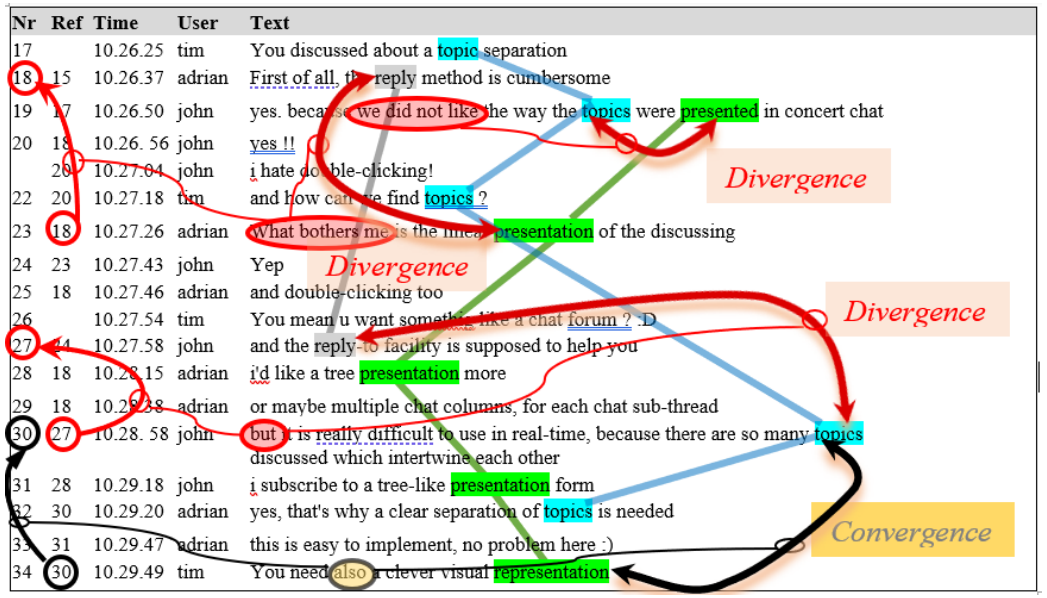
- Human-AI collaboration and co-creation
  - AI in **education**
    - Analysis of human conversations with the polyphonic model operationalized with AI
    - Analysing students' essays
    - Intelligent Tutoring Systems
  - AI for analysing conversations in **medicine**
  - AI for **creativity fostering**
- **Ethical**, unbised AI
- **Human in the loop** – Hermenophore tools
  - **Explainable** AI
  - Detection of hallucinations in Generative AI
  - Fake news detection
- Using and detecting **human touch** in AI – stylometry and polyphonic model

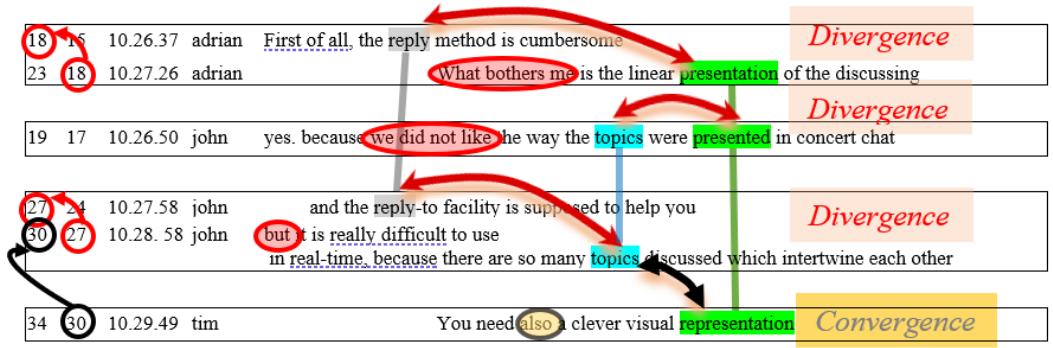
- Results used in applications developed at the NST University Politehnica of Bucharest

