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/

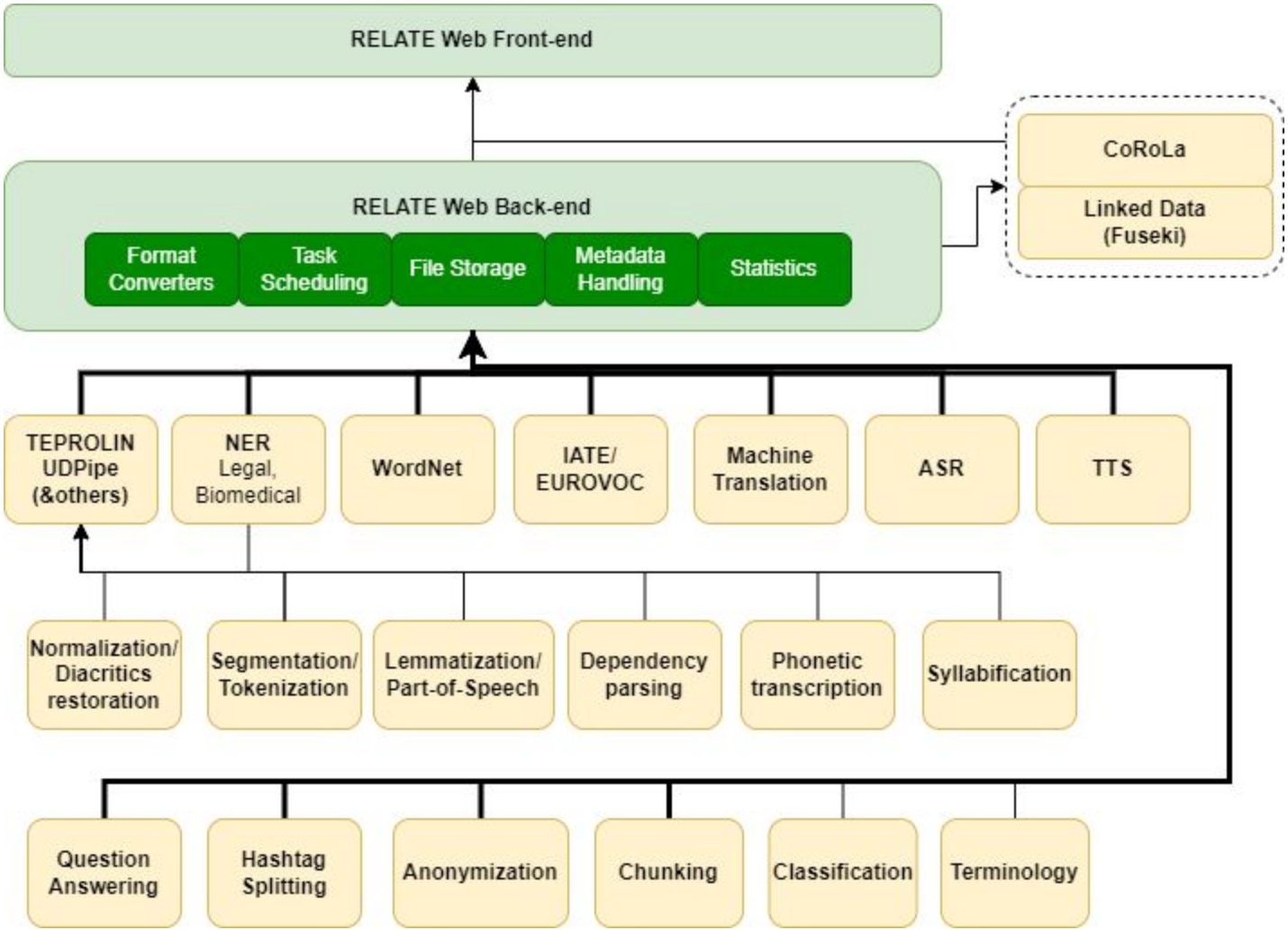
NATO OTAN
This project is supported by: The NATO Science for Peace and Security Programme

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.iwltp-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 .


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

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 F1	Sentences F1	Lemma Acc	UPOS Acc	XPOS 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

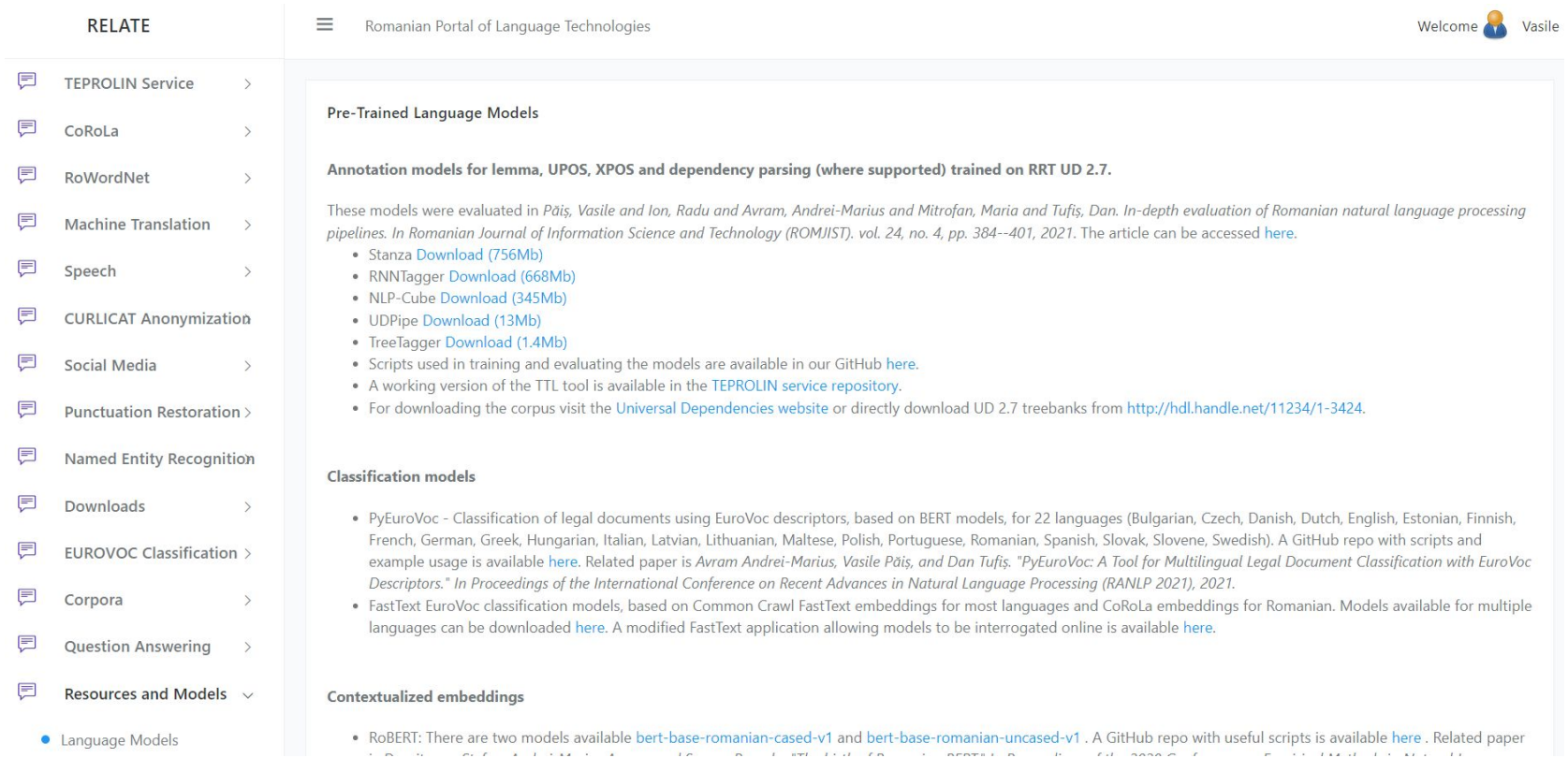
Table 6: Training duration and running speed considering tokenization and complete annotation for RRT-Test and SiMoNERo

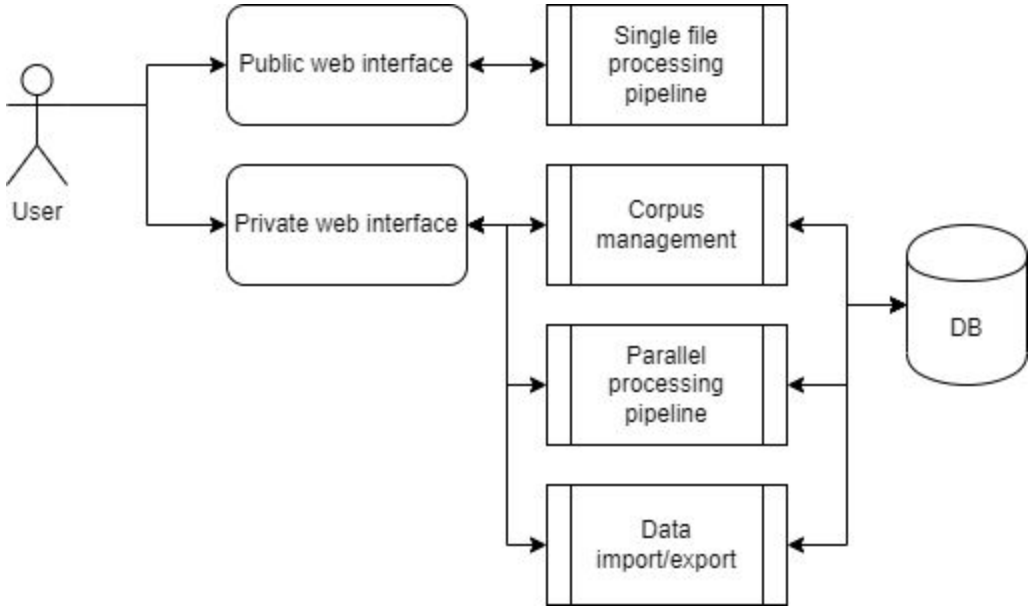
	RRT Train	RRT-test		RRT All	SiMoNERo	
	Train time	Tokenize tok/s	Run tok/s	Train time	Tokenize tok/s	Run tok/s
TreeTagger	70s	32648	**8162**	**21s**	43201	**24337**
TTL	**58s**	2332	583	63s	279	202
UDPipe	3h3m	9775	2720	3h15m	22024	2585
NLP-Cube	13h20m	191	60	14h54m	188	64
RNNTagger	1d3h44m	**42958**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

