



The RELATE Platform

Vasile Păiș

Research Institute for Artificial Intelligence, Romanian Academy

vasile@racai.ro

18.09.2025

Overview

- Introduction
 - Platform Architecture
 - Language resources and pre-trained models
 - Features
- Text Processing
- Multimodal Processing
- Linked Data
- Development

RELATE - A platform for Romanian language technologies and resources

Includes technologies developed at ICIA and by Partners in multiple projects:

CoRoLa, ReTeRom, ROBIN, Presidency, MARCELL, CURLICAT,

Enrich4All, ELE, USPDATRO,

SARoj (System for the Anonymization of Romanian Jurisprudence)

Follows the European Language Grid - ELG philosophy:

- web services, REST APIs, dockers when possible
- services can be distributed across different physical nodes
- services can be consumed directly from partners

Development continues in current projects

<https://relate.racai.ro>

Large Language Models for the EU - LLMs4EU

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

LLMs4EU

alt-edic



**Funded by
the European Union**

<https://www.alt-edic.eu/projects/llms4eu/>

<https://www.racai.ro/p/llms4eu/>

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

RELATE will be used:

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

Defending against deep fake news with large language and image models (DeepNewsDef)

Develop a software platform for detection of deep fakes (text and image) in news content from Romania and the Republic of Moldova, considering the language varieties and characteristics (as reflected in text and images) in the two countries.

RELATE will be used for:

- resource creation, statistics
- automatic annotation and processing of multimodal data

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

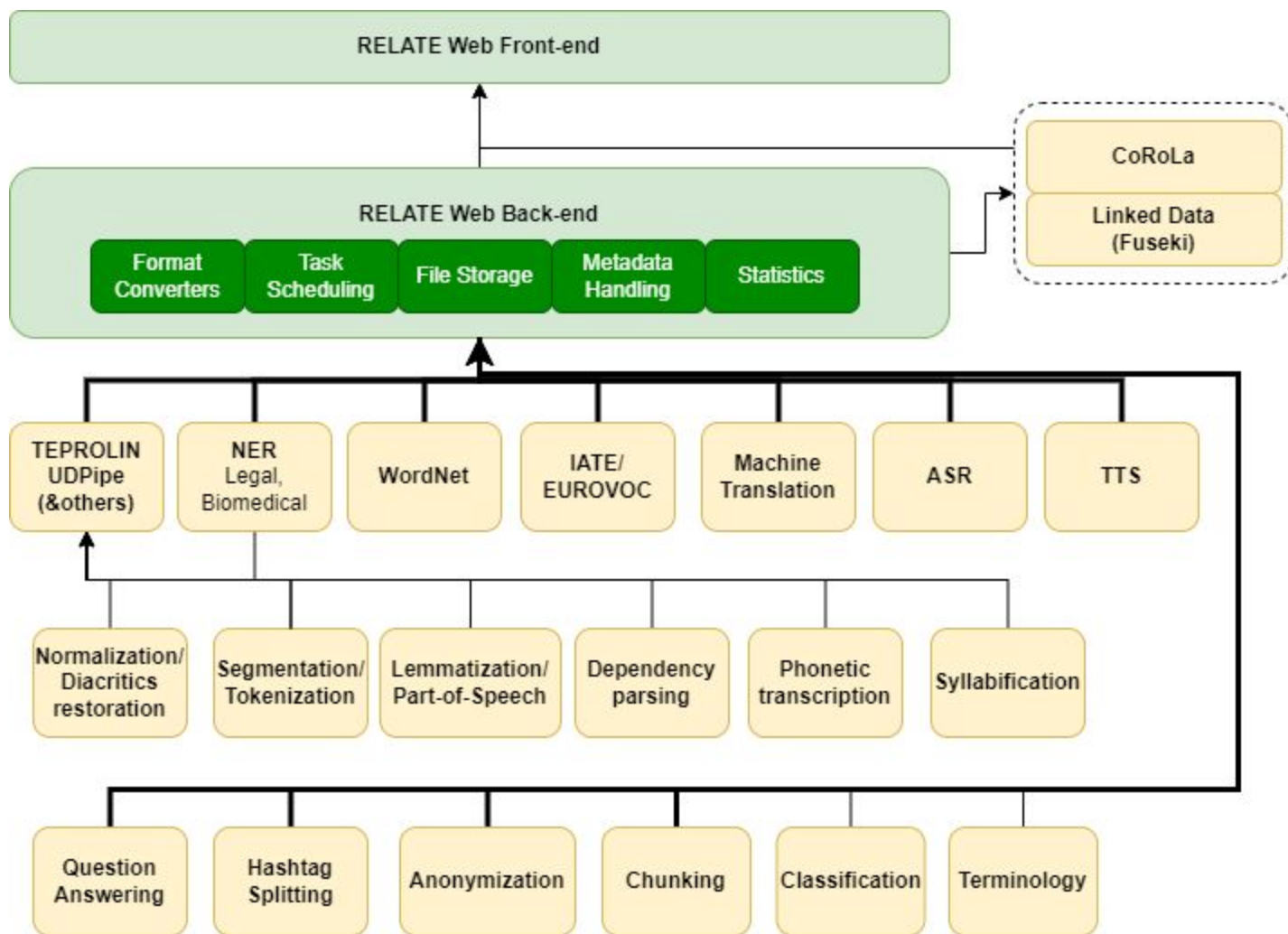


RELATE - overview papers

Vasile Păiș, Radu Ion, and Dan Tufiș. “**A Processing Platform Relating Data and Tools for Romanian Language**”. English. In: Proceedings of the 1st International Workshop on Language Technology Platforms. Marseille, France: European Language Resources Association, 2020, pp. 81–88. URL: <https://www.aclweb.org/anthology/2020.iwlt-1.13>

Vasile Păiș. “**Multiple annotation pipelines inside the RELATE platform**”. In: The 15th International Conference on Linguistic Resources and Tools for Natural Language Processing. 2020, pp. 65–75. URL: <https://profs.info.uaic.ro/~consilr/wp-content/uploads/2021/03/volum-ConsILR-v-4-final-revizuit.pdf#page=73> .

Vasile Păiș, Dan Tufiș, and Radu Ion. “**Integration of Romanian NLP tools into the RELATE platform**”. In: International Conference on Linguistic Resources and Tools for Natural Language Processing. 2019, pp. 181–192. URL: https://profs.info.uaic.ro/~consilr/2019/wp-content/uploads/2020/01/ConsILR2019_final_BTT-60-ex-B5.pdf#page=189 .



Comparison of RO - BLARK

Table 2: Performance comparison of different models trained and evaluated against the RRT corpus for tokenization, sentence splitting, lemmatization, part-of-speech tagging.

	Tokens	Sentences	Lemma	UPOS	XPOS
	F1	F1	Acc	Acc	Acc
TreeTagger	99.72	98.37	89.62	95.58	92.97
TTL	99.66	96.19	95.84	95.54	95.18
UDPipe	99.88	97.39	95.91	97.15	96.24
NLP-Cube	99.86	98.62	51.39	97.73	96.94
RNNTagger	99.23	95.47	98.30	97.82	97.19
Stanza	99.94	98.42	98.90	97.64	97.07

Table 6: Training duration and running speed considering tokenization and complete annotation for RRT-Test and SiMoNERo

	RRT Train		RRT-test		RRT All		SiMoNERo	
	Train	Tokenize	Run		Train	Tokenize	Run	
	time	tok/s	tok/s		time	tok/s	tok/s	
TreeTagger	70s	32648	8162		21s	43201	24337	
TTL	58s	2332	583		63s	279	202	
UDPipe	3h3m	9775	2720		3h15m	22024	2585	
NLP-Cube	13h20m	191	60		14h54m	188	64	
RNNTagger	1d3h44m	42958	281		1d8h46m	73010	453	
Stanza	4d16h8m	1484	430		5d2h10m	186	44	

Păiș, Vasile and Ion, Radu and Avram, Andrei-Marius and Mitrofan, Maria and Tufiș, Dan (2021). *In-depth evaluation of Romanian natural language processing pipelines*. In Romanian Journal of Information Science and Technology (ROMJIST). vol. 24, no. 4, pp. 384--401, <https://www.romjist.ro/full-texts/paper700.pdf>

	UPOS	XPOS
	Acc	Acc
BERT-base-ro	98.01	96.43
RoBERT-small	97.52	96.01
RoBERT-base	98.02	97.18
RoBERT-large	98.10	97.65

Language resources and pre-trained models

RELATE

TEPROLIN Service

CoRoLa

RoWordNet

Machine Translation

Speech

CURLICAT Anonymization

Social Media

Punctuation Restoration

Named Entity Recognition

Downloads

EUROVOC Classification

Corpora

Question Answering

Resources and Models

Language Models

Romanian Portal of Language Technologies

Welcome  Vasile

Pre-Trained Language Models

Annotation models for lemma, UPOS, XPOS and dependency parsing (where supported) trained on RRT UD 2.7.

These models were evaluated in Păiş, Vasile and Ion, Radu and Avram, Andrei-Marius and Mitrofan, Maria and Tufiş, Dan. *In-depth evaluation of Romanian natural language processing pipelines. In Romanian Journal of Information Science and Technology (ROMJIST). vol. 24, no. 4, pp. 384--401, 2021.* The article can be accessed [here](#).

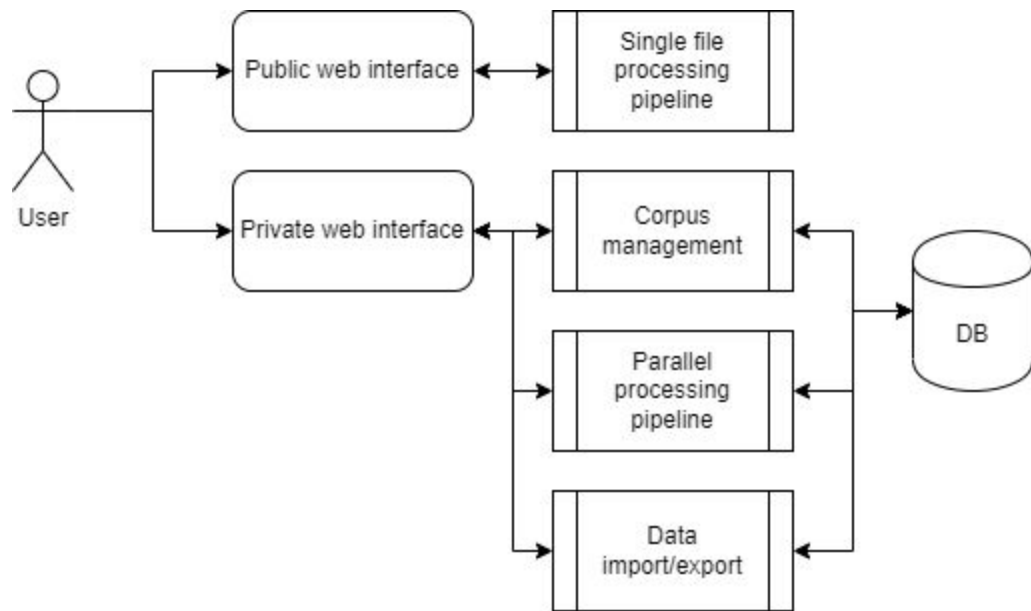
- Stanza [Download \(756Mb\)](#)
- RNNTagger [Download \(668Mb\)](#)
- NLP-Cube [Download \(345Mb\)](#)
- UDPipe [Download \(13Mb\)](#)
- TreeTagger [Download \(1.4Mb\)](#)
- Scripts used in training and evaluating the models are available in our GitHub [here](#).
- A working version of the TTL tool is available in the [TEPROLIN service repository](#).
- For downloading the corpus visit the [Universal Dependencies website](#) or directly download UD 2.7 treebanks from <http://hdl.handle.net/11234/1-3424>.