The HCAI-based polyphonic model of collaboration and discourse analysis

- The polyphonic model of **collaboration** and **discourse analysis** as result of the ICIA research
- 447 papers refer to it:

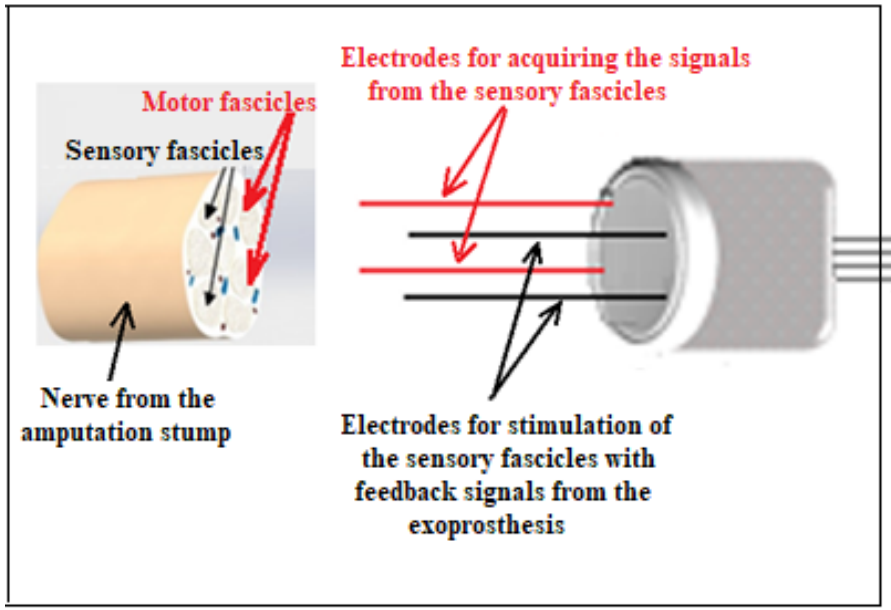
  https://scholar.google.ro/scholar?start=50&q=trausan+polyphony&hl=en&as_sdt=0,5
- Basis for EU projects (LTfLL, RAGE), PolyCafe, ReaderBench systems, national projects, and sonification of conversations
- "… Any true **understanding** is **dialogic** in nature" (Bakhtin), dialog is essential in human life
- **Dialog** is essential in **knowledge construction**
- There are important connections between **creativity, language, and music.**
- In creativity fostering there are two phases: **divergence followed by convergence,** like in **polyphonic music**
- Typical method for creativity fostering is **brainstorming = debates in dialog**