	UPOS Acc	XPOS 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**

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

Language resources and pre-trained models




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

RELATE - public view

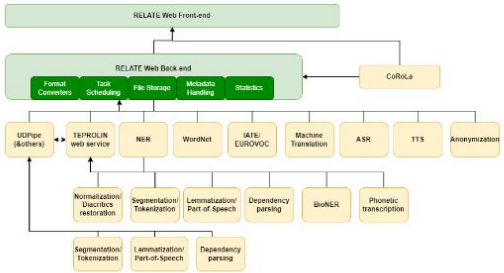
RELATE

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

Romanian Portal of Language Technologies

Login

RELATE - Romanian Portal of Language Technologies



RELATE Architecture
(click to zoom)

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.

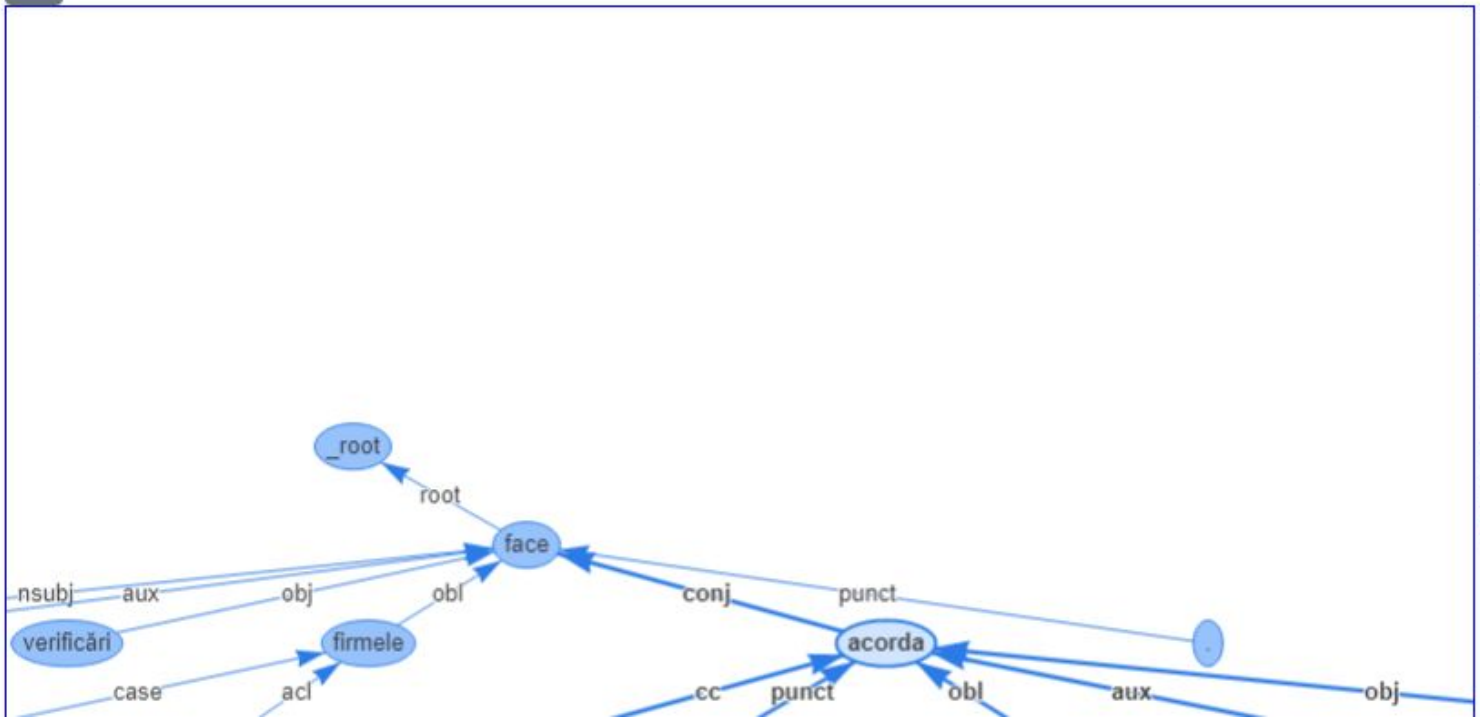
Complete run

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

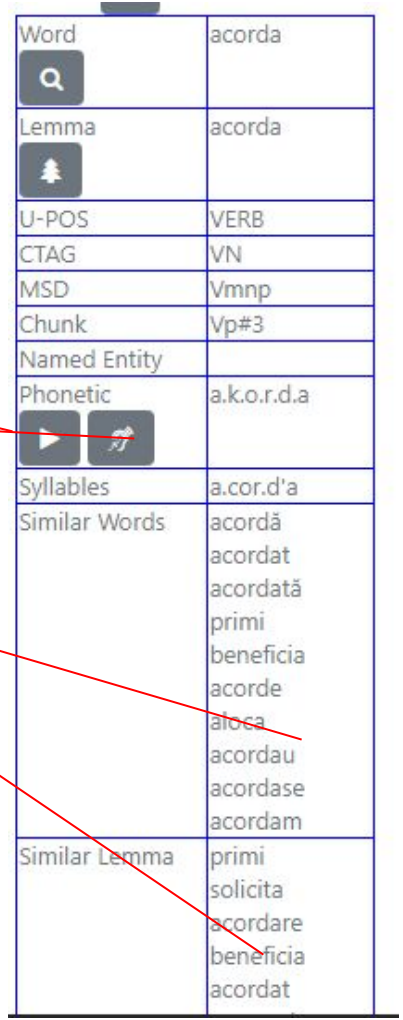
Search in Korap

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

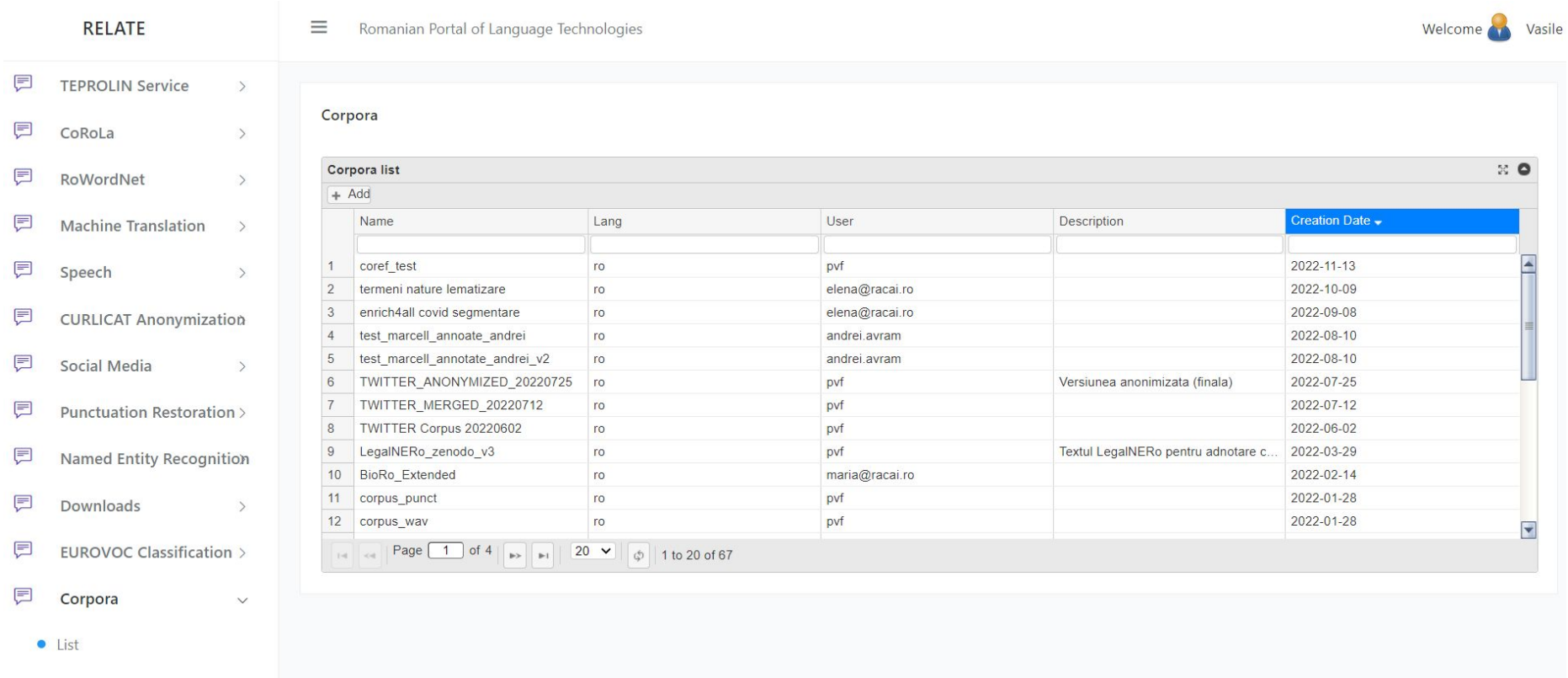
Media type:

- [ ] Text
- [ ] Speech
- [ ] Image

Filter

http://relate.racai.ro/index.php?path=teprolin/doc_dev

RELATE - authenticated view



Corpus Marcell | Corpora

Files | Tasks | Basic Tags | Statistics | Archives

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_00000G3W547PDTJOUI01U6U...	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_00000G3W53QZCYSPF6P15G...	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=Masc|Number=Sing	2	nmod	
9	7	2007	2007	NUM	Mc	_	6	nummod	
10									
11	# sent_id = ro_legal.2								

Page 1 of 18 20 1 to 20 of 350

RELATE - Large corpora

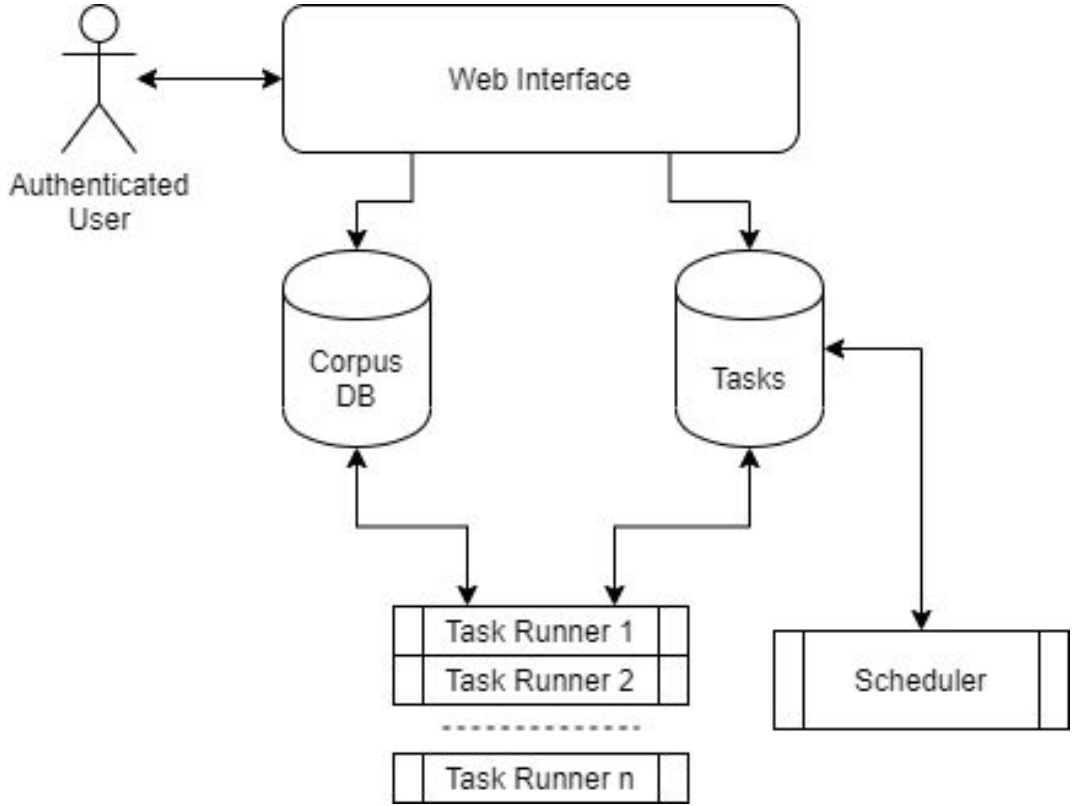
Efficiently stores and handles large corpora:

- 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

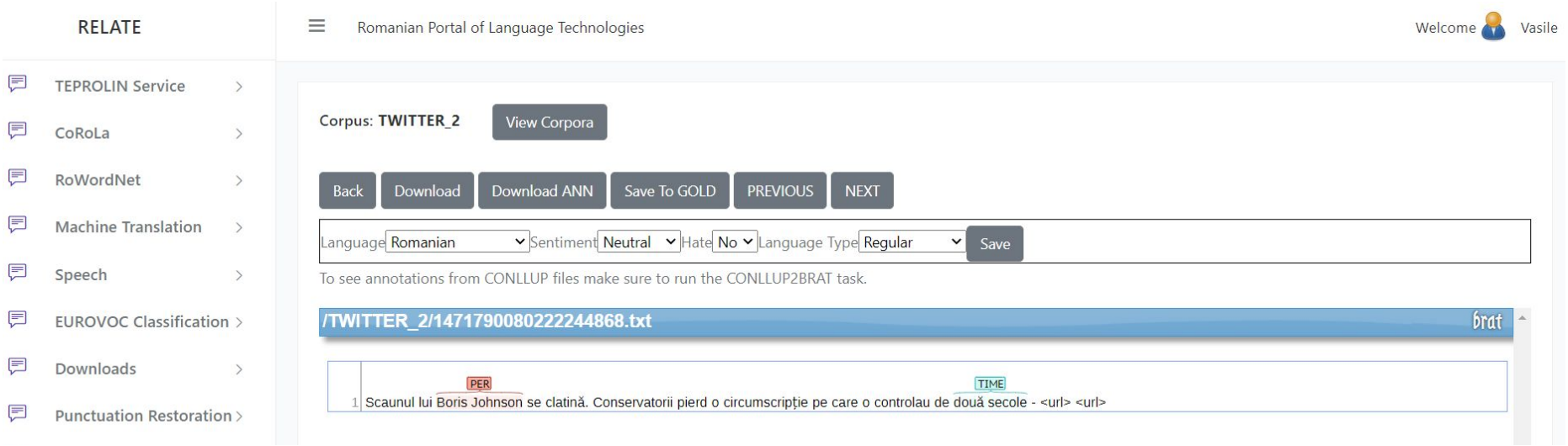
- **Anonymization**: from the CURLICAT project
- **Legal NER**: Person, Organization, Location, Time, Legal references (from a gold annotated sub-corpus derived from the MARCELL corpus)
- **Pre-trained language models**: available for download
- **Standard formats**: CoNLL-U, CoNLL-U Plus, XML, JSON,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

- **A modular architecture which continues to expand with new projects**

Text processing

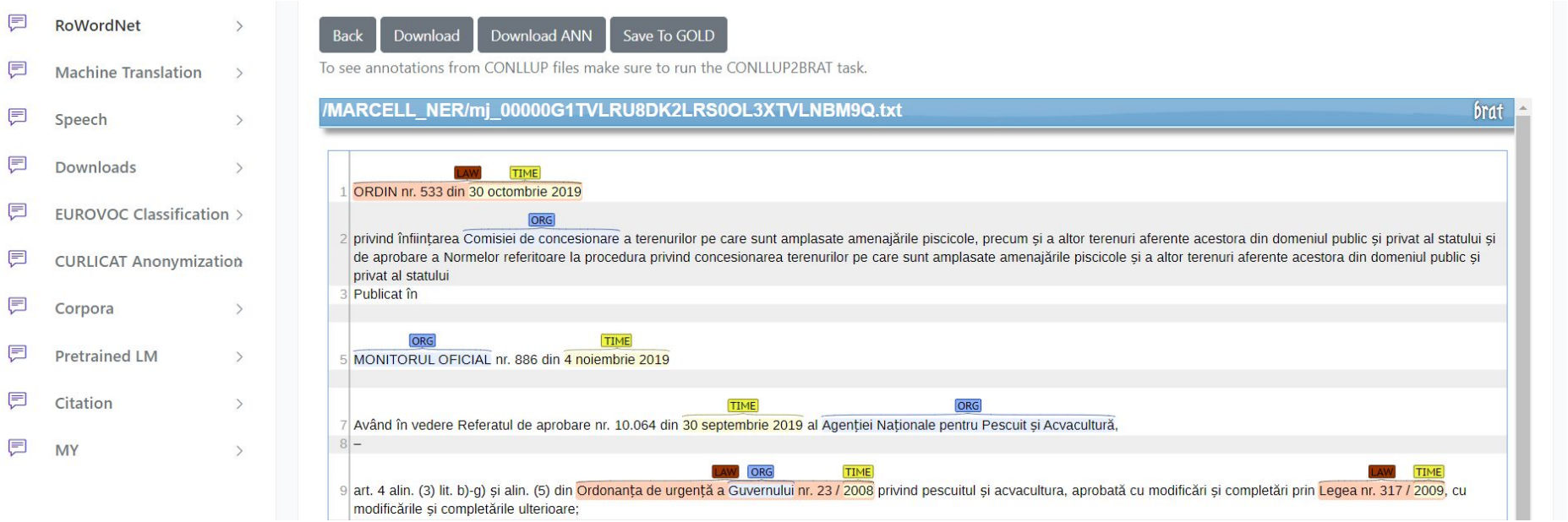
- Gold corpora creation
- Task-based processing
- Standoff Metadata
- Import/Export