Classification models

- PyEuroVoc - Classification of legal documents using EuroVoc descriptors, based on BERT models, for 22 languages (Bulgarian, Czech, Danish, Dutch, English, Estonian, Finnish, French, German, Greek, Hungarian, Italian, Latvian, Lithuanian, Maltese, Polish, Portuguese, Romanian, Spanish, Slovak, Slovene, Swedish). A GitHub repo with scripts and example usage is available [here](#). Related paper is Avram Andrei-Marius, Vasile Păiş, and Dan Tufiş. "PyEuroVoc: A Tool for Multilingual Legal Document Classification with EuroVoc Descriptors." *In Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, 2021.
- FastText EuroVoc classification models, based on Common Crawl FastText embeddings for most languages and CoRoLa embeddings for Romanian. Models available for multiple languages can be downloaded [here](#). A modified FastText application allowing models to be interrogated online is available [here](#).

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



RELATE - public view

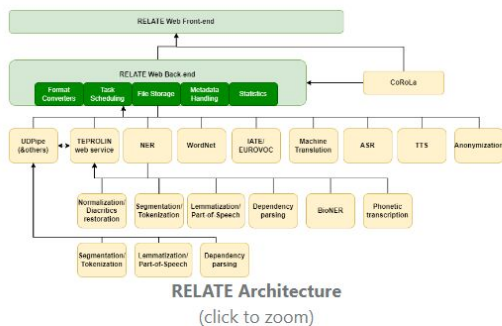
RELATE

☰ Romanian Portal of Language Technologies

Login

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

RELATE - Romanian Portal of Language Technologies



RELATE is a Romanian language technology platform integrating different state-of-the-art tools, algorithms, models and language resources for processing the Romanian language. The modules are developed either in-house or by our partners in different research projects. Please check each page for appropriate references. The modular architecture (see the diagram) allows chaining the available modules into custom pipelines providing advanced language processing capabilities.

The platform allows direct interaction with Romanian language tools for annotating and processing data. For small data sizes it is possible to directly invoke the modules from the web interface in an interactive way. For larger data volumes, the internal platform components allow creating corpora of any size and execute parallel processing pipelines. The platform is open to the public for research purposes (including the internal part, following an account request). The platform is developed for research purposes and may not be suitable for any commercial or production use.

Platform development takes place at GitHub: <https://github.com/racai-ai/relate>

Featured components

TEPROLIN is a web service providing lemmatization, part-of-speech tagging, dependency parsing. The processing flow can be [customized](#) if needed.

Complete run

JSON

CoNLL-U

CoNLL-X

XML

Text

Chunks

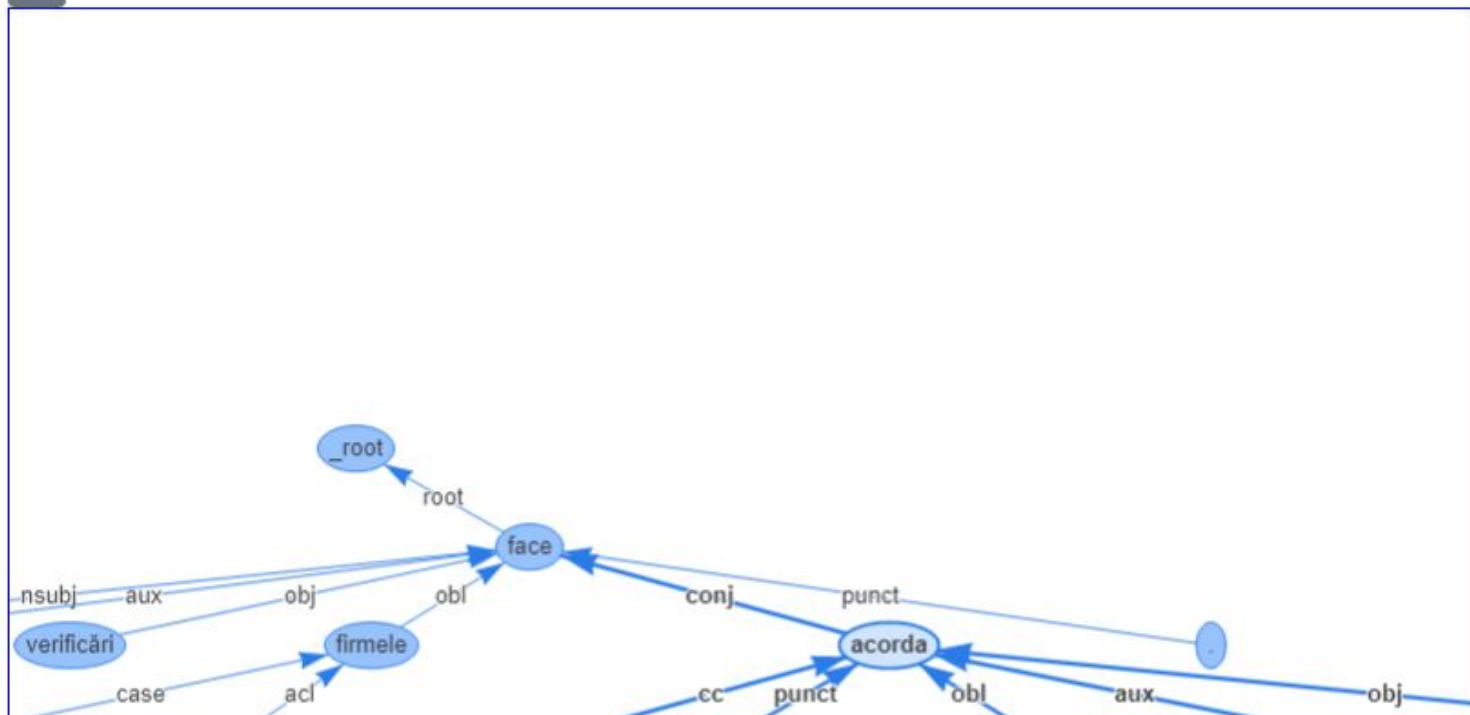
Tree

Entities

Fiscul va face verificări la firmele indicate de CNSP, iar pe zona de dezvoltare va acorda granturi, precum cele pentru primării.

<

>



Word	acorda
Lemma	acorda
U-POS	VERB
CTAG	VN
MSD	Vmnp
Chunk	Vp#3
Named Entity	
Phonetic	a.k.o.r.d.a
Syllables	a.cor.d'a
Similar Words	
Similar Lemma	primi solicita acordare beneficia acordat neacordare

Search in Korap

Search in Wordnet

Search in CoRoLa

Text to Speech

Word embeddings from CoRoLa

Word	acorda
	
Lemma	acorda
	
U-POS	VERB
CTAG	VN
MSD	Vmnp
Chunk	Vp#3
Named Entity	
Phonetic	a.k.o.r.d.a
 	
Syllables	a.cor.d'a
Similar Words	acordă acordat acordată primi beneficia acorde aloca acordau acordase acordam
Similar Lemma	primi solicita acordare beneficia acordat

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

Language Models

Language Resources

 Repository

Romanian Language Resources Repository

Showing 1 - 10 out of 215 >>

AGROVOC - AGROVOC Multilingual Thesaurus**Description:**

Multilingual thesaurus of Agriculture and food. Searching interface and XML/RDF full download.

[View resource](#)**Anatomie-TB - Anatomie-Termbank****Description:**

terminological database about anatomy in various languages

[View resource](#)**Annotated corpora and tools of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions (edition 1.0)****Description:**

The PARSEME shared task aims at identifying verbal MWEs in running texts. Verbal MWEs include idioms (let the cat out of the bag), light verb constructions (make a decision), verb-particle constructions (give up), and inherently reflexive verbs (se suicider 'to suicide' in French). VMWEs were annotated according to the universal guidelines in 18 languages. The corpora are provided in the parsemetsv format, inspired by the CONLL-U format. For most languages, paired files in the CONLL-U format - not necessarily using UD tagsets - containing parts of speech, lemmas, morphological features and/or syntactic dependencies are also provided. Depending on the language, the information

Search expression:

Resource type:

- ☐ Language Resource
☐ Language Model

Media type:

- ☐ Text
☐ Speech
☐ Image

Filter

- TEPROLIN Service ▾
 - Complete Flow
 - Custom Flow
 - Operations & Statistics
 - Developer Documentation
- CoRoLa >
- RoWordNet >
- Machine Translation >
- Speech >
- EUROVOC Classification >
- CURLICAT Anonymization
- Named Entity Recognition
- Punctuation Restoration >
- Social Media >

The TEPROLIN Web Service

Radu Ion (radu@racai.ro)

Introduction

The TEPROLIN Web Service (WS) was developed and is maintained in the [ReTeRom project](#). The backend is the TEPROLIN text preprocessing platform which incorporates several NLP applications for which it provides a unified access interface as a [Python 3](#) object.

TEPROLIN currently offers 15 text preprocessing operations for Romanian, 13 of which are described in (Ion, 2018). These are:

1. text-normalization
2. diacritics-restoration
3. word-hyphenation
4. word-stress-identification
5. word-phonetic-transcription
6. numeral-rewriting
7. abbreviation-rewriting
8. sentence-splitting
9. tokenization
10. pos-tagging
11. lemmatization
12. named-entity-recognition 
13. biomedical-named-entity-recognition 
14. chunking
15. dependency-parsing

RELATE - authenticated view

RELATE

☰ Romanian Portal of Language Technologies

Welcome  Vasile

- TEPROLIN Service >
- CoRoLa >
- RoWordNet >
- Machine Translation >
- Speech >
- CURLICAT Anonymization
- Social Media >
- Punctuation Restoration >
- Named Entity Recognition
- Downloads >
- EUROVOC Classification >
- Corpora >

• List

Corpora

Corpora list

+ Add

	Name	Lang	User	Description	Creation Date ▼
1	coref_test	ro	pvf		2022-11-13
2	termeni nature lematizare	ro	elena@racai.ro		2022-10-09
3	enrich4all covid segmentare	ro	elena@racai.ro		2022-09-08
4	test_marcell_annoate_andrei	ro	andrei.avram		2022-08-10
5	test_marcell_annotate_andrei_v2	ro	andrei.avram		2022-08-10
6	TWITTER_ANONYMIZED_20220725	ro	pvf	Versiunea anonimizata (finala)	2022-07-25
7	TWITTER_MERGED_20220712	ro	pvf		2022-07-12
8	TWITTER Corpus 20220602	ro	pvf		2022-06-02
9	LegalNERo_zenodo_v3	ro	pvf	Textul LegalNERo pentru adnotare c...	2022-03-29
10	BioRo_Extended	ro	maria@racai.ro		2022-02-14
11	corpus_punct	ro	pvf		2022-01-28
12	corpus_wav	ro	pvf		2022-01-28

Page 1 of 4 20 1 to 20 of 67

Files list

[+ Add TEXT](#) [+ Add CSV/TSV](#) [+ Add ZIP TEXT](#)

	Name ▾	Type	Description	User	Creation Date
1	mj_00000G3W5B04K68KG8E2BJ...	text		pvf	2019-10-23 15:04:43
2	mj_00000G3W5A1UQW3MCZG1U...	text		pvf	2019-10-23 15:04:43
3	mj_00000G3W58159BQ4SSF1FC...	text		pvf	2019-10-23 15:04:43
4	mj_00000G3W547PDTJOU1U6U...	text		pvf	2019-10-23 15:04:43
5	mj_00000G3W544EFEG91D4145P...	text		pvf	2019-10-23 15:04:43
6	mj_00000G3W54086WCLHHF371...	text		pvf	2019-10-23 15:04:43
7	mj_00000G3W53QZCYSFP6P15G...	text		pvf	2019-10-23 15:04:43
8	mj_00000G3W51V24NCDHUZ367...	text		pvf	2019-10-23 15:04:43
9	mj_00000G3W4XPRW2QM1XT38...	text		pvf	2019-10-23 15:04:43
10	mj_00000G3W4XFXSAN4T0R39K...	text		pvf	2019-10-23 15:04:43
11	mj_00000G3W4WNUDI2YL0X04P...	text		pvf	2019-10-23 15:04:43
12	mj_00000G3W4UYQKLSX5JG3EH...	text		pvf	2019-10-23 15:04:43
13	mj_00000G3W4UU6GRLNRJG3E8...	text		pvf	2019-10-23 15:04:43
14	mj_00000G3W4SA4N7P0J0R0CF...	text		pvf	2019-10-23 15:04:43