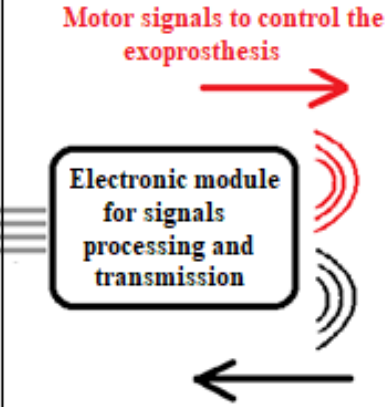


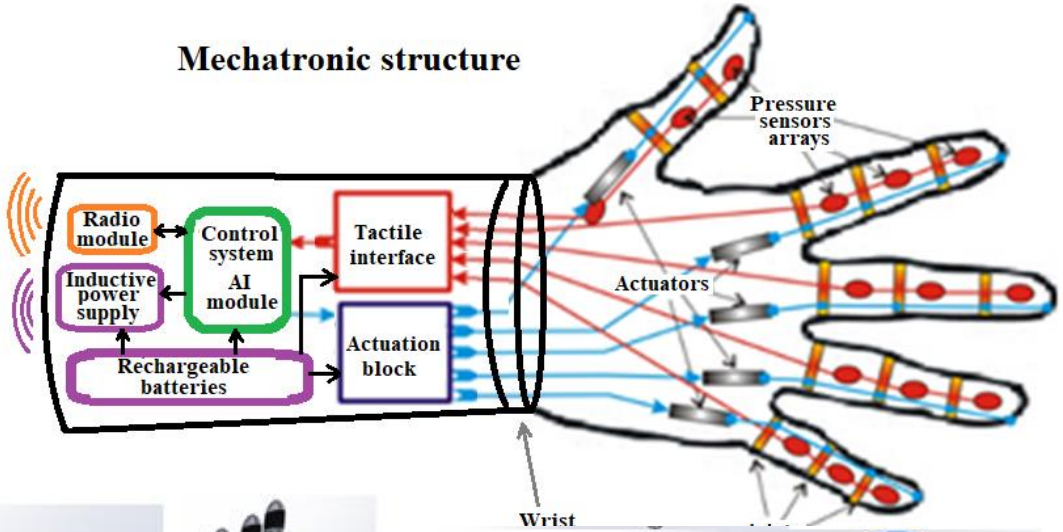


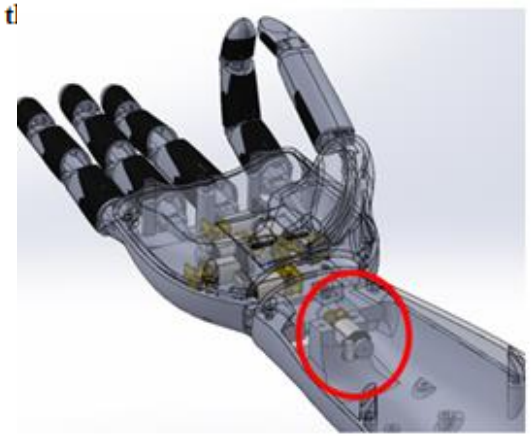
NerveRepack project, HORIZON-KDT-JU-2022-2-RIA, 2023-2027

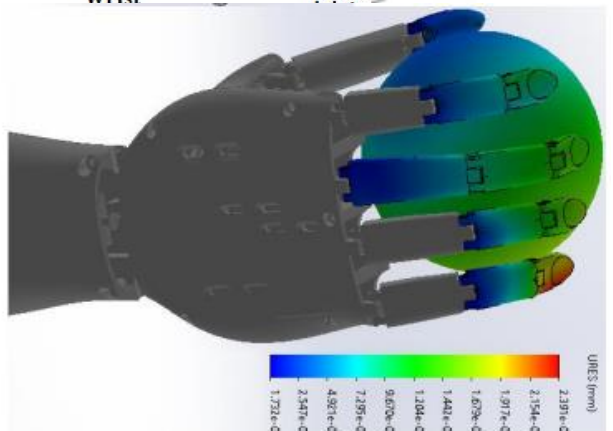
In this project will be designed and fabricated neural implants for exoprostheses. The neural implant will contain: microelectrodes, signals processing module, radio transmission module and inductive power supply module.