Gold corpora creation - NER - MicroBloggingNERo



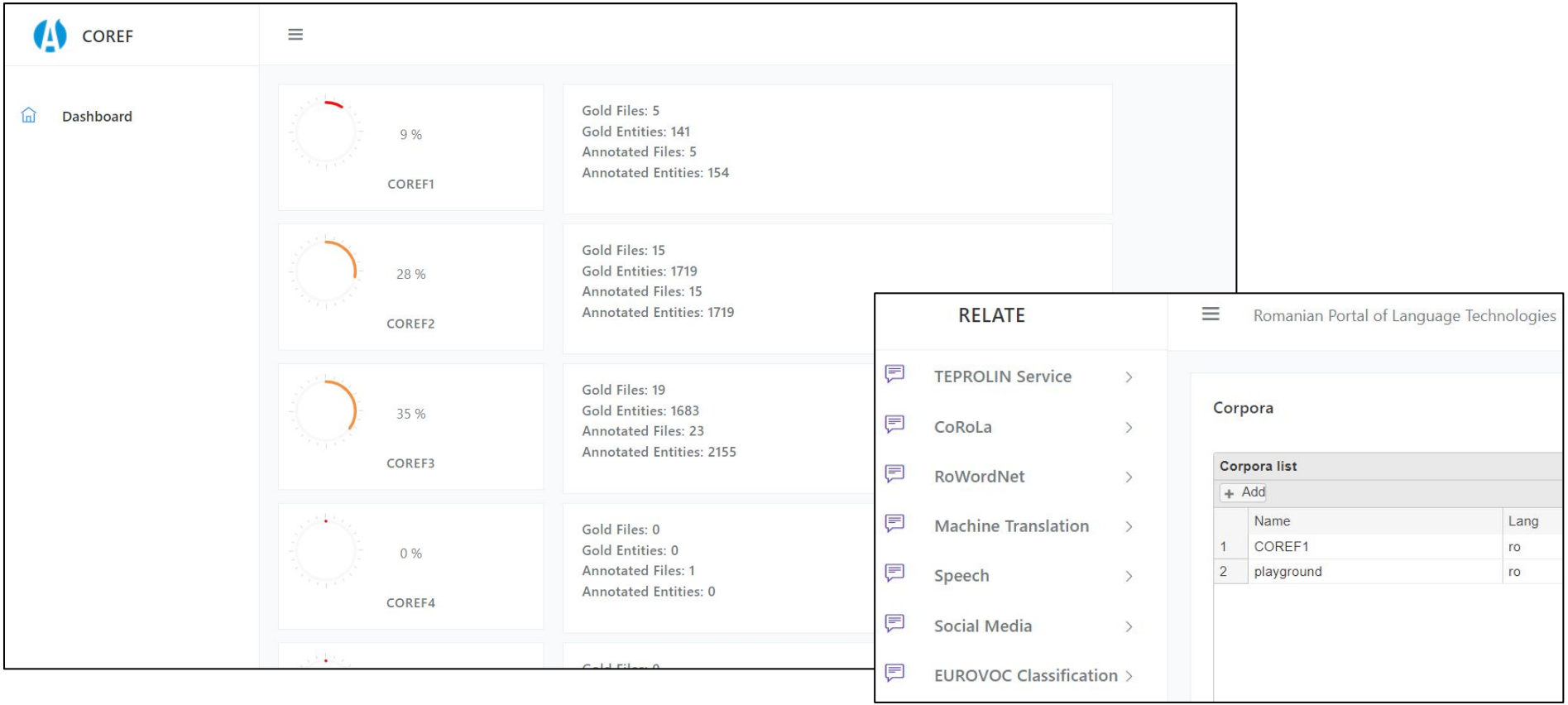
- Păiș, Vasile and Mitrofan, Maria and Barbu-Mititelu, Verginica and Irimia, Elena and Gasan, Carol Luca and Micu, Roxana and Marin, Laura and Dicusar, Maria and Florea, Bianca and Badila, Ana (2022). *Romanian micro-blogging named entity recognition (MicroBloggingNERo)*. Zenodo, https://doi.org/10.5281/zenodo.6905235
- 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

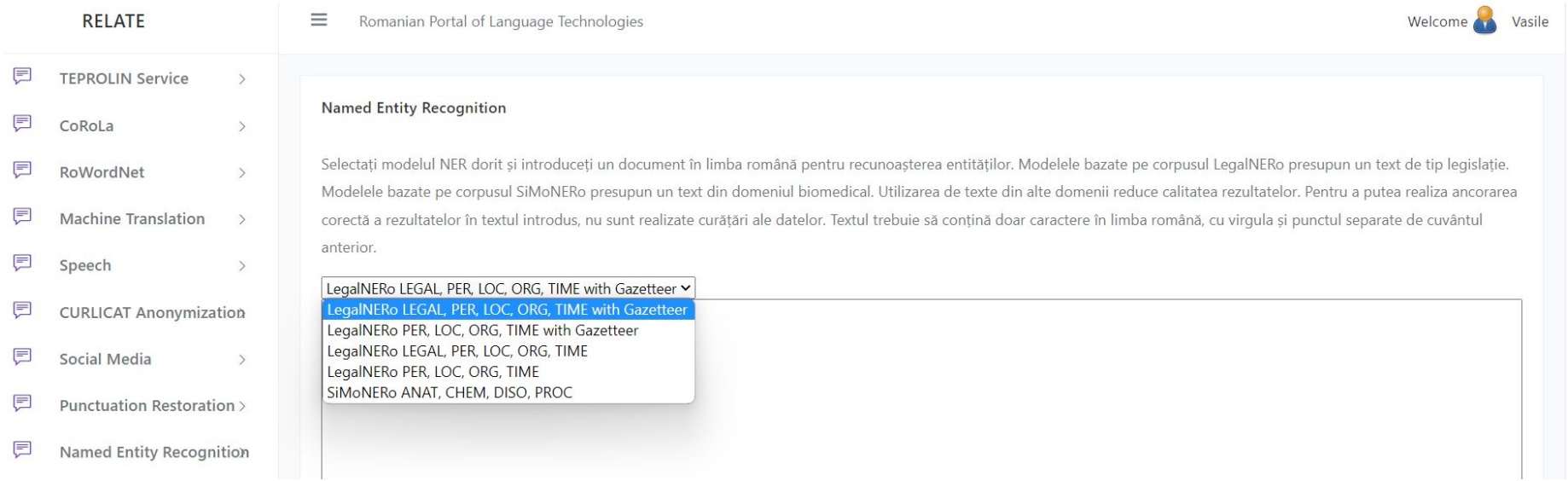


- Păiș, Vasile and Mitrofan, Maria and Gasan, Carol Luca and Ianov, Alexandru and Ghiță, Corvin and Coneschi, Vlad Silviu and Onuț, Andrei (2021). *Romanian Named Entity Recognition in the Legal domain (LegalNERo)*. Zenodo, https://doi.org/10.5281/zenodo.4772094
- Păiș, Vasile and Mitrofan, Maria. (2021) *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



- 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

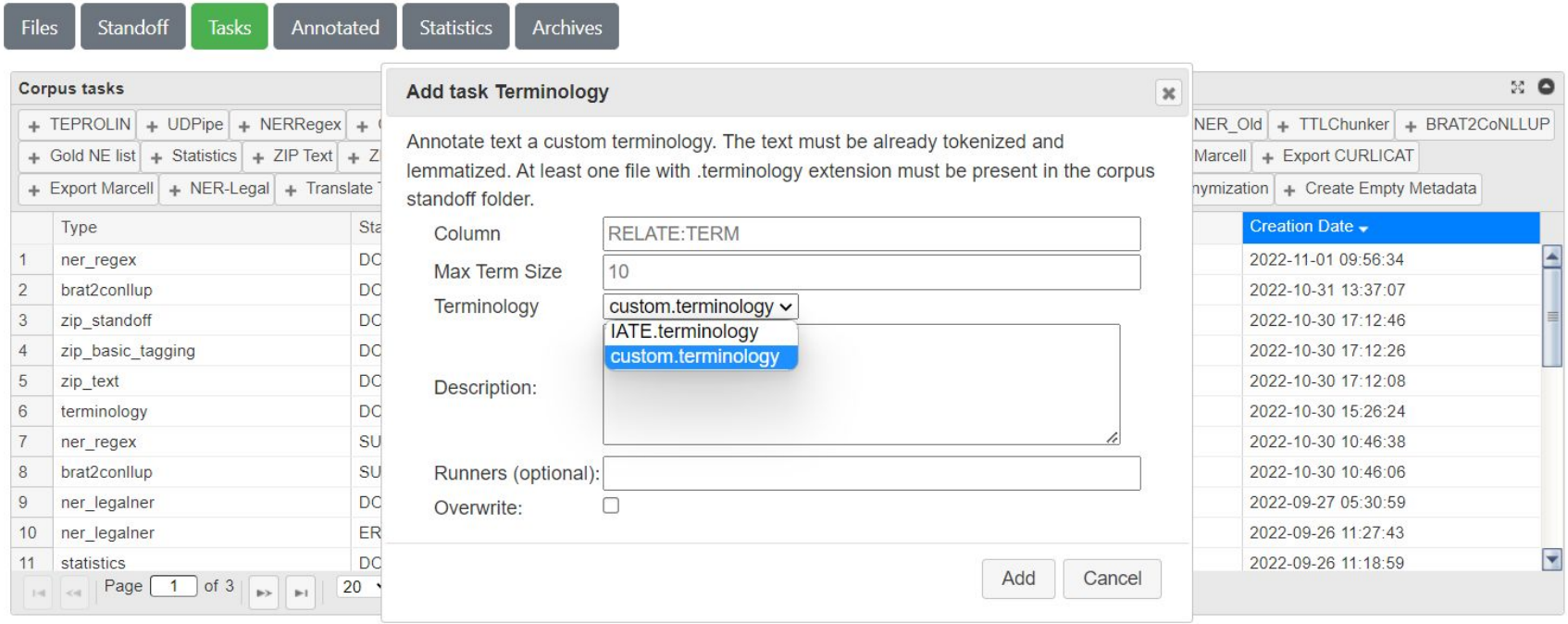
Corpus tasks

+ 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 Pęzik and Barbu Mititelu, Verginica and Radu Ion and Irimia, Elena and Mitrofan, Maria and Păiș, Vasile and Tufiș, 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/
- Tufiș, Dan and Mitrofan, Maria and Păiș, 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