Page

1

of 7207



20



1 to 20 of 144131

[Back](#)[Download](#)[View as Text](#)

File View

	ID	Form	Lemma	UPOS	XPOS	Feats	Head	Deprel	De
1	# sent_id = ro_legal.1								
2	# text = HOTĂRÂRE nr. 1.182 din 4 octombrie 2007								
3	1	HOTĂRÂRE	hotărâre	NOUN	Ncfsrn	Case=Nom Definite=Ind Gender=Fem Number=Sing	0	root	
4	2	nr.	nr.	NOUN	Yn	Abbr=Yes	1	nmod	
5	3	1.182	1.182	NUM	Mc	—	2	nummod	
6	4	din	din	ADP	Spsa	AdpType=Prep Case=Acc	6	case	
7	5	4	4	NUM	Mc	—	6	nummod	
8	6	octombrie	octombrie	NOUN	Ncms-n	Definite=Ind Gender=Masculine Number=Sing	2	nmod	
9	7	2007	2007	NUM	Mc	—	6	nummod	
10									
11	# sent_id = ro_legal.2								

RELATE - Large corpora

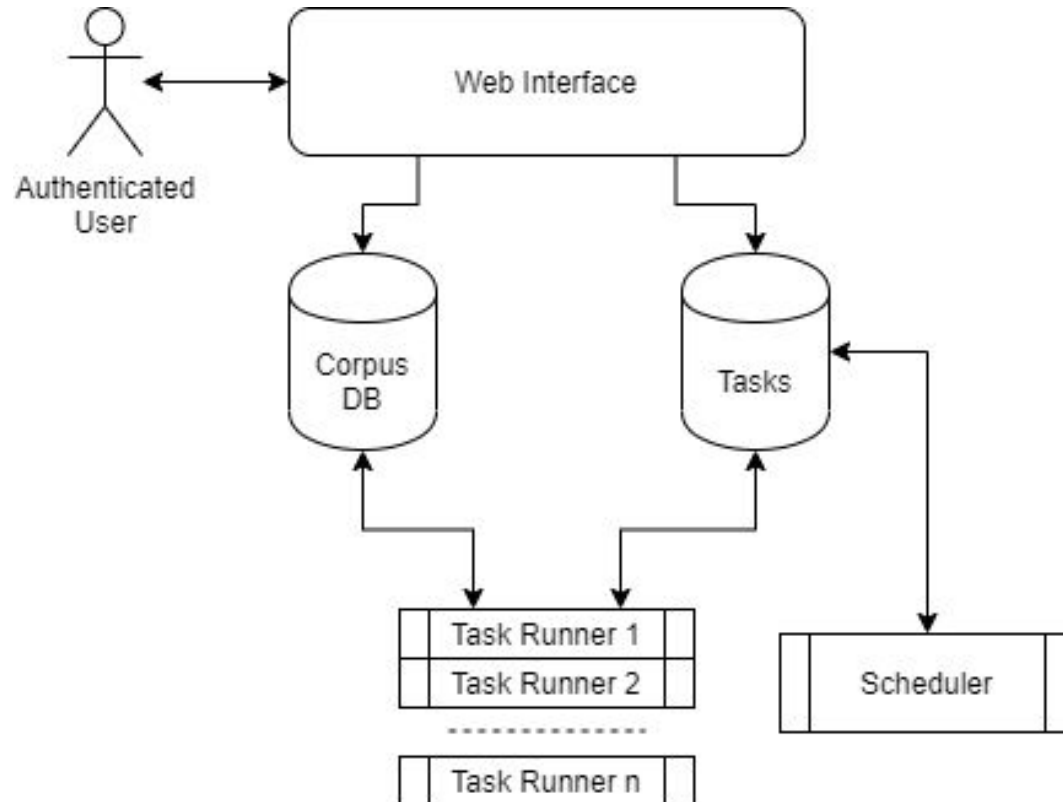
Efficiently stores and handles large corpora:

- text only
- multimodal (text+audio, text+image, text+video)
- text + PDF
 - + metadata

Parallel processing

- task scheduling
- services from multiple nodes

RELATE Task-based processing



RELATE Features

- **Corpora management:** create, upload, download, archive, annotate, statistics, visualize
- **Creation of gold corpora:** integrates BRAT for NER and other tasks, speech recorder for speech-text aligned corpora
- **Annotation:** lemma, part-of-speech, dependency parsing, syllabification, phonetic transcription, diacritics restoration, NER (Legal/BioNER), IATE, EuroVoc, generic terminology
- **Machine translation:** text-to-text, speech-to-speech
- **Speech:** ASR, TTS, speech-to-speech translation
- **WordNet interface:** RoWordNet, aligned query with EnWordNet
- **CoRoLa interface:** Korap, speech component, word embeddings

RELATE Features

- **Anonymization:** from the CURLICAT project
 - **Legal NER:** Person, Organization, Location, Time, Legal references (from a gold annotated sub-corpus derived from the MARCELL corpus)
 - **Pre-trained language models:** available for download
 - **Standard formats:** CoNLL-U, CoNLL-U Plus, XML, JSON, RDF
-
- **A modular architecture which continues to expand with new projects**

Text processing

- Gold corpora creation
- Task-based processing
- Standoff Metadata
- Import/Export

Gold corpora creation - NER - MicroBloggingNERo

RELATE

TEPROLIN Service >

CoRoLa >

RoWordNet >

Machine Translation >

Speech >

EUROVOC Classification >

Downloads >

Punctuation Restoration >

Romanian Portal of Language Technologies

Welcome  Vasile

Corpus: TWITTER_2 View Corpora

Back

Download

Download ANN

Save To GOLD

PREVIOUS

NEXT

Language Romanian

Sentiment Neutral

Hate No

Language Type Regular

Save

To see annotations from CONLLUP files make sure to run the CONLLUP2BRAT task.

/TWITTER_2/147179008022244868.txt

brat

1	Scaunul lui PER Boris Johnson se clatină. Conservatorii pierd o circumscriptie pe care o controlau de două secole - TIME <url> <url>
---	--

- Păiș, Vasile and Mitrofan, Maria and Barbu-Mititelu, Verginica and Irimia, Elena and Gasan, Carol Luca and Micu, Roxana and Marin, Laura and Dicusar, Maria and Florea, Bianca and Badila, Ana (2022). *Romanian micro-blogging named entity recognition (MicroBloggingNERo)*. Zenodo, <https://doi.org/10.5281/zenodo.6905235>
- Păiș, Vasile and Barbu Mititelu, Verginica and Irimia, Elena and Mitrofan, Maria and Gasan, Carol Luca and Micu, Roxana (2022). *Romanian micro-blogging named entity recognition including health-related entities*. In Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, pp. 190--196, <https://aclanthology.org/2022.smm4h-1.49>

Gold corpora creation - NER - LegalNERo

RoWordNet

Machine Translation

Speech

Downloads

EUROVOC Classification

CURLICAT Anonymization

Corpora

Pretrained LM

Citation

MY

Back

Download

Download ANN

Save To GOLD

To see annotations from CONLLUP files make sure to run the CONLLUP2BRAT task.

/MARCELL_NER/mj_00000G1TVLRU8DK2LRS0OL3XTVLNBM9Q.txt

brat

1

ORDIN nr. 533 din 30 octombrie 2019

LAW

TIME

2

privind înființarea Comisiei de concesionare a terenurilor pe care sunt amplasate amenajările piscicole, precum și a altor terenuri aferente acestora din domeniul public și privat al statului și de aprobare a Normelor referitoare la procedura privind concesionarea terenurilor pe care sunt amplasate amenajările piscicole și a altor terenuri aferente acestora din domeniul public și privat al statului

ORG

3

Publicat în

5

MONITORUL OFICIAL nr. 886 din 4 noiembrie 2019

ORG

TIME

7

Având în vedere Referatul de aprobare nr. 10.064 din 30 septembrie 2019 al Agenției Naționale pentru Pescuit și Acvacultură,

TIME

ORG

8

–

9

art. 4 alin. (3) lit. b)-g) și alin. (5) din Ordonanța de urgență a Guvernului nr. 23 / 2008 privind pescuitul și acvacultura, aprobată cu modificări și completări prin Legea nr. 317 / 2009, cu modificările și completările ulterioare;

LAW

ORG

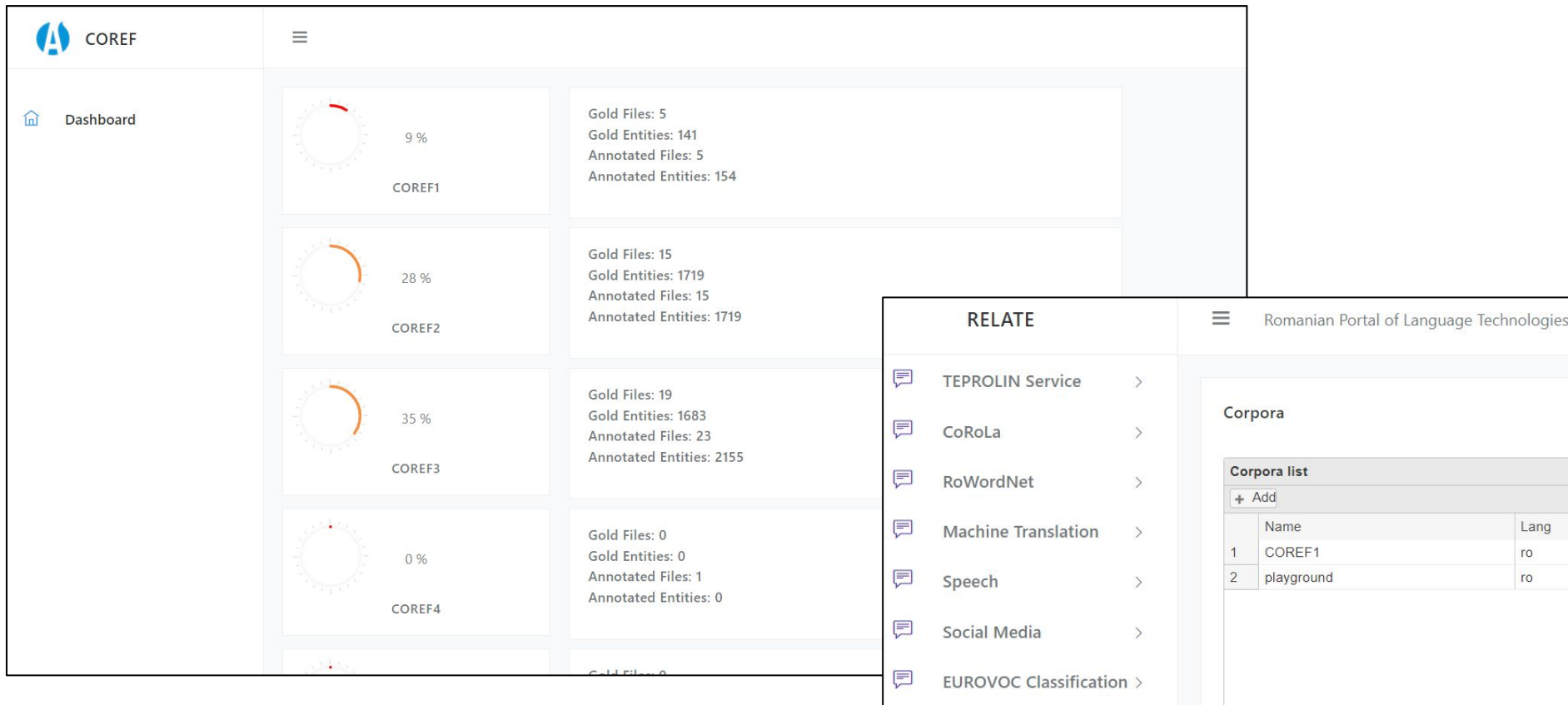
TIME

LAW

TIME

- Păiș, Vasile and Mitrofan, Maria and Gasan, Carol Luca and Ianov, Alexandru and Ghiță, Corvin and Coneschi, Vlad Silviu and Onuț, Andrei (2021). *Romanian Named Entity Recognition in the Legal domain (LegalNERo)*. Zenodo, <https://doi.org/10.5281/zenodo.4772094>
- Păiș, Vasile and Mitrofan, Maria. (2021) *Towards a named entity recognition system in the Romanian legal domain using a linked open data corpus*. In Workshop on Deep Learning and Neural Approaches for Linguistic Data, pp. 16--17.
- Păiș, Vasile and Mitrofan, Maria and Gasan, Carol Luca and Coneschi, Vlad and Ianov, Alexandru (2021). *Named Entity Recognition in the Romanian Legal Domain*. In Proceedings of the Natural Legal Language Processing Workshop, pp. 9--18. <https://aclanthology.org/2021.nllp-1.2>