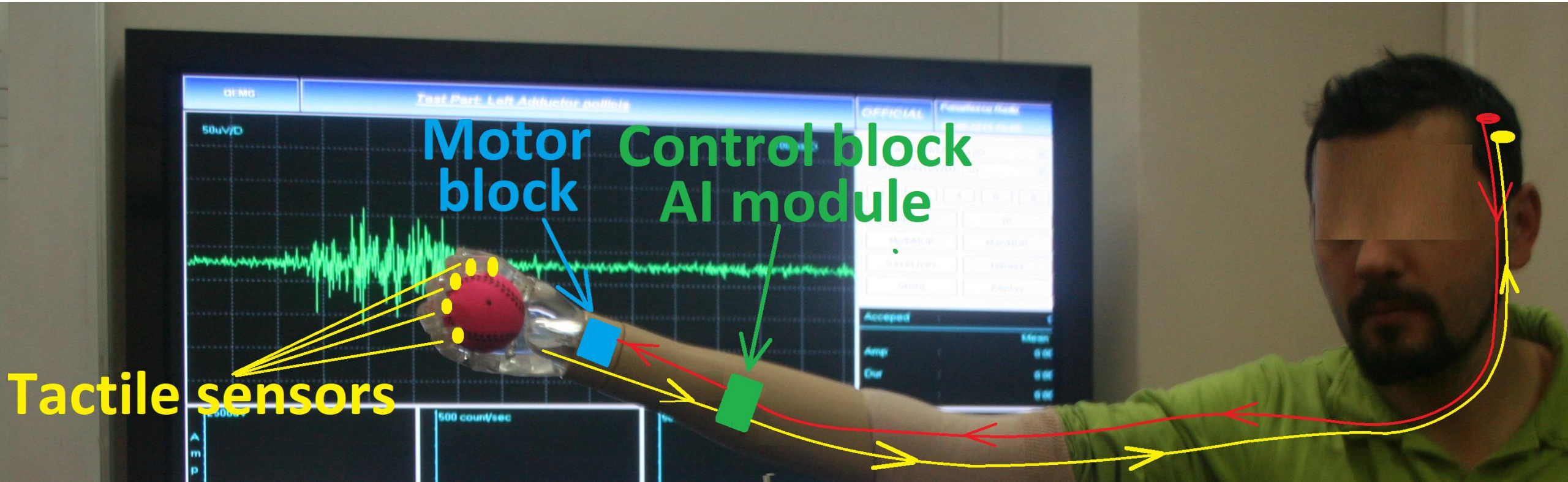




Conclusions

The methodology cannot be applied to all patients who are going to have a forearm amputation but it must be customized for each individual case.

This method will facilitate the electrodes' implantation in the motor nerve branches from the patient's stump and their wireless connection to a neural exoprosthesis that will be able to perform movements similar to a healthy hand.



Dunstan Baby Language Classification with CNN

According to Dustan's theory, before crying, the babies try to communicate their needs using a special „language" that consists of five "words" (or specific utterances) associated with five basic needs:

- "Neh" = hungry;
- "Eh" = need to burp;
- "Oah(Owh)" = tired (sleepy);
- "Eairh (Eargghh)" = stomach cramp (lower gas);
- "Heh" = physical discomfort at skin kevel (feeling hot or wet, for example).

ICIA designed a new architecture for classifying the audio material coming from Romanian babies.

The database loaded with the sounds made by Romanian babies was labelled by doctors in the maternity hospitals and two Dunstan experts, separately.

Finally, the results of the CNN automatic classification were compared to those obtained by the Dunstan coaches

The CNN architecture

A CNN architecture consists of many layers of linked neurons.

A CNN classifies a large number of files (video recordings or audio recordings) into different categories automatically.

In our research we used CNN for the recognition and classification of the words from the so-called “Dunstan baby language”.