- Váradi, Tamás and Nyéki, Bence and Koeva, Svetla and Tadić, Marko and Štefanec, Vanja and Ogrodniczuk, Maciej and Nitoń, Bartlomiej and Pęzik, 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 | Annotated | 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
```

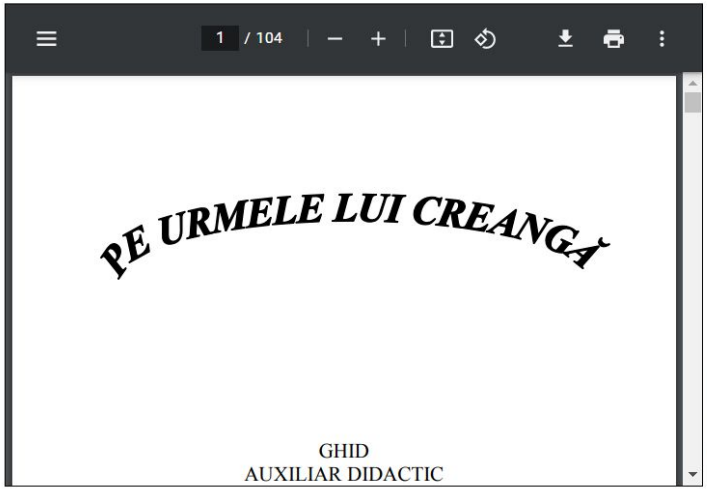
1021_a_2529.xml Save

```
<?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




```
            GHID
       AUXILIAR DIDACTIC
PENTRU ÎNVĂȚĂMÂNTUL PREȘCOLAR



     Prof. BOBU VICTORIA
     Prof. STURZU MARIANA



    Iași, Editura PIM, 2011
Editura PIM
Editură acreditată CNCSIS – 66/2010
Șoseaua Ștefan cel Mare și Sfânt nr. 4, Iași – 700497
Tel.: 0730.086.676; Fax: 0332.440.730
```

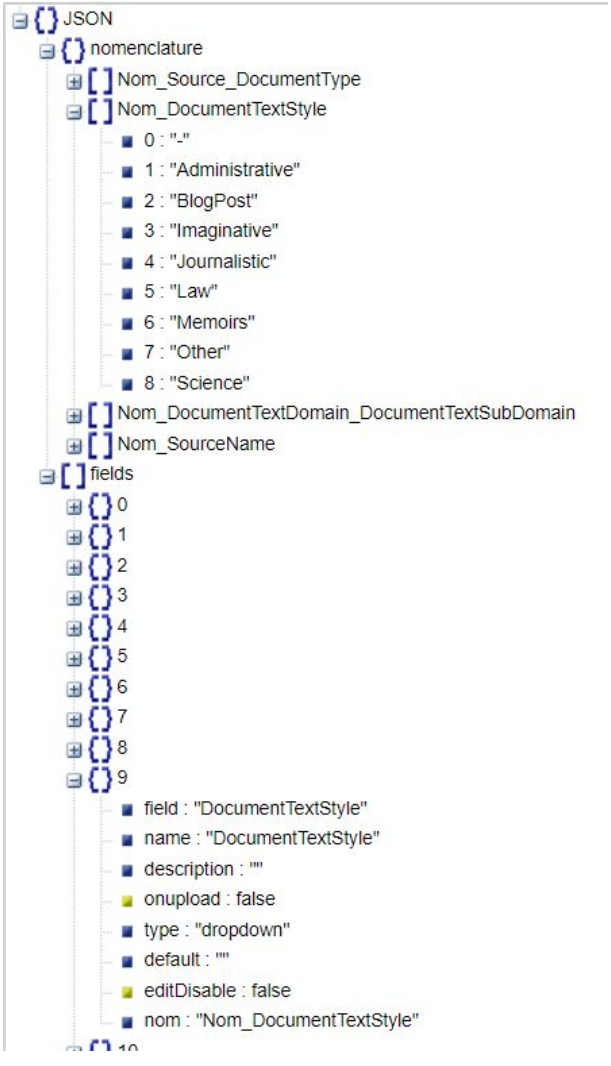
Pe urmele lui Creanga FINAL.xml Save

DocumentTitle	Pe urmele lui Creangă	Titlul documentului. Pentru o colecţie: 'Titlul cărţii/[Volumul I, vol. I]:Titlul volumului'. Exemplu: 'Chirurgie generală/Volumul II, ediţia a II-a' sau 'Moromeţii/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	

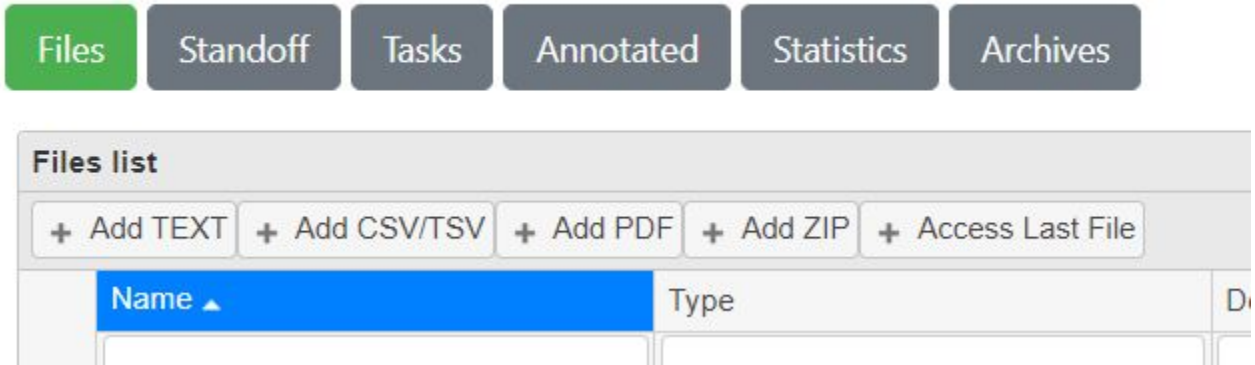
Standoff Metadata - CoRoLa

Pe urmele lui Creanga FINAL.xml Save

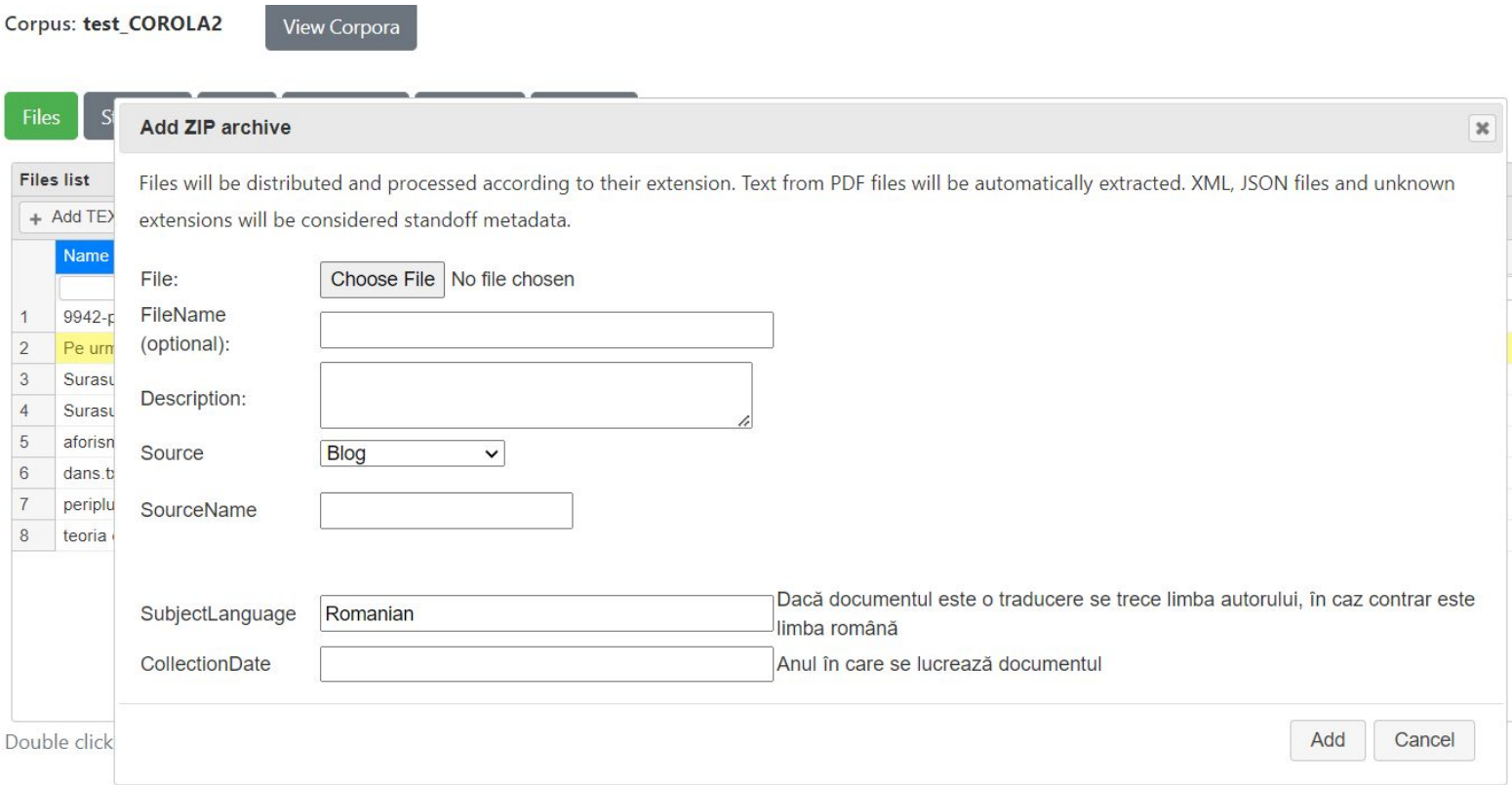
DocumentTitle	Pe urmele lui Creangă	Titlul documentului. Pentru o colecţie: 'T ediţia a II-a' sau 'Morometii/vol. I' (Se va
ArticleTitle	Pe urmele lui Creangă	Pentru cărţi introduceţi valoarea din Doc
AuthorName	Bobu Victoria, Sturzu Mar	Autorul sau autorii
TranslatorName	-	Numele traducătorului sau '-'
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 tre
ISSN-ISBN	978-606-13-0496-7	Dacă exista 2 ISBN, se prefera ISBN 13
CollectionDate	2023	Anul în care se lucrează documentul