Gold corpora creation - Annotator Instances + Dashboard



Multiple NER models

RELATE

TEPROLIN Service >

CoRoLa >

RoWordNet >

Machine Translation >

Speech >

CURLICAT Anonymization

Social Media >

Punctuation Restoration >

Named Entity Recognition

Romanian Portal of Language Technologies

Welcome  Vasile

Named Entity Recognition

Selectați modelul NER dorit și introduceți un document în limba română pentru recunoașterea entităților. Modelele bazate pe corpusul LegalNERo presupun un text de tip legislație. Modelele bazate pe corpusul SiMoNERo presupun un text din domeniul biomedical. Utilizarea de texte din alte domenii reduce calitatea rezultatelor. Pentru a putea realiza ancorarea corectă a rezultatelor în textul introdus, nu sunt realizate curățări ale datelor. Textul trebuie să conțină doar caractere în limba română, cu virgula și punctul separate de cuvântul anterior.

LegalNERo LEGAL, PER, LOC, ORG, TIME with Gazetteer ▼

LegalNERo LEGAL, PER, LOC, ORG, TIME with Gazetteer

LegalNERo PER, LOC, ORG, TIME with Gazetteer

LegalNERo LEGAL, PER, LOC, ORG, TIME

LegalNERo PER, LOC, ORG, TIME

SiMoNERo ANAT, CHEM, DISO, PROC

- Păiș, Vasile (2022). RACAI at SemEval-2022 Task 11: Complex named entity recognition using a lateral inhibition mechanism. In Proceedings SemEval, pp. 1562--1569, <https://aclanthology.org/2022.semeval-1.215>
- Mitrofan, Maria and Păiș, Vasile (2022). Improving Romanian BioNER Using a Biologically Inspired System. In Proceedings of the 21st Workshop on Biomedical Language Processing, pp. 316--322, <https://aclanthology.org/2022.bionlp-1.30>

Task-based processing

Corpus: **Marcell**

[View Corpora](#)

[Files](#)

[Standoff](#)

[Tasks](#)

[Annotated](#)

[Statistics](#)

[Archives](#)

Files list

[+ Add TEXT](#) [+ Add CSV/TSV](#) [+ Add PDF](#) [+ Add ZIP](#) [+ Access Last File](#)

	Name ▲	Type	Description	User	Creation Date
1	mj_00000G0001I4Q9GISD13TF7D...	text		pvf	2019-10-23 15:04:43
2	mj_00000G0001L9DUTVXXJ37F4F...	text		pvf	2019-10-23 15:04:43
3	mj_00000G0002CA2MKRE8V1QE...	text		pvf	2019-10-23 15:04:43
4	mj_00000G0002I4DEC59G23QLY...	text		pvf	2019-10-23 15:04:43
5	mj_00000G0005HOC5GKLDY0DN...	text		pvf	2019-10-23 15:04:43
6	mj_00000G0007WC5AJK76I255P9...	text		pvf	2019-10-23 15:04:43
7	mj_00000G000827N7KZV7704W5...	text		pvf	2019-10-23 15:04:43
8	mj_00000G0008L3G0KSO062ZUJ...	text		pvf	2019-10-23 15:04:43
9	mj_00000G0008LFFWVAOBH37E...	text		pvf	2019-10-23 15:04:43
10	mj_00000G0008NVRHIFAVJ3R6X...	text		pvf	2019-10-23 15:04:43
11	mj_00000G000BRZ93RU63D3NY0...	text		pvf	2019-10-23 15:04:43
12	mj_00000G000D2PI9LA0VR3NB3T...	text		pvf	2019-10-23 15:04:43

Page 1 of 7207 20 1 to 20 of 144131

MARCELL-RO Corpus

- Part of the CEF Telecom project Multilingual Resources for CEF.AT in the legal domain (MARCELL)
- 7 legislative corpora containing the total body of national legislative documents in effect for 7 countries included in the consortium: Bulgaria, Croatia, Hungary, Poland, Romania, Slovakia and Slovenia
- Tokenized, lemmatized, morphologically annotated, dependency parsed, named entities, nominal phrases, term annotation with IATE and EuroVoc.
- annotated in the RELATE platform
- statistics computed in the platform (4.3 million sentences, 144k documents, 67 million tokens)
- exported as archives (text, conllu plus and xml) from the platform

Task-based processing

Corpus: Marcell

View Corpora

Files

Standoff

Tasks

Annotated

Statistics

Archives

Corpus tasks

+ TEPROLIN	+ UDPIPE	+ NERRegex	+ Classify EuroVoc	+ Terminology	+ IATE/EuroVoc	+ Cleanup	+ NER Baseline	+ CoNLLUP2BRAT	+ NER_Old	+ TTLChunker	+ BRAT2CoNLLUP
+ Gold NE list	+ Statistics	+ ZIP Text	+ ZIP Annotated	+ ZIP Standoff	+ ZIP Gold Standoff	+ ZIP Gold Annotated	+ ZIP Audio	+ Change Terms Marcell	+ Export CURLICAT		
+ Export Marcell	+ NER-Legal	+ Translate Text	+ ASR	+ Punctuation	+ TTS	+ Translate S2S	+ DeAnonymization	+ NER-SiMoNERo	+ Anonymization	+ Create Empty Metadata	
	Type	Status	Description	User	Creation Date ▼						
1	marcell	DONE		pvf	2021-03-12 19:03:48						
2	marcell	DONE		pvf	2021-03-12 15:40:04						
3	eurovoc_class	DONE		pvf	2021-03-12 11:09:05						

- Tamás Váradi, Svetla Koeva, Martin Yamalov, Marko Tadić, Bálint Sass, Bartłomiej Nitoń, Maciej Ogrodniczuk, Piotr Pezik and Barbu Mititelu, Verginica and Radu Ion and Irimia, Elena and Mitrofan, Maria and Păiş, Vasile and Tufiş, Dan and Radovan Garabík, Simon Krek, Andraz Repar, Matjaž Rihtar and Janez Brank (2020). *The MARCELL Legislative Corpus*. In Proceedings of The 12th LREC, pp. 3754-3761. <https://www.aclweb.org/anthology/2020.lrec-1.464/>
- Tufiş, Dan and Mitrofan, Maria and Păiş, Vasile and Ion, Radu and Coman, Andrei (2020). *Collection and Annotation of the Romanian Legal Corpus*. In Proceedings of The 12th LREC, pp. 2766-2770. <https://www.aclweb.org/anthology/2020.lrec-1.337/>
- <https://elrc-share.eu/repository/search/?q=marcell>

Task-based processing

The screenshot displays the CURLICAT web interface. At the top, there are navigation tabs: Files, Standoff, Tasks (highlighted in green), Annotated, Statistics, and Archives. Below these, the 'Corpus tasks' section shows a list of tasks with checkboxes: TEPROLIN, UDPipe, NERRegex, Gold NE list, Statistics, ZIP Text, Export Marcell, NER-Legal, and Translate. A table below this lists tasks with columns for 'Type' and 'Status'. The 'Add task Terminology' dialog is open, showing instructions: 'Annotate text a custom terminology. The text must be already tokenized and lemmatized. At least one file with .terminology extension must be present in the corpus standoff folder.' The dialog fields include: Column (RELATE:TERM), Max Term Size (10), Terminology (custom.terminology), Description (empty text area), Runners (optional) (empty text area), and Overwrite (checkbox). The dialog has 'Add' and 'Cancel' buttons. In the background, another panel shows a list of files with columns for 'NER_Old', 'TTLChunker', 'BRAT2CoNLLUP', 'Marcell', 'Export CURLICAT', 'Nymization', and 'Create Empty Metadata'. A table below this shows a list of files with a 'Creation Date' column.

Creation Date
2022-11-01 09:56:34
2022-10-31 13:37:07
2022-10-30 17:12:46
2022-10-30 17:12:26
2022-10-30 17:12:08
2022-10-30 15:26:24
2022-10-30 10:46:38
2022-10-30 10:46:06
2022-09-27 05:30:59
2022-09-26 11:27:43
2022-09-26 11:18:59

- Váradi, Tamás and Nyéki, Bence and Koeva, Svetla and Tadić, Marko and Štefanec, Vanja and Ogrodniczuk, Maciej and Nitoń, Bartłomiej and Pezik, Piotr and Barbu Mititelu, Verginica and Irimia, Elena and Mitrofan, Maria and Păiș, Vasile and Tufiș, Dan and Garabík, Radovan and Krek, Simon and Repar, Andraž (2022). **Introducing the CURLICAT Corpora: Seven-language Domain Specific Annotated Corpora from Curated Sources**. In Proceedings LREC, pp. 100--108. <https://aclanthology.org/2022.lrec-1.11/>
- <https://elrc-share.eu/repository/search/?q=curlicat>

CURLICAT-RO Corpus

- Part of the CEF Telecom project Curated Multilingual Resources for CEF.AT(CURLICAT)
 - an ongoing project (will finish this year)
 - 7 monolingual corpora containing texts from the following fields: culture, education, economy, health, nature, politics, science
 - Bulgaria, Croatia, Hungary, Poland, Romania, Slovakia and Slovenia
 - Anonymized, tokenized, lemmatized, morphologically annotated, dependency parsed, named entities, nominal phrases, term annotation with IATE and automatic term recognition.
-
- annotated in the RELATE platform
 - statistics computed in the platform (4.4 million sentences)
 - will be exported as archives (text, conllu plus) from the platform

Task-based processing - Statistics

Files Standoff Tasks **Statistics** Archives

Statistics	
+ Download Stats + View WordForm Stats + View Lemma Stats + View WordForm Doc Freq + View Letters Stats + View Lemma UPOS Stats + View IATE Stats + View IATE Doc Freq + View EUROVOC ID Stats + View EUROVOC ID Doc Freq + View EUROVOC MT Stats + View EUROVOC MT Doc Freq	
Key	Value
1 Basic.Dis Legomena	133276
2 Basic.Hapax Legomena	653298
3 Basic.Number of Annotated Documents	26477
4 Basic.Number of Characters	559682566
5 Basic.Number of EUROVOC IDs	0
6 Basic.Number of EUROVOC MTs	0
7 Basic.Number of IATE terms	18772350
8 Basic.Number of Lines	3390728
9 Basic.Number of Raw Documents	26477
10 Basic.Number of Romanian letters	426159374
11 Basic.Number of Sentences	3610766

Page 1 of 30

20

1 to 20 of 583

Standoff Metadata

[Back](#)[PREVIOUS](#)[NEXT](#)

Peștera Proserv SA / 205 3.3. Parodia literară și asumarea modern
Capitolul 4. Parodii românești ale secolului XX: de la practicile
Drumul spre postmodernism/ 249 4.2. Postmodernismul: un construct
Concluzii / 329
Bibliografie / 337 Abstract / 343 Apălimăriței / 347
Cuvânt înainte
Parodia literară: șapte rescrieri românești reprezintă, în peisaj
Autoarea încearcă să răspundă unei duble provocări. de ordin teor

1021_a_2529.xml

[Save](#)