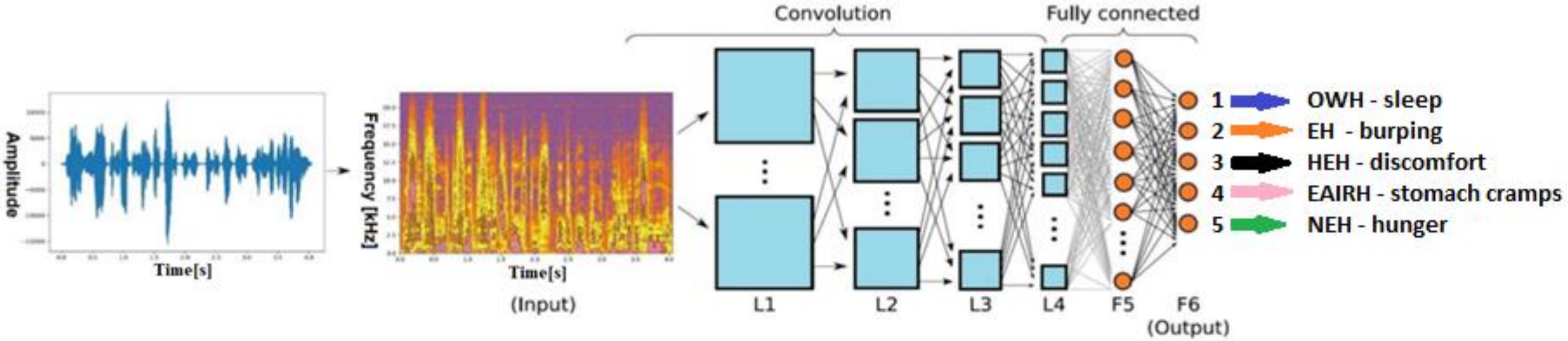
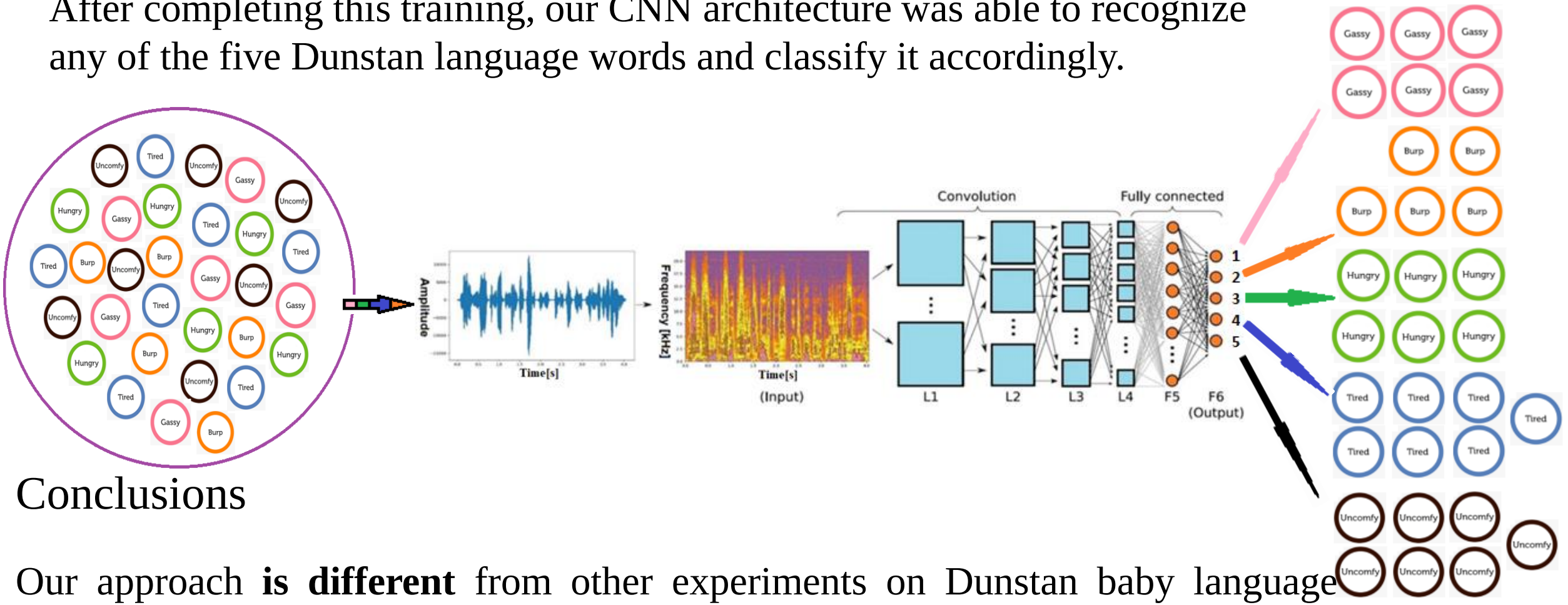


Fig 2. Generic representation of the CNN architecture [1]

The clasification of the Dunstan words with CNN

After completing this training, our CNN architecture was able to recognize any of the five Dunstan language words and classify it accordingly.