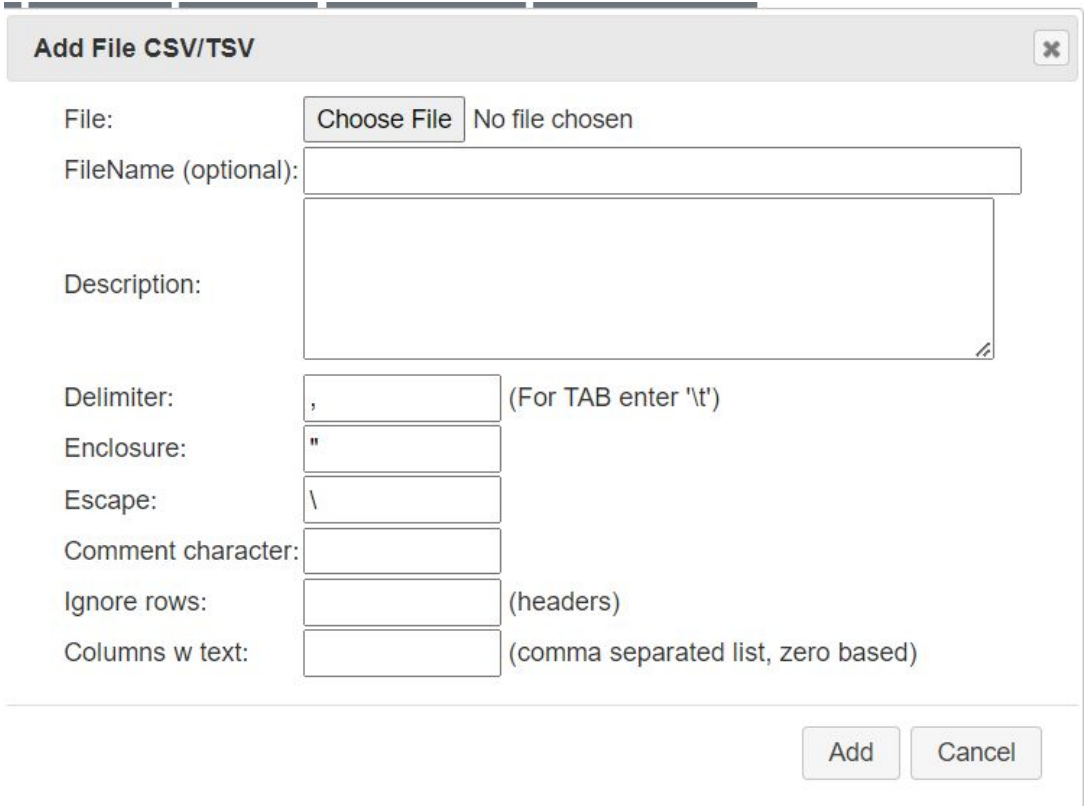
Import/Export



Import/Export

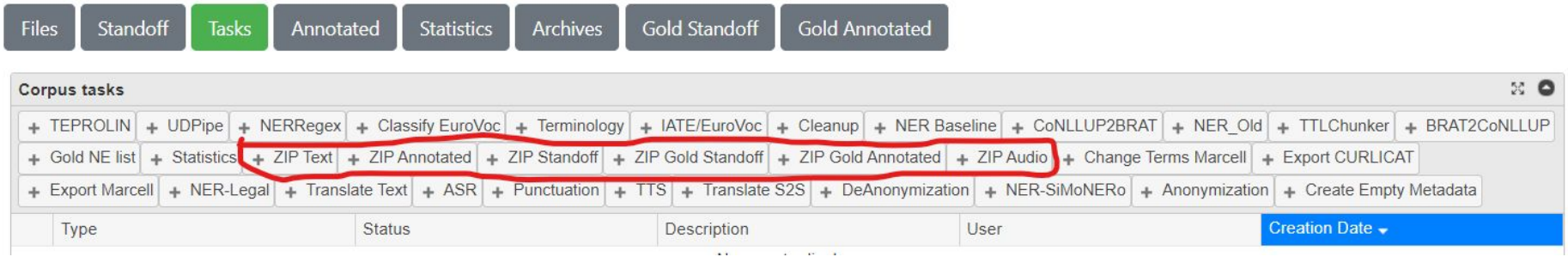


Import/Export



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

Import/Export



Import/Export

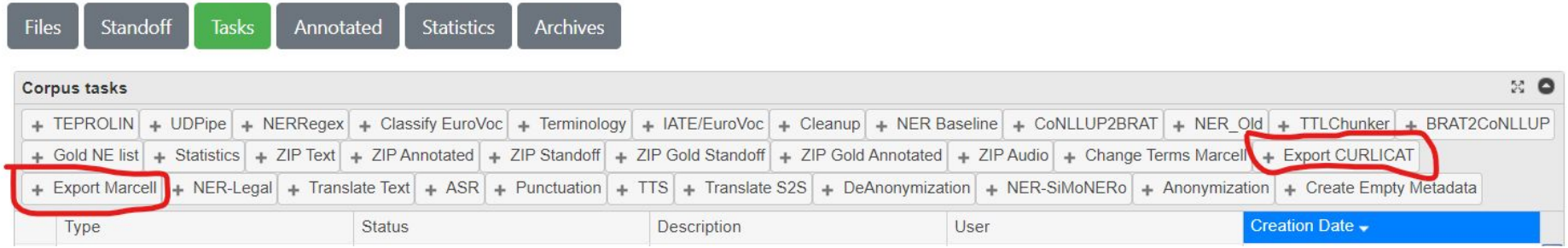
Corpus: **CURLICAT_Anonymized** View Corpora