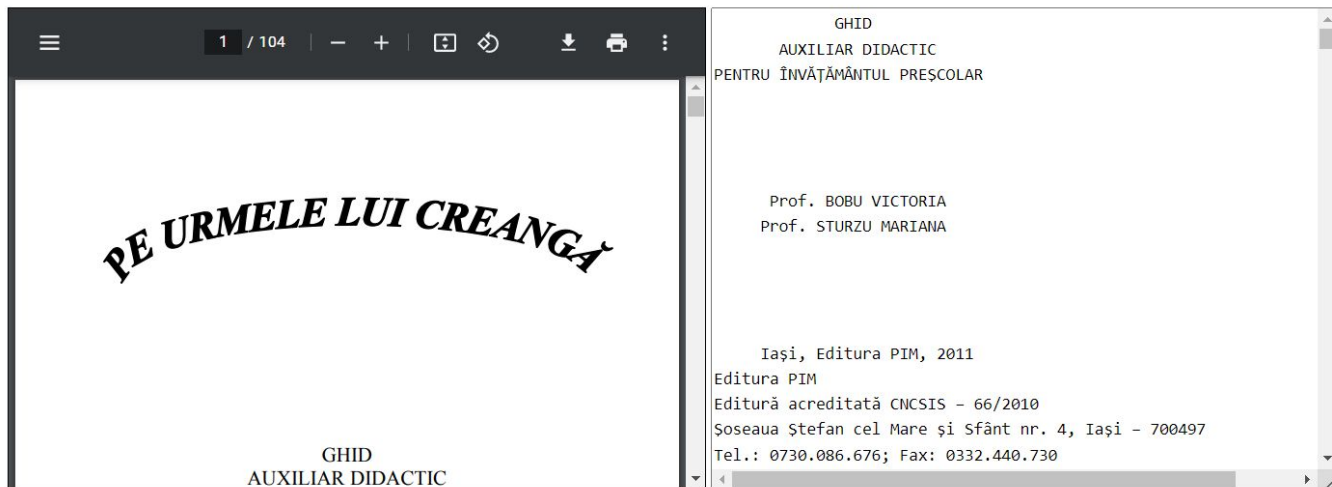
```
<?xml version="1.0" encoding="UTF-8"?>
  <Metadata>
    <DocumentTitle>Parodia literară. Șapte rescrieri românești</DocumentTitle>
    <ArticleTitle>-</ArticleTitle>
    <AuthorName>Livia Iacob</AuthorName>
    <PublicationDate>2011</PublicationDate>
    <Source>Publishing House</Source>
    <SourceName>Editura Europeană</SourceName>
    <TranslatorName>-</TranslatorName>
```

1021_a_2529.ann

[Save](#)

T1	LOC 125 132	România
T2	LOC 435 442	România
T3	ORG 857 873	Mahtab Impex SRL

Standoff Metadata - CoRoLa



Pe urmele lui Creanga FINAL.xml

Save

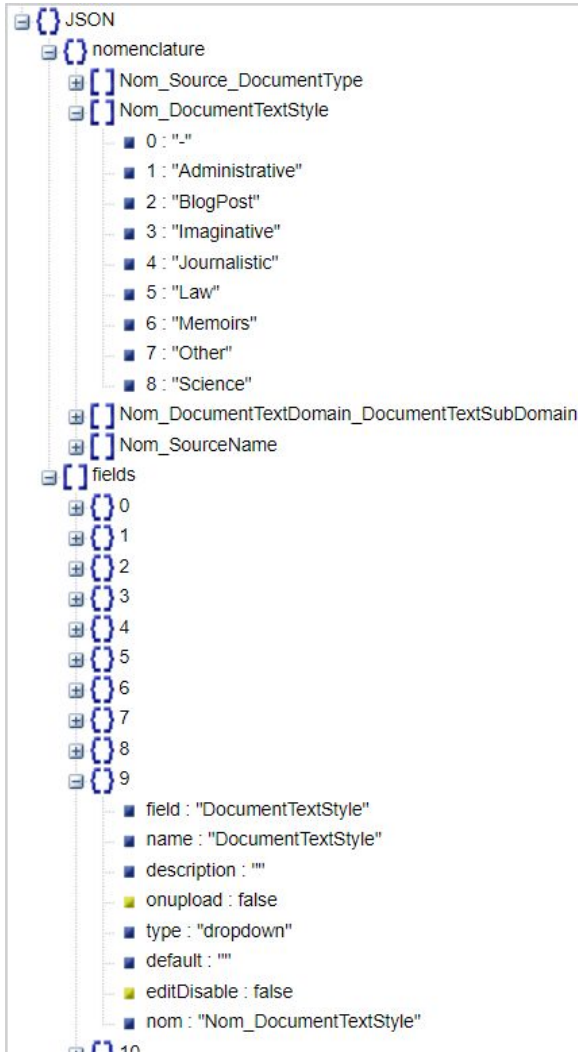
DocumentTitle	Pe urmele lui Creangă	Titlul documentului. Pentru o colecție: 'Titlul cărții/[Volumul I, vol. I]:Titlul volumului'. Exemplu: 'Chirurgie generală/Volumul II, ediția a II-a' sau 'Moromeții/vol. I' (Se va trece 'Volumul' sau 'Vol.' după cum este în documentul electronic)
ArticleTitle	Pe urmele lui Creangă	Pentru cărți introduceți valoarea din DocumentTitle
AuthorName	Bobu Victoria, Sturzu Mar	Autorul sau autorii
TranslatorName	-	Numele traducătorului sau '-'
PublicationDate	2011	Anul de publicare
Source	Publishing House ▾	

Standoff Metadata - CoRoLa

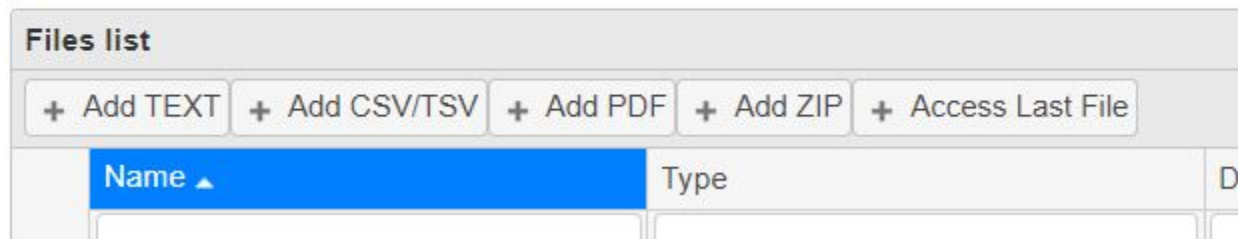
Pe urmele lui Creanga FINAL.xml

Save

DocumentTitle	Pe urmele lui Creangă	Titlul documentului. Pentru o colecție: 'I' ediția a II-a' sau 'Moromeții/vol. I' (Se va
ArticleTitle	Pe urmele lui Creangă	Pentru cărți introduceți valoarea din Doc
AuthorName	Bobu Victoria, Sturzu Mar	Autorul sau autorii
TranslatorName	-	Numele traducătorului sau '-'
PublicationDate	2011	Anul de publicare
Source	Publishing House	
SourceName	Editura Pim	
Medium	Written	
DocumentType	Book	
DocumentTextStyle	Other	
DocumentTextDomain	Arts and Culture	
DocumentTextSubDomain	Literature	
SubjectLanguage	Romanian	Dacă documentul este o traducere se tr
ISSN-ISBN	978-606-13-0496-7	Dacă exista 2 ISBN, se prefera ISBN 13
CollectionDate	2023	Anul în care se lucrează documentul



Import/Export



Import/Export

Corpus: test_COROLA2

View Corpora

Files

S

Files list

+ Add TEX

	Name
1	9942-p
2	Pe urn
3	Surasu
4	Surasu
5	aforism
6	dans.b
7	periplu
8	teoria

Double click

Add ZIP archive

Files will be distributed and processed according to their extension. Text from PDF files will be automatically extracted. XML, JSON files and unknown extensions will be considered standoff metadata.

File:

Choose File

No file chosen

FileName (optional):

Description:

Source

Blog

SourceName

SubjectLanguage

Romanian

Dacă documentul este o traducere se trece limba autorului, în caz contrar este limba română

CollectionDate

Anul în care se lucrează documentul

Add

Cancel

Import/Export

Add File CSV/TSV ✕

File:

Choose File

No file chosen

FileName (optional):

Description:

Delimiter:

,

(For TAB enter "\t")

Enclosure:

"

Escape:

\

Comment character:

Ignore rows:

(headers)

Columns w text:

(comma separated list, zero based)

Add

Cancel

- CSV/TSV for small texts
- each line is processed as an independent text document

Import/Export

- Files
- Standoff
- Tasks
- Annotated
- Statistics
- Archives
- Gold Standoff
- Gold Annotated

Corpus tasks											
+ TEPROLIN	+ UDPipe	+ NERRegex	+ Classify EuroVoc	+ Terminology	+ IATE/EuroVoc	+ Cleanup	+ NER Baseline	+ CoNLLUP2BRAT	+ NER_Old	+ TTLChunker	+ BRAT2CoNLLUP
+ Gold NE list	+ Statistics	+ ZIP Text	+ ZIP Annotated	+ ZIP Standoff	+ ZIP Gold Standoff	+ ZIP Gold Annotated	+ ZIP Audio	+ Change Terms Marcell	+ Export CURLICAT		
+ Export Marcell	+ NER-Legal	+ Translate Text	+ ASR	+ Punctuation	+ TTS	+ Translate S2S	+ DeAnonymization	+ NER-SiMoNERo	+ Anonymization	+ Create Empty Metadata	
Type	Status			Description			User		Creation Date		

Import/Export

Corpus: **CURLICAT_Anonymized**

[View Corpora](#)

[Files](#)

[Standoff](#)

[Tasks](#)

[Annotated](#)

[Statistics](#)

[Archives](#)

Archives

	File	Size
1	zip_text/CURLICAT.zip	266.18 Mb
2	zip_text/20220312_text_anonymized.zip	298.39 Mb
3	zip_text/meta.zip	17.57 Mb
4	zip_text/20221030_text.zip	242.9 Mb
5	zip_basic_tagging/20211227_annotated_before_chunking.zip	1.86 Gb
6	zip_basic_tagging/20211231_annotated_before_IATE.zip	2.09 Gb
7	zip_basic_tagging/annotated_fara_iate_20220321.zip	2.16 Gb
8	zip_basic_tagging/20220926_annotated.zip	1.8 Gb
9	zip_basic_tagging/20221030_annotated.zip	1.82 Gb
10	zip_standoff/20220114_standoff.zip	14 Mb
11	zip_standoff/metadate_curlicat.zip	14 Mb
12	zip_standoff/20221030_meta.zip	56.3 Mb
13	curlicat-out/ro-raw.zip	207 Mb
14	curlicat-out/ro-annotated.zip	1.23 Gb
15	curlicat-out/ro-xml.zip	1.79 Gb

Import/Export

- Files Standoff **Tasks** Annotated Statistics Archives

Corpus tasks

+ TEPROLIN

+ UDPipe

+ NERRegex

+ Classify EuroVoc

+ Terminology

+ IATE/EuroVoc

+ Cleanup

+ NER Baseline

+ CoNLLUP2BRAT

+ NER_Old

+ TTLChunker

+ BRAT2CoNLLUP

+ Gold NE list

+ Statistics

+ ZIP Text

+ ZIP Annotated

+ ZIP Standoff

+ ZIP Gold Standoff

+ ZIP Gold Annotated

+ ZIP Audio

+ Change Terms Marcell

+ Export CURLICAT

+ Export Marcell

+ NER-Legal

+ Translate Text

+ ASR

+ Punctuation

+ TTS

+ Translate S2S

+ DeAnonymization

+ NER-SiMoNERo

+ Anonymization

+ Create Empty Metadata

Type	Status	Description	User	Creation Date ▾
------	--------	-------------	------	-----------------

Import/Export

Corpus: **Marcell**

[View Corpora](#)

[Files](#)

[Standoff](#)

[Tasks](#)

[Annotated](#)

[Statistics](#)

[Archives](#)