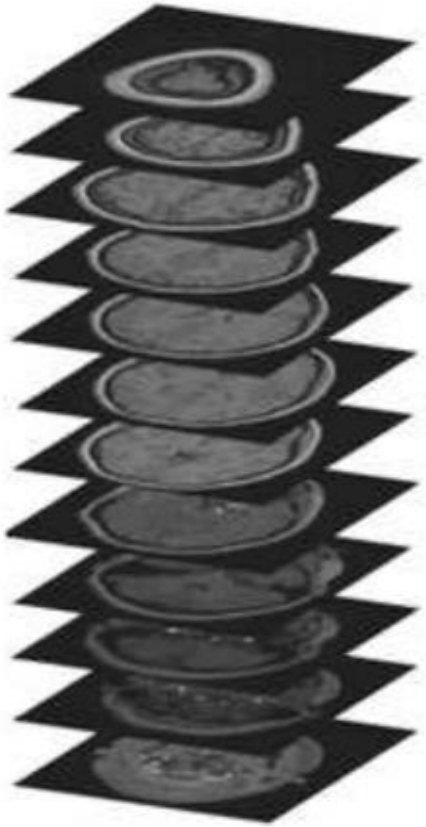
Conclusions

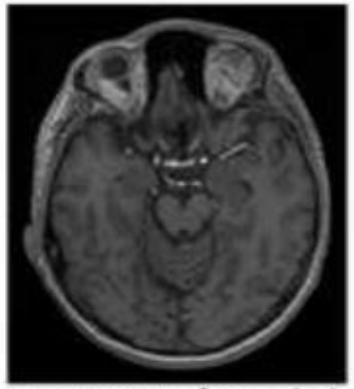
Our approach **is different** from other experiments on Dunstan baby language recognition as our aim was the classification of the “words” of that language (the utterances of the babies that precede the crying), while other results published in the scientific literature aim the recognition of the five types of cries.

Initial Insights into Deep Learning Analysis for Detecting Brain Oxygenation Changes from MRI

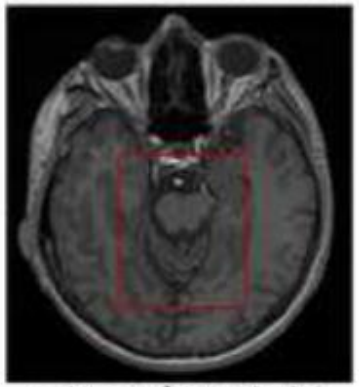
Can deep learning analysis uncover brain oxygenation changes during human screaming?
Investigating Brain Functional Characteristics through Magnetic Resonance Imaging

Methods and Outcomes

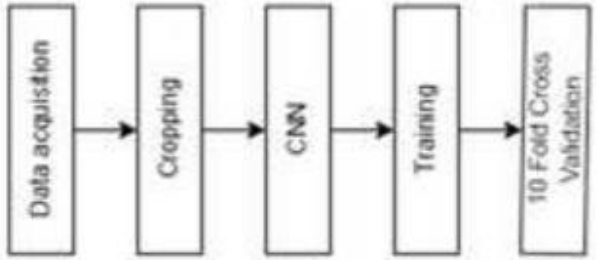




pre-screaming state



screaming state



An accuracy of 89.92% highlights clear differences between MRI scans taken under typical, non-screaming circumstances and those acquired during screaming episodes.

#subject	Age	Gender	Screaming type
#1	57	F	Surprise with disgust/contempt
#1	57	F	Fear
#1	57	F	Happiness
#2	44	F	Frustration
#3	18	F	Anger
#4	25	M	Sadness

Alexandra-Georgiana Andrei, Costin Andrei Brătan, Claudia Tocilă-Mătăşel et al.