Files | Standoff | Tasks | Annotated | Statistics | Archives

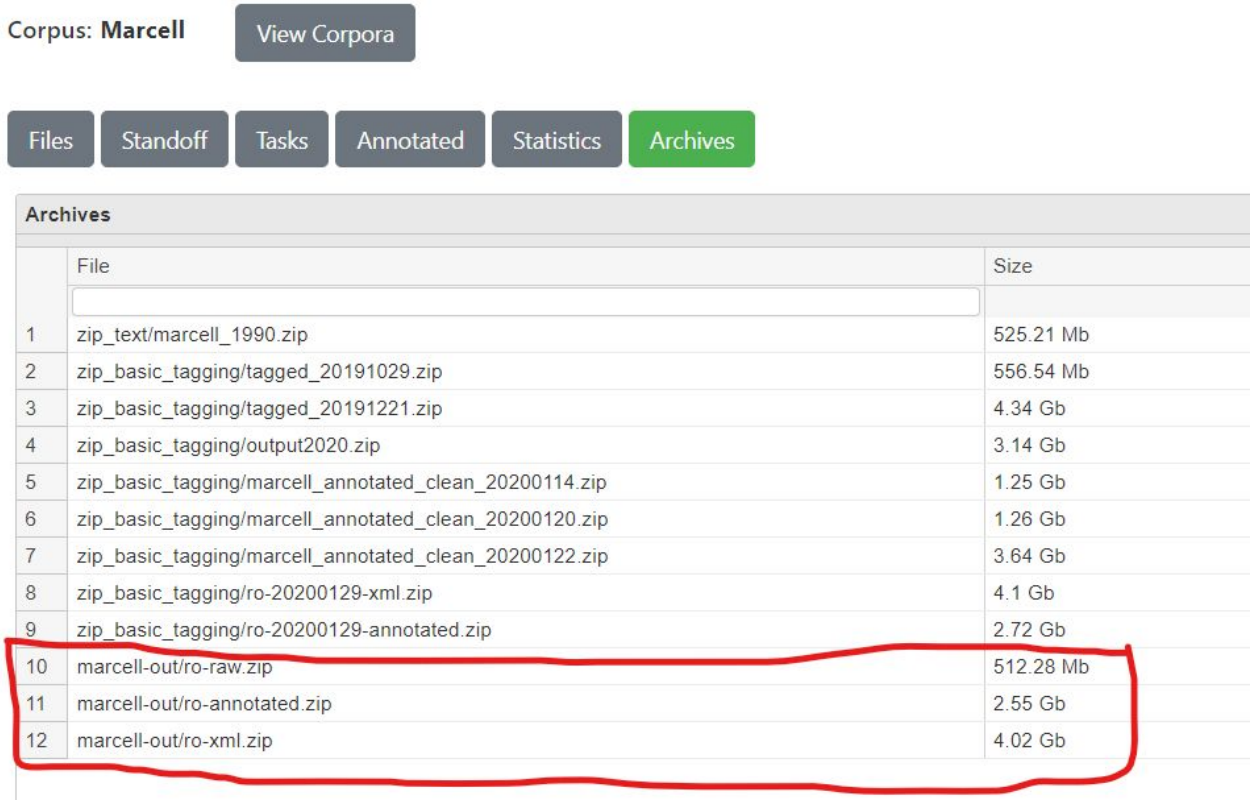
Archives

	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



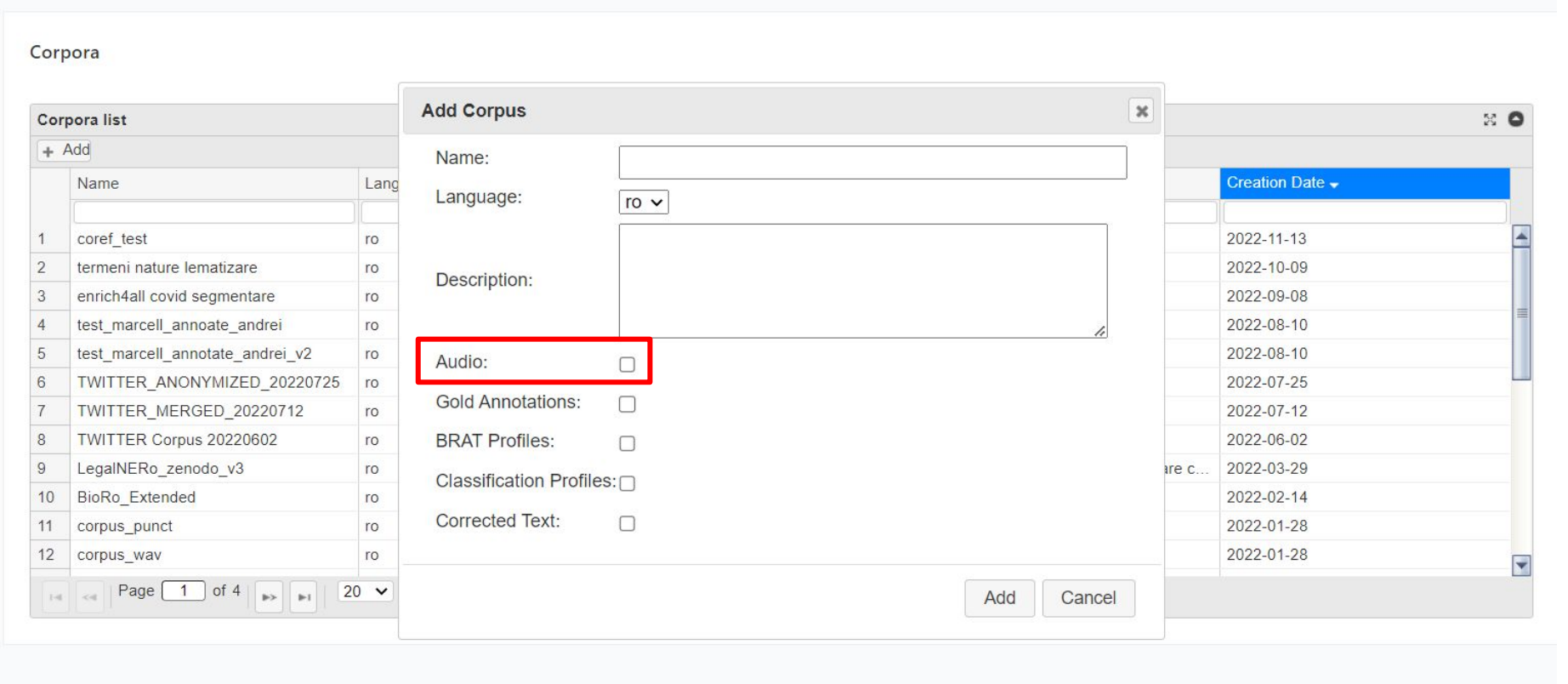
Import/Export



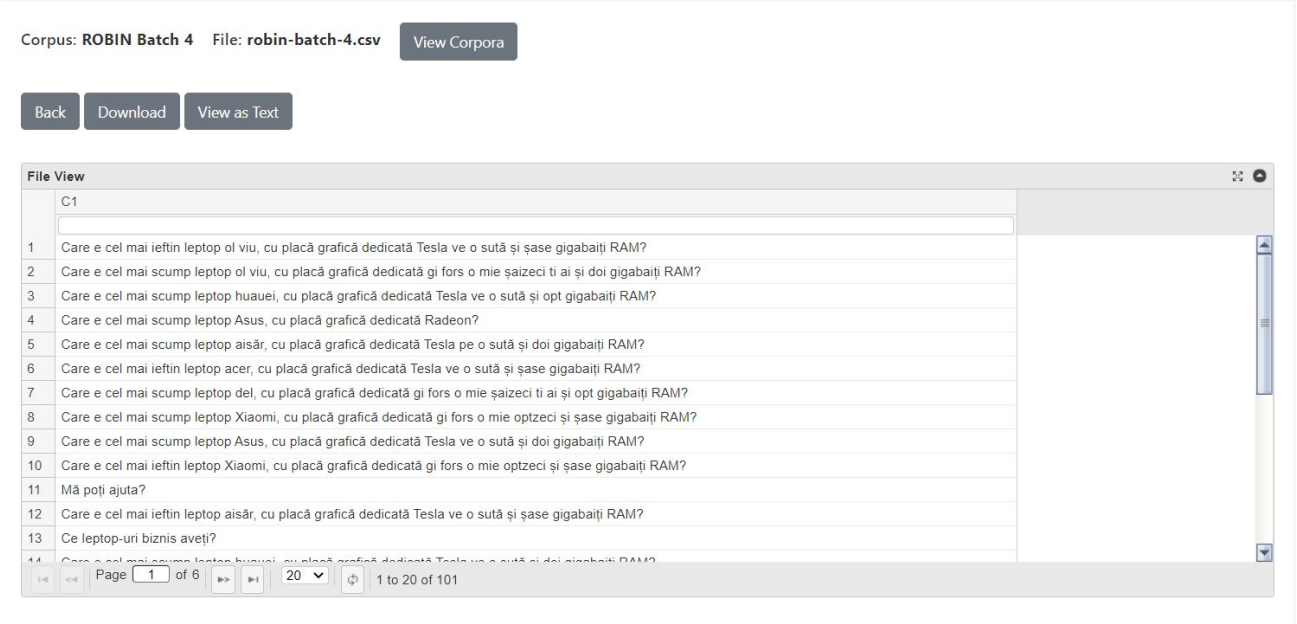
Multimodal Processing

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

Speech - Corpora creation



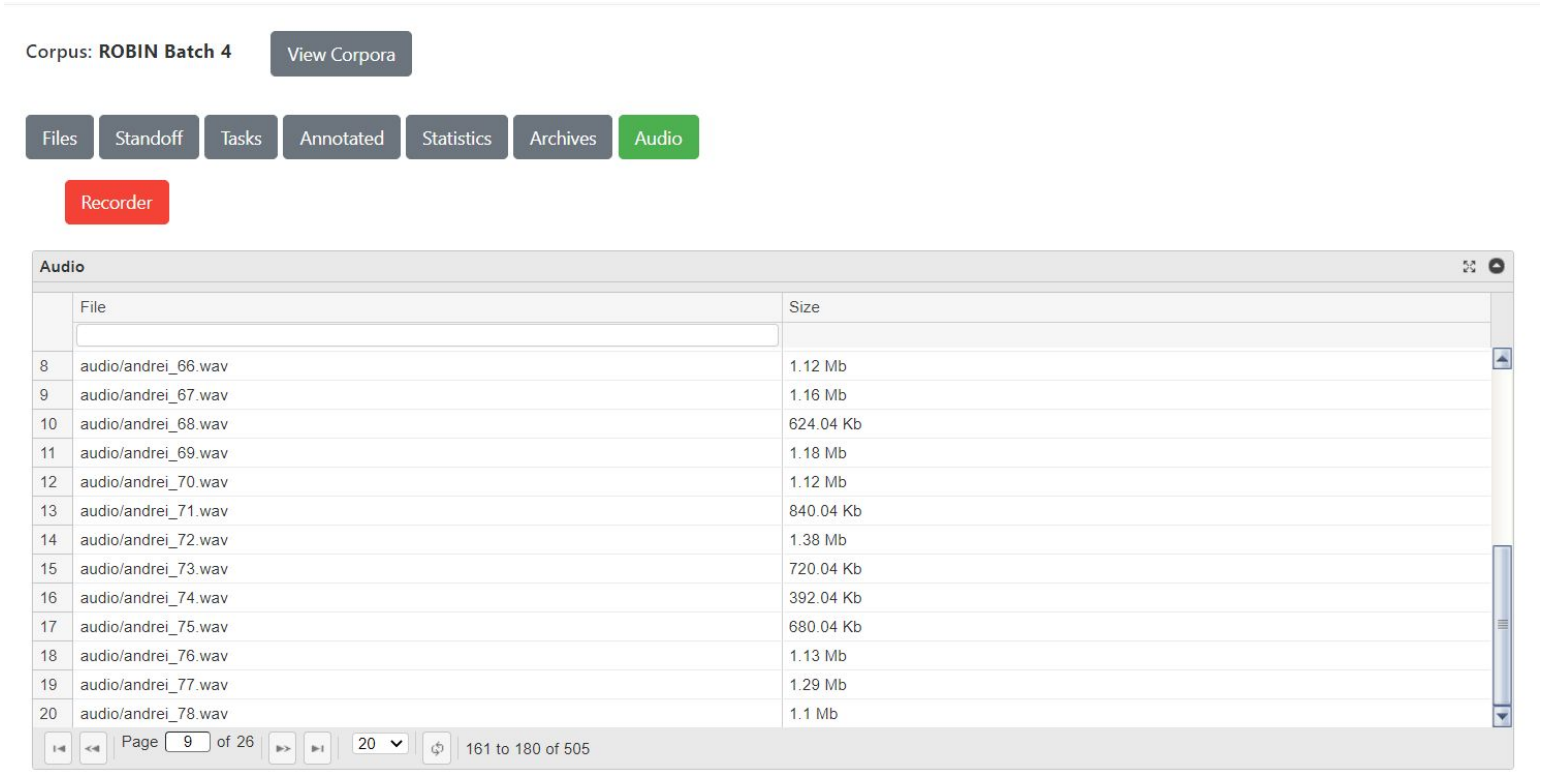
Speech - Corpora creation



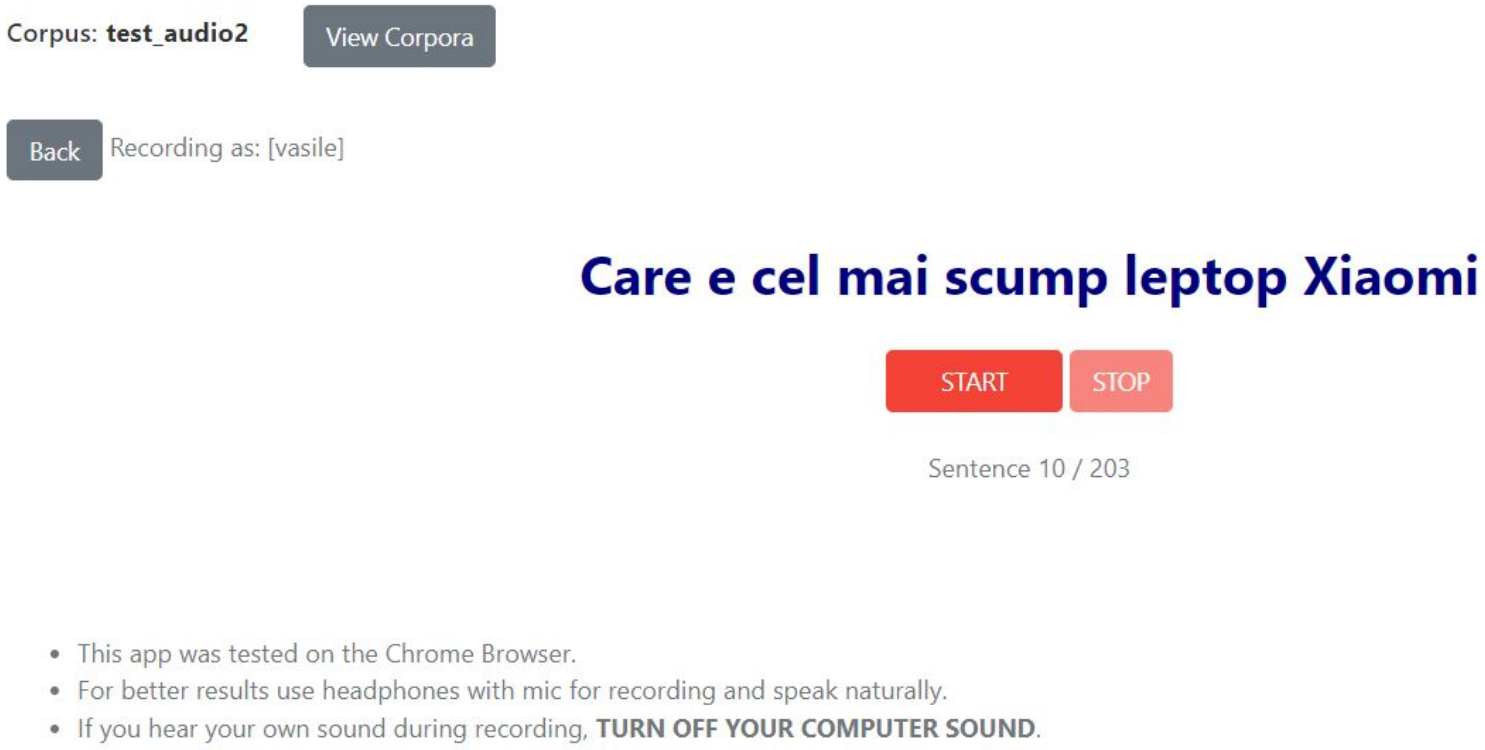
Păiș, Vasile and Ion, Radu and Barbu Mititelu, Verginica and Irimia, Elena and Mitrofan, Maria and Avram, Andrei (2021). ***ROBIN Technical Acquisition Speech Corpus***. Zenodo, https://doi.org/10.5281/zenodo.4626539

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