Archives

	File	Size
1	zip_text/marcell_1990.zip	525.21 Mb
2	zip_basic_tagging/tagged_20191029.zip	556.54 Mb
3	zip_basic_tagging/tagged_20191221.zip	4.34 Gb
4	zip_basic_tagging/output2020.zip	3.14 Gb
5	zip_basic_tagging/marcell_annotated_clean_20200114.zip	1.25 Gb
6	zip_basic_tagging/marcell_annotated_clean_20200120.zip	1.26 Gb
7	zip_basic_tagging/marcell_annotated_clean_20200122.zip	3.64 Gb
8	zip_basic_tagging/ro-20200129-xml.zip	4.1 Gb
9	zip_basic_tagging/ro-20200129-annotated.zip	2.72 Gb
10	marcell-out/ro-raw.zip	512.28 Mb
11	marcell-out/ro-annotated.zip	2.55 Gb
12	marcell-out/ro-xml.zip	4.02 Gb

Multimodal Processing

- Corpora creation
 - speech, images, video (+ text)
- Modules
- Task-based processing

Speech - Corpora creation

Corpora

Corpora list		
+ Add		
	Name	Lang
1	coref_test	ro
2	termeni nature lematizare	ro
3	enrich4all covid segmentare	ro
4	test_marcell_annoate_andrei	ro
5	test_marcell_annotate_andrei_v2	ro
6	TWITTER_ANONYMIZED_20220725	ro
7	TWITTER_MERGED_20220712	ro
8	TWITTER Corpus 20220602	ro
9	LegalNERo_zenodo_v3	ro
10	BioRo_Extended	ro
11	corpus_punct	ro
12	corpus_wav	ro

Add Corpus

Name:

Language:

Description:

Audio: ☐

Gold Annotations: ☐

BRAT Profiles: ☐

Classification Profiles: ☐

Corrected Text: ☐

Add Cancel

Creation Date	
	2022-11-13
	2022-10-09
	2022-09-08
	2022-08-10
	2022-08-10
	2022-07-25
	2022-07-12
	2022-06-02
are c...	2022-03-29
	2022-02-14
	2022-01-28
	2022-01-28

Speech - Corpora creation

Corpus: **ROBIN Batch 4** File: **robin-batch-4.csv** [View Corpora](#)

[Back](#) [Download](#) [View as Text](#)

File View

	C1
1	Care e cel mai ieftin laptop ol viu, cu placă grafică dedicată Tesla ve o sută și șase gigabaiți RAM?
2	Care e cel mai scump laptop ol viu, cu placă grafică dedicată gi fors o mie șazezi ti ai și doi gigabaiți RAM?
3	Care e cel mai scump laptop huawei, cu placă grafică dedicată Tesla ve o sută și opt gigabaiți RAM?
4	Care e cel mai scump laptop Asus, cu placă grafică dedicată Radeon?
5	Care e cel mai scump laptop aisăr, cu placă grafică dedicată Tesla pe o sută și doi gigabaiți RAM?
6	Care e cel mai ieftin laptop acer, cu placă grafică dedicată Tesla ve o sută și șase gigabaiți RAM?
7	Care e cel mai scump laptop del, cu placă grafică dedicată gi fors o mie șazezi ti ai și opt gigabaiți RAM?
8	Care e cel mai scump laptop Xiaomi, cu placă grafică dedicată gi fors o mie optzeci și șase gigabaiți RAM?
9	Care e cel mai scump laptop Asus, cu placă grafică dedicată Tesla ve o sută și doi gigabaiți RAM?
10	Care e cel mai ieftin laptop Xiaomi, cu placă grafică dedicată gi fors o mie optzeci și șase gigabaiți RAM?
11	Mă poți ajuta?
12	Care e cel mai ieftin laptop aisăr, cu placă grafică dedicată Tesla ve o sută și șase gigabaiți RAM?
13	Ce laptop-uri biznes aveți?
14	Care e cel mai scump laptop huawei, cu placă grafică dedicată Tesla ve o sută și doi gigabaiți RAM?

Page **1** of 6 **20** 1 to 20 of 101

Păiș, Vasile and Ion, Radu and Barbu Mititelu, Verginica and Irimia, Elena and Mitrofan, Maria and Avram, Andrei (2021). ***ROBIN Technical Acquisition Speech Corpus***. Zenodo, <https://doi.org/10.5281/zenodo.4626539>

Păiș, Vasile and Ion, Radu and Avram, Andrei-Marius and Irimia, Elena and Mititelu, Verginica Barbu and Mitrofan, Maria (2021). ***Human-Machine Interaction Speech Corpus from the ROBIN project***. In Proceedings SPED, pp. 91-96. <https://ieeexplore.ieee.org/document/9587355>

Speech - Corpora creation

Corpus: **ROBIN Batch 4**

[View Corpora](#)

[Files](#)

[Standoff](#)

[Tasks](#)

[Annotated](#)

[Statistics](#)

[Archives](#)

[Audio](#)

[Recorder](#)

Audio			
	File	Size	
8	audio/andrei_66.wav	1.12 Mb	
9	audio/andrei_67.wav	1.16 Mb	
10	audio/andrei_68.wav	624.04 Kb	
11	audio/andrei_69.wav	1.18 Mb	
12	audio/andrei_70.wav	1.12 Mb	
13	audio/andrei_71.wav	840.04 Kb	
14	audio/andrei_72.wav	1.38 Mb	
15	audio/andrei_73.wav	720.04 Kb	
16	audio/andrei_74.wav	392.04 Kb	
17	audio/andrei_75.wav	680.04 Kb	
18	audio/andrei_76.wav	1.13 Mb	
19	audio/andrei_77.wav	1.29 Mb	
20	audio/andrei_78.wav	1.1 Mb	

Page 9 of 26 20 161 to 180 of 505

Speech - Corpora creation

Corpus: **test_audio2**

View Corpora

Back

Recording as: [vasile]

Care e cel mai scump laptop Xiaomi

START

STOP

Sentence 10 / 203

- This app was tested on the Chrome Browser.
- For better results use headphones with mic for recording and speak naturally.
- If you hear your own sound during recording, **TURN OFF YOUR COMPUTER SOUND.**

Speech - Modules

- ASR - DeepSpeech2

Avram, Andrei-Marius and Păiș, Vasile and Tufis, Dan (2020). *Romanian speech recognition experiments from the ROBIN project*. In Proceedings CONSILR, pp. 103--114,

<https://profs.info.uaic.ro/~consilr/2021/wp-content/uploads/2021/03/volum-ConsILR-v-4-final-revizuit.pdf#page=111>

Avram, Andrei-Marius and Păiș, Vasile and Tufis, Dan (2020). *Towards a Romanian end-to-end automatic speech recognition based on Deepspeech2*. In Proceedings of the Romanian Academy Series A. vol. 21, pp. 395--402,

https://acad.ro/sectii2002/proceedings/doc2020-4/11-Avram_Tufis.pdf

- ASR - WAV2VEC2

Avram, Andrei-Marius and Păiș, Vasile and Tufis, Dan (2022). *Self-Supervised Pre-Training in Speech Recognition Systems*. Chapter in Speech Recognition Technology and Applications (ed. Păiș, Vasile). Nova Science Publishers, pp. 27--56, <https://doi.org/10.52305/BKWM899>

Speech - Modules

RELATE

- TEPROLIN Service >
- CoRoLa >
- RoWordNet >
- Machine Translation >
- Speech ▾
 - Translate RO - EN
 - Translate EN - RO
 - ROBIN ASR
 - ROBIN ASR Development
 - ROBIN TTS
 - ASR WAV2VEC2
- CURLICAT Anonymization

☰ Romanian Portal of Language Technologies

Welcome  Vasile

WAV2VEC2 ASR (Automatic Speech Recognition) Development

File

Recording

Results

Selectati un fisier WAV asociat unui text in limba romana. Este indicat sa fie inregistrat cat mai clar.

This model is constructed based on fine-tuning a WAV2VEC2 model on a Romanian speech corpus ([RTASC](#)). We are working on an improved model making use of multiple resources.

Choose File

No file chosen

Run ASR

Scientific papers

Avram, Andrei-Marius and Păis, Vasile and Tufiş, Dan. 2022. [Self-Supervised Pre-Training in Speech Recognition Systems](#). In *Speech Recognition Technology and Applications*, pages 27-56, Nova Science Publishers.

Speech - Modules

- Punctuation Restoration

Păiș, Vasile and Tufiș, Dan (2022). *Capitalization and punctuation restoration: a survey*. Artificial Intelligence Review 55(3), 1681–1722, <https://rdcu.be/cpOch>

Păiș, Vasile (2022). *Punctuation Recovery for Romanian Transcribed Documents*. Chapter in Speech Recognition Technology and Applications (ed. Păiș, Vasile). Nova Science Publishers, pp. 119–154, <https://doi.org/10.52305/BKWM899>

Punctuation Restoration

Selectați modelul dorit și introduceți un document în limba română, fără punctuație și cu litere mici. Modelele bazate pe corpusul MARCELL presupun un text de tip legislație. Utilizarea de texte din alte domenii reduce calitatea rezultatelor. Nu sunt realizate curățări ale datelor, astfel încât textul trebuie să conțină doar caractere în limba română.

MARCELL Punctuation Restoration ▼

președintele a vizitat România Suedia și Danemarca

Punctuation

Demo text

Speech - Modules

- TTS - RACAI SSLA

Zafiu, Adrian and Dumitrescu, Ștefan Daniel and Boroș, Tiberiu (2015). *Modular Language Processing framework for Lightweight Applications (MLPLA)*. In Proceedings of LTC 2015.

Boroș, Tiberiu (2013). *Contributions to the modeling and implementation of Text-To-Speech Synthesis System*. Case Study: Romanian Language. PhD thesis, Romanian Academy.

Boroș, Tiberiu and Ion, Radu and Dumitrescu, Ștefan Daniel (2013). *The RACAI text-to-speech synthesis system*. In Blizzard Challenge Workshop 2013.

- TTS - ROBIN

Tufiș, Dan and Barbu Mititelu, Verginica and Irimia, Elena and Mitrofan, Maria and Ion, Radu and George, Cioroiu (2019). *Making Pepper Understand and Respond in Romanian*. In the 22nd International Conference on Control Systems and Computer Science.

- bazat pe TTS-Cube : <https://tiberiu44.github.io/TTS-Cube/>

Speech - Modules

- TTS - RACAI SSLA

Zafiu, Adrian and Dumitrescu, Ștefan Daniel and Boroș, Tiberiu (2015). *Modular Language Processing framework for Lightweight Applications (MLPLA)*. In Proceedings of LTC 2015.

Boroș, Tiberiu (2013). *Contributions to the modeling and implementation of Text-To-Speech Synthesis System*. Case Study: Romanian Language. PhD thesis, Romanian Academy.