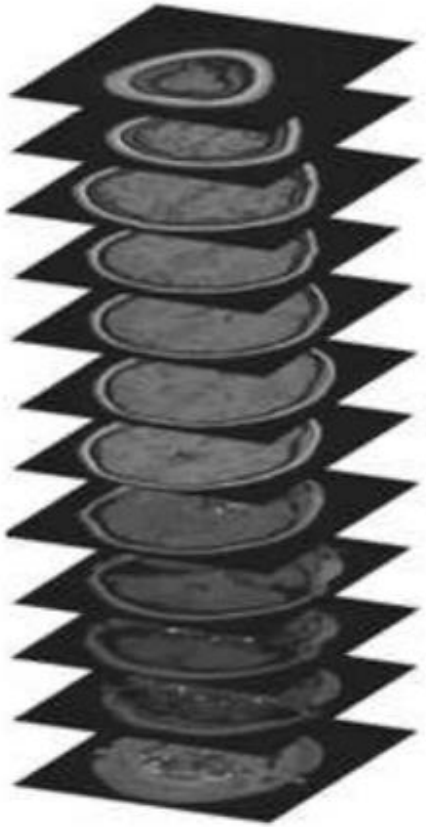
Journal of Biomedical Informatics

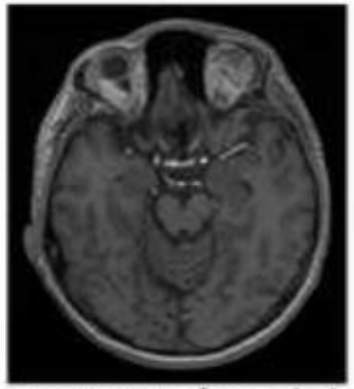
Figure 1: The equipment used in this study - 3.0 T Signa Pioneer MRI scanner produced by GE Healthcare

Initial Insights into Deep Learning Analysis for Detecting Brain Oxygenation Changes from MRI

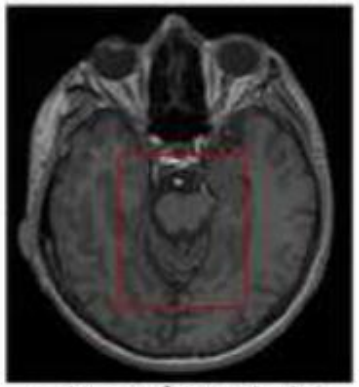
Can deep learning analysis uncover brain oxygenation changes during human screaming?
Investigating Brain Functional Characteristics through Magnetic Resonance Imaging

Methods and Outcomes

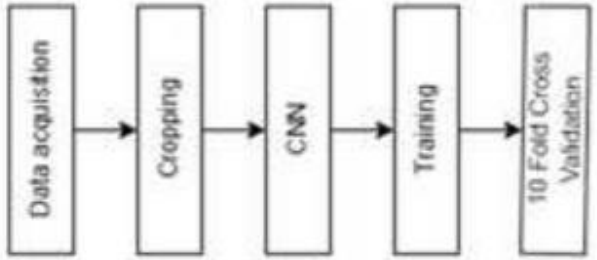




pre-screaming state



screaming state



An accuracy of 89.92% highlights clear differences between MRI scans taken under typical, non-screaming circumstances and those acquired during screaming episodes.

#subject	Age	Gender	Screaming type
#1	57	F	Surprise with disgust/contempt
#1	57	F	Fear
#1	57	F	Happiness
#2	44	F	Frustration
#3	18	F	Anger
#4	25	M	Sadness

Alexandra-Georgiana Andrei, Costin Andrei Brătan, Claudia Tocilă-Mătăşel et al.

Journal of Biomedical Informatics

Figure 1: The equipment used in this study - 3.0 T Signa Pioneer MRI scanner produced by GE Healthcare

What is next?

Research Hot Topics in Language-centered AI

- Automatic Detection of Fake News & Deep Fakes (supported by fact checkers as Politifact, Factcheck.org, Snopes etc.)
- Detection of documents produced by generative language models (ex. ChatGPT)
- Automatic Detection of biases in LLM and eventually eliminate or reduce their overall contribution