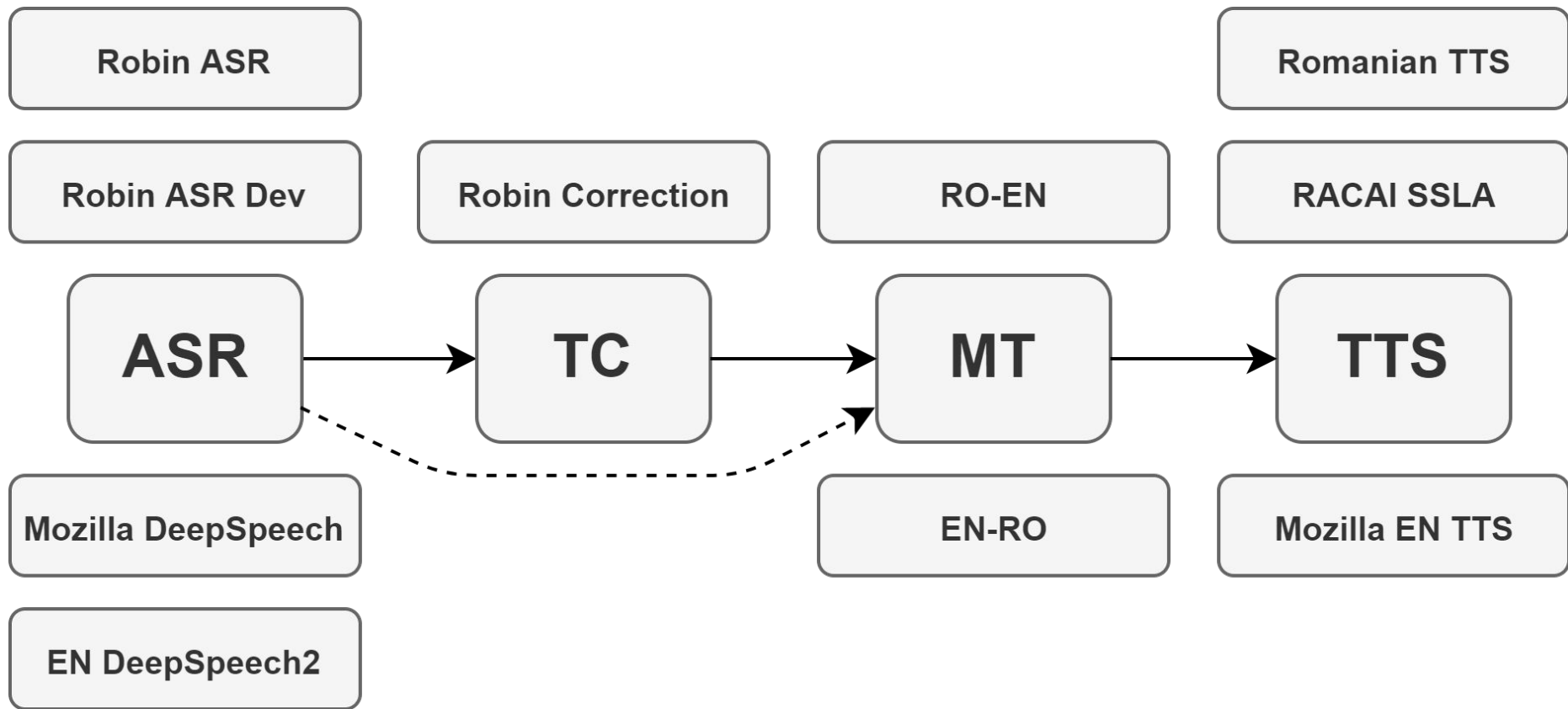
Boroș, Tiberiu and Ion, Radu and Dumitrescu, Ștefan Daniel (2013). *The RACAI text-to-speech synthesis system*. In Blizzard Challenge Workshop 2013.

- TTS - ROBIN - bazat pe TTS-Cube : <https://tiberiu44.github.io/TTS-Cube/>

Tușiș, Dan and Barbu Mititelu, Verginica and Irimia, Elena and Mitrofan, Maria and Ion, Radu and George, Cioroiu (2019). *Making Pepper Understand and Respond in Romanian*. In the 22nd International Conference on Control Systems and Computer Science.

- TTS - RomanianTTS - <http://romaniantts.com>

Adriana Stan, Junichi Yamagishi, Simon King, Matthew Aylett (2011) "The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate", In Speech Communication, vol. 53, no. 3, pp. 442-450.



A Modular Approach for Romanian-English Speech Translation, Avram, Andrei-Marius and Păiș, Vasile and Tufiș, Dan, NLDB 2021

Speech-to-Speech Translation

File

Recording

Results

Selecțai un fișier WAV asociat unui text în limba română. Este indicat să fie înregistrat cât mai clar.

Choose File

No file chosen

lanț de prelucrare:

ROBIN ASR

ROBIN Correction

RO Presidency

Mozilla EN TTS

Traducere

For this implementation we used the following:

- Romanian ASR from the ROBIN Project: Andrei-Marius Avram, Vasile Păiș, Dan Tufiș, 2020. Towards a Romanian end-to-end automatic speech recognition based on DeepSpeech2. Proc. Ro. Acad., Series A, Volume 21, No. 4, pp. 395-402.
- Romanian TTS from <http://romaniantts.com> : Adriana Stan, Junichi Yamagishi, Simon King, Matthew Aylett, "The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate", In Speech Communication, vol. 53, no. 3, pp. 442-450, 2011.
- Romanian SSLA TTS: Tiberiu Boroș, Ștefan D. Dumitrescu, Vasile Păiș, "Tools and resources for Romanian text-to-speech and speech-to-text applications", CoRR, vol. abs/1802.05583, 2018. <https://arxiv.org/pdf/1802.05583.pdf>
- Translation using the EU Council Presidency Translator developed by Tilde with support from RACAI during the Romanian presidency.
- English DeepSpeech2 ASR from: Amodei et al. 2016. Deep Speech 2 : End-to-End Speech Recognition in English and Mandarin. In Proceedings of The 33rd International Conference on Machine Learning (PMLR), 48:173-182.
- English Mozilla DeepSpeech ASR: Hannun et al. 2016. Deep Speech: Scaling up end-to-end speech recognition. arXiv:1412.5567 [cs.CL]
- English Mozilla TTS: <https://github.com/mozilla/TTS>

Text Translation

RELATE

TEPROLIN Service >

CoRoLa >

RoWordNet >

Machine Translation ▾

- Text RO - EN
- Text EN - RO
- Speech RO - EN
- Speech EN - RO
- Text Dev Documentation

Speech >

Downloads >

☰ Romanian Portal of Language Technologies

Machine translation

Introduceți un text în limba ROMÂNĂ

Submit

CEF Automated Translation
toolkit for the Rotating
Presidency of the Council of the
EU

Developed by Tilde with support
from ICIA during the Romanian
presidency

The API is integrated in
RELATE

Speech - Task-based processing

Files Standoff **Tasks** Annotated Statistics Archives Audio

Corpus tasks

+ TEPROLIN + UDPipe + NERRegex + Classify EuroVoc + Terminology + IATE/EuroVoc + Cleanup + NER Baseline + CoNLLUP2BRAT + NER_Old + TTLChunker + BRAT2Co
+ Gold NE list + Statistics + ZIP Text + ZIP Annotated + ZIP Standoff + ZIP Gold Standoff + ZIP Gold Annotated + ZIP Audio + Change Terms Marcell + Export CURLICAT
+ Export Marcell + NER-Legal + Translate Text + ASR + Punctuation + TTS + Translate S2S + DeAnonymization + NER-SiMoNERo + Anonymization + Create Empty Metadata

Type

Creation Date ▾

Add task translate S2S ✕

Perform speech to speech translation.

ASR RO DeepSpeech2 ▾

Punctuation MARCELL ▾

Translate RO-EN ▾

TTS EN Mozilla TTS ▾

Description:

Runners (optional):

Overwrite: ☐

Add Cancel

- The correct sequence for C
- The correct sequence for M

could be added one at a time and the results cl

Export Marcell. Tasks should be added one at

Image processing

Corpus: NewsImages_BATCH1_Vasile

View Corpora

Specific interface elements
activated for image datasets

Files Standoff Tasks Annotated Statistics Archives **Images** Properties Rights

Images			
+ Access Last File			
	File	Size	
1	image/BATCH1_00001.jpg	153.64 Kb	
2	image/BATCH1_00002.jpg	308.87 Kb	
3	image/BATCH1_00003.jfif	116.78 Kb	
4	image/BATCH1_00004.jfif	132.74 Kb	
5	image/BATCH1_00005.jpg	924.55 Kb	
6	image/BATCH1_00006.jpg	814.63 Kb	
7	image/BATCH1_00007.png	59.7 Kb	
8	image/BATCH1_00008.jpg	314.54 Kb	
9	image/BATCH1_00009.jpg	182.96 Kb	
10	image/BATCH1_00010.jpg	192.22 Kb	
11	image/BATCH1_00011.webp	375.94 Kb	
12	image/BATCH1_00012.jpg	752.16 Kb	
Page 1 of 30 20 1 to 20 of 600			

Image processing

Corpus: NewsImages_BATCH1_Vasile File: BATCH1_00005.jpg

[View Corpora](#)

[Back](#)

[Download](#)

[PREVIOUS](#)

[NEXT](#)

Realistic Image Generation

5-Excellent ▾

[Save](#)



Video processing

Corpus: **Test_Video_Corpus**

View Corpora

Specific interface elements
activated for video datasets

- Files
- Standoff
- Tasks
- Annotated
- Statistics
- Archives
- Videos
- Properties
- Rights

Video

+ Access Last File

	File	Size
1	video/04.mp4	1.97 Mb

Video processing

Corpus: **Test_Video_Corpus** File: 04.mp4

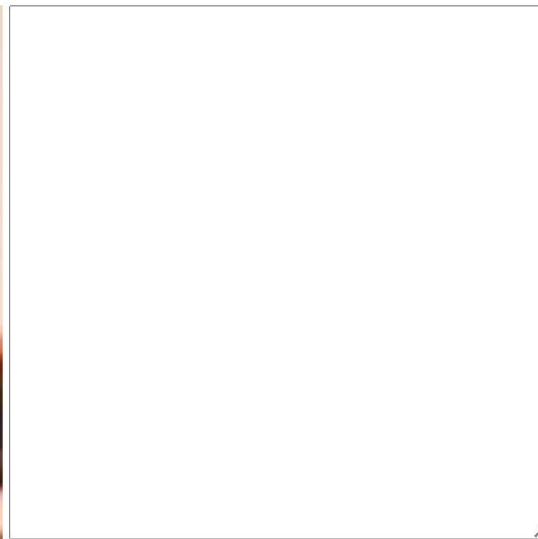
[View Corpora](#)

[Back](#)

[Download](#)

[PREVIOUS](#)

[NEXT](#)



Multimodal Statistics

File level information

Corpus: **Test_Audio_Corpus** File: **statistics/audio_list.csv**

View Corpora

[Back](#)

Download

[View as Text](#)

File View									
	file	duration_seconds	duration_formatted	channels	bits_per_sample	codec	sample_rate	mime	filesize
1									
2	vasile_102.wav	3.37	00:00:03.371	2	16	Pulse Code Modulation (PCM)	48000	audio/wav	647212
3	vasile_102_1.wav	3.37	00:00:03.371	2	16	Pulse Code Modulation (PCM)	48000	audio/wav	647212

Multimodal Statistics

File level information

Corpus: NewsImages_BATCH1_Vasile File: statistics/image_list.csv

[View Corpora](#)[Back](#)[Download](#)[View as Text](#)

File View

	file	width	height	mime	channels	bits	filesize
1	BATCH1_00001.jpg	1024	1024	image/jpeg	3	8	157329
2	BATCH1_00021.jpg	1024	1024	image/jpeg	3	8	150670
3	BATCH1_00041.jpeg	1000	1500	image/jpeg	3	8	282757
4	BATCH1_00061.jpg	1024	1024	image/jpeg	3	8	222206
5	BATCH1_00081.jpg	1024	1024	image/jpeg	3	8	154272
6	BATCH1_00101.jpg	1024	1024	image/jpeg	3	8	234513
7	BATCH1_00121.png	1024	1024	image/webp	3	8	78228
8	BATCH1_00141.jpg	1024	1024	image/jpeg	3	8	173911
9	BATCH1_00161.jpg	1024	1024	image/jpeg	3	8	321920
10	BATCH1_00181.jpg	1024	1024	image/jpeg	3	8	54455
11	BATCH1_00201.jpg	1024	1024	image/jpeg	3	8	359323
12	BATCH1_00221.jfif	896	1152	image/jpeg	3	8	89242

Multimodal Statistics

File level information

Corpus: NewsImages_BATCH1_Vasile File: statistics/image_list.csv

[View Corpora](#)[Back](#)[Download](#)[View as Text](#)

File View

	file	width	height	mime	channels	bits	filesize
1	BATCH1_00001.jpg	1024	1024	image/jpeg	3	8	157329
2	BATCH1_00021.jpg	1024	1024	image/jpeg	3	8	150670
3	BATCH1_00041.jpeg	1000	1500	image/jpeg	3	8	282757
4	BATCH1_00061.jpg	1024	1024	image/jpeg	3	8	222206
5	BATCH1_00081.jpg	1024	1024	image/jpeg	3	8	154272
6	BATCH1_00101.jpg	1024	1024	image/jpeg	3	8	234513
7	BATCH1_00121.png	1024	1024	image/webp	3	8	78228
8	BATCH1_00141.jpg	1024	1024	image/jpeg	3	8	173911
9	BATCH1_00161.jpg	1024	1024	image/jpeg	3	8	321920
10	BATCH1_00181.jpg	1024	1024	image/jpeg	3	8	54455
11	BATCH1_00201.jpg	1024	1024	image/jpeg	3	8	359323
12	BATCH1_00221.jfif	896	1152	image/jpeg	3	8	89242

Multimodal Statistics

Dataset level information

Corpus: **Test_Audio_Corpus**

[View Corpora](#)

[Files](#)

[Standoff](#)

[Tasks](#)

[Annotated](#)

[Statistics](#)

[Archives](#)

[Audio](#)

[Properties](#)

[Rights](#)

Statistics

[+ Download Stats](#) [+ WordForm Stats](#) [+ Words First Upper](#) [+ Words First Lower](#) [+ Lemma Stats](#) [+ WordForm Doc Freq](#) [+ Letters Stats](#) [+ Lemma UPOS Stats](#) [+ IATE](#)
[+ EUROVOC MT Stats](#) [+ EUROVOC MT Doc Freq](#) [+ Text Data](#) [+ Annotated Data](#) [+ Image Data](#) [+ Audio Data](#)

	Key	Value
	audio	
1	audio.bits_per_sample.16	2
2	audio.bytes	1294424
3	audio.channels.2	2
4	audio.codec.Pulse Code Modulation (PCM)	2
5	audio.duration_formatted	00:00:06.741
6	audio.duration_seconds	6.74
7	audio.mime.audio/wav	2
8	audio.number	2
9	audio.sample_rate.48000	2

[1](#) [<<](#) Page **1** of 1 [>>](#) [20](#) [⌵](#) 1 to 9 of 9

Multimodal Statistics

Dataset level information

Statistics		
+ Download Stats + WordForm Stats + Words First Upper + Words First Lower + Lemma Stats + WordForm Doc Freq + Letters Stats + Lemma UPOS Stats + IATE Stats + IATE Doc Freq		
+ Annotated Data + Image Data + Audio Data		
	Key	Value
	image	
1	image.bits.8	600
2	image.bytes	175767552
3	image.channels.3	600
4	image.height.1024	379
5	image.height.1152	82
6	image.height.1418	3
7	image.height.1440	23
8	image.height.1500	22
9	image.height.1536	4
10	image.height.2048	1
11	image.height.2304	80
12	image.height.2560	1
13	image.height.512	4
14	image.height.6000	1
15	image.max_height	6000
16	image.max_width	4000
17	image.mime.image/jpeg	501
18	image.mime.image/webp	99
19	image.min_height	512
20	image.min_width	512
21	image.number	600

RoMEMEs dataset



- usually funny
- tend to spread on social media
- image(s) + text
- may be targeted against individuals or groups

Indicator	Value
Number of memes	462
Min Width	259
Max Width	3,839
Min Height	194
Max Height	3,166
Complexity Simple	358
Complexity Combination	104
Real Images	293
Fake Images	153
Deep Fake Images	16

Indicator	Value
Polarity Negative	221
Polarity Positive	89
Polarity Neutral	152
Emotion Anger	79
Emotion Fear	38
Emotion Joy	87
Emotion Love	14
Emotion Sadness	106
Emotion Surprise	138
Political	147

RoMEMES

Dataset in Zenodo <https://zenodo.org/records/13120216>



CommunitiesMy dashboard



 vasilie@ra...

Published July 29, 2024 | Version v1

DatasetOpen

 Edit New version Share**71**
VIEWS**13**
DOWNLOADS Show more details

Versions

Version v1Jul 29, 2024
10.5281/zenodo.13120216**Cite all versions?** You can cite all versions by using the DOI [10.5281/zenodo.13120215](https://doi.org/10.5281/zenodo.13120215). This DOI represents all versions, and will always resolve to the latest one. [Read more.](#)

Files

RoMEMES

Pâiș, Vasile (Contact person)¹ ; Niță, Sara (Data collector)¹; Jerpelea, Alexandru (Data collector)¹; Pană, Luca (Data collector)¹; Cărea, Eric (Data collector)¹

Show affiliations

RoMEMES is a dataset of Romanian language memes, collected from public social media platforms. The dataset was manually annotated with: associated text in Romanian language, image complexity, sentiment, political content. In addition, the dataset contains associated metadata and the text part was automatically annotated in the RELATE platform with part-of-speech tags, lemmas, and dependency parsing.

Files and folders in this dataset:

- metadata.tsv - contains metadata and annotations; the first column is the file ID
- text - folder containing text files, following the file naming convention ID.txt
- images - folder containing image files, following the file naming convention ID.extension, where extension is the original file extension (sometimes this may not correspond with the mime/type of the file, as indicated in metadata.tsv)
- conllup - folder containing automatic text annotations, created in the RELATE platform, following the file naming convention ID.conllup

Paper presented at CONSILR 2024

RoMEMEs v2 - at SPED 2025

Cluj 19-22
October with
LDS Workshop

Corpus: ROMEMES_v2_studenti File: 00600006.jpg

[View Corpora](#)

[Back](#)

[Download](#)

[PREVIOUS](#)

[NEXT](#)

Complexity

Combination

Real/Fake

Real

Polarity

Neutral

Sentiment

Joy

Political

Yes

[Save](#)



-Pot să îți copiezi tema? -Da, dar ai grijă să nu semene. -OK

Linked Data

<https://www.racai.ro/p/lod/>

- Corpora converted to linked data (RDF)
- Apache Jena Fuseki available in RELATE
 - several Romanian corpora loaded in Fuseki
- RELATE supporting RDF in certain places

Barbu Mititelu, Verginica and Irimia, Elena and Păiș, Vasile and Avram, Andrei-Marius and Mitrofan, Maria (2022). ***Use case: Romanian language resources in the LOD paradigm***. In Proceedings of the Linked Data in Linguistics Workshop, pp. 35--44, <http://www.lrec-conf.org/proceedings/lrec2022/workshops/LDL/2022.ldl2022-1.0.pdf#page=43>

Barbu Mititelu, Verginica and Irimia, Elena and Păiș, Vasile and Mitrofan, Maria and Avram, Andrei-Marius and Curea, Eric (2021). ***Linked open data resources for Romanian***. In Proceedings CONSILR. pp. 7--19, https://profs.info.uaic.ro/~consilr/2022/wp-content/uploads/2022/04/consilr2021_14_03_2022_P.pdf#page=15

Linked Data

[Back](#)[View as RDF](#) [View as JSON](#)

Synset: RO30-09728285-n român

(Noun) Persoană care aparține populației de bază a României sau care este

- hypernym [RO30-09686536-n european](#)
- member_holonym [RO30-08813978-n România](#)

[Back](#)

```
@prefix dc: <http://purl.org/dc/terms/> .
@prefix ili: <http://ili.globalwordnet.org/ili/> .
@prefix lime: <http://www.w3.org/ns/lemon/lime#> .
@prefix ontolex: <http://www.w3.org/ns/lemon/ontolex#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix schema: <http://schema.org/> .
@prefix skos: <http://www.w3.org/2004/02/skos/core#> .
@prefix synsem: <http://www.w3.org/ns/lemon/synsem#> .
@prefix wn: <http://wordnet-rdf.princeton.edu/ontology#> .
@prefix pwnlemma: <http://wordnet-rdf.princeton.edu/rdf/lemma/> .
@prefix pwnid: <http://wordnet-rdf.princeton.edu/id/> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
```

```
<#român-n>
  ontolex:canonicalForm [
    ontolex:writtenRep "român"@ro
  ] ;
  ontolex:sense <#român-09728285-n> ;
  wn:partOfSpeech wn:noun ;
  a ontolex:LexicalEntry .

<#român-09728285-n>
  ontolex:isLexicalizedSenseOf pwnid:09728285-n ;
  a ontolex:LexicalSense .

pwnid:09728285-n
  wn:partOfSpeech wn:noun ;
  wn:definition [ rdf:value "Persoană care aparține populației de bază a României" ] ;
  wn:09686536-n pwnid:hypernym ;
  wn:08813978-n pwnid:member_holonym ;
  a ontolex:LexicalConcept .
```

Linked Data

<https://relate.racai.ro/datasets/>

Allows interrogation via
SPARQL queries

relate.racai.ro/datasets/

 Apache Jena Fuseki

  dataset  manage datasets  help

Apache Jena Fuseki

Version 4.0.0. Uptime: 6d 2h 11m 45s

Datasets on this server

dataset name	actions
/legalnero	 query  add data  info
/parseme	 query  add data  info
/rolex	 query  add data  info
/rown	 query  add data  info
/rtasc	 query  add data  info
/simonero	 query  add data  info

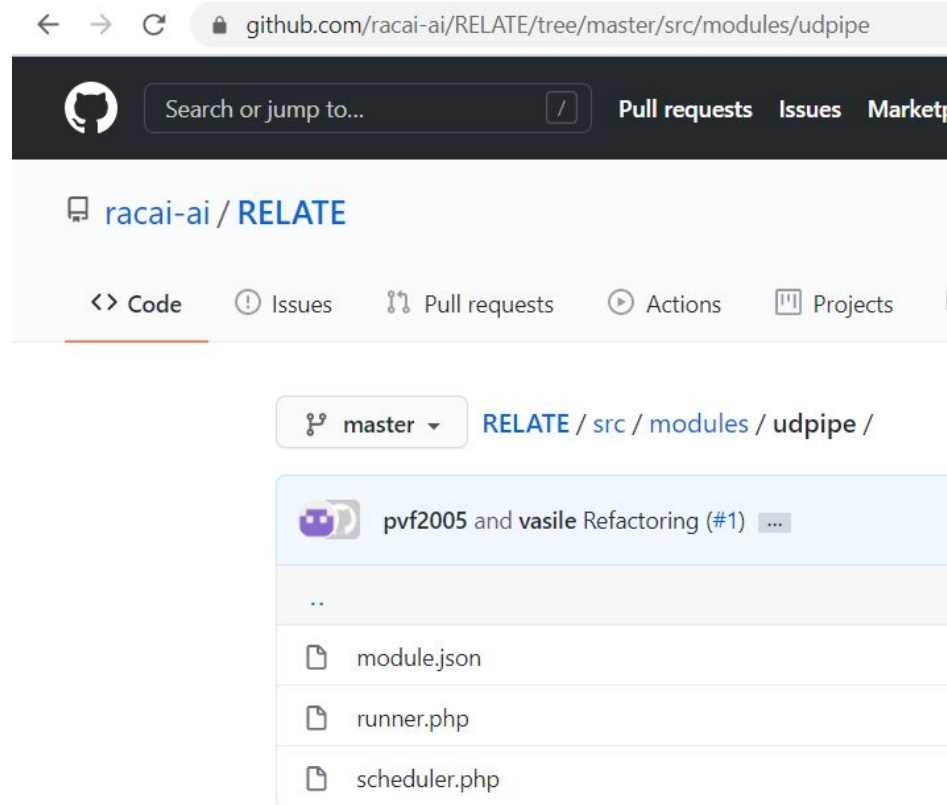
Păiș, Vasile and Barbu-Mititelu, Verginica (2022). ***Linguistic Linked Open Data for Speech Processing***. Chapter in Speech Recognition Technology and Applications (ed. Păiș, Vasile). Nova Science Publishers, pp. 155--188, <https://doi.org/10.52305/BKWM899>

RELATE development

- open source

<https://github.com/racai-ai/RELATE>

- component-based
 - each functionality developed as a separate component
 - usually no direct interactions between components



LDS Workshop - 22 October 2025 - Cluj



Collocated with SPED 2025



https://language-data-space.ec.europa.eu/events/lds-country-workshop-romania-2025-10-22_en



LDS Workshop