Speech - Corpora creation



Speech - Modules

- ASR - DeepSpeech2

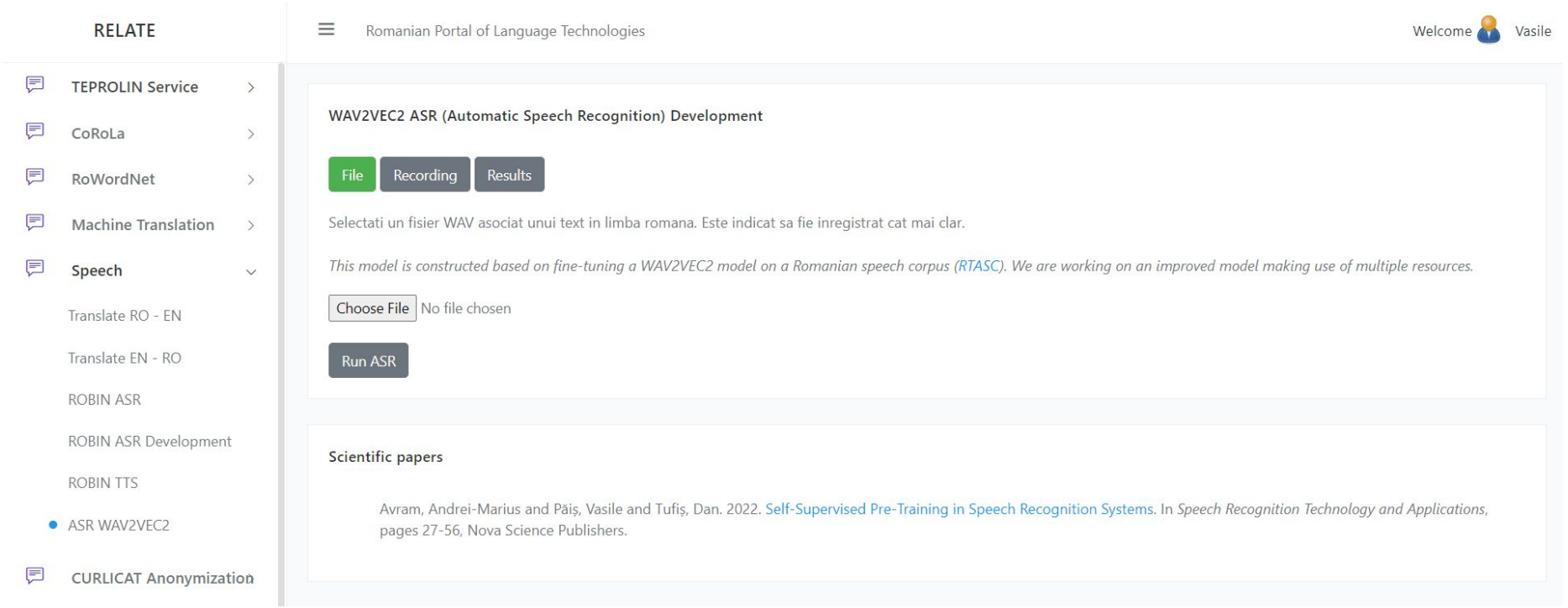
Avram, Andrei-Marius and Păiș, Vasile and Tufis, Dan (2020). ***Romanian speech recognition experiments from the ROBIN project***. In Proceedings CONSILR, pp. 103--114, https://profs.info.uaic.ro/~consilr/2021/wp-content/uploads/2021/03/volum-ConsILR-v-4-final-revizuit.pdf#page=111

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 Tufiș, 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

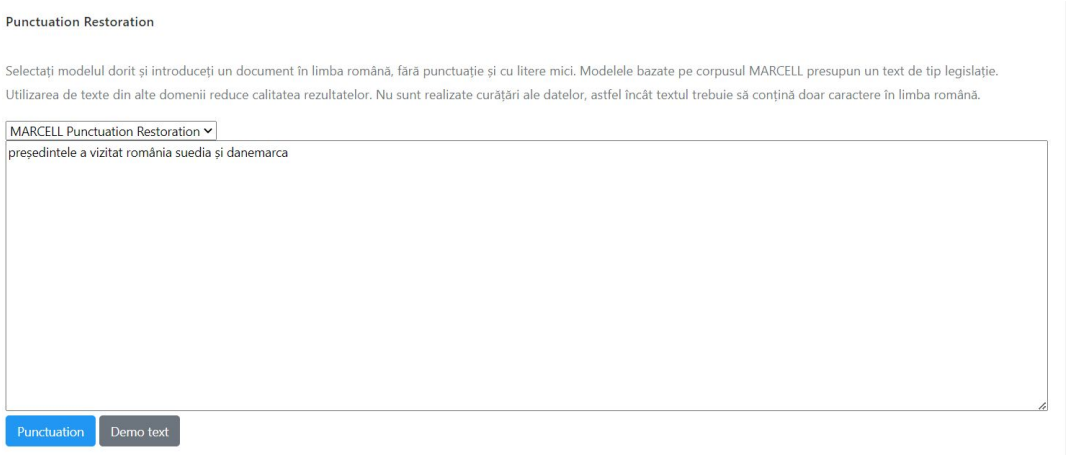


Speech - Modules

- Punctuation Restoration

Păiş, Vasile and Tufiş, 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



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.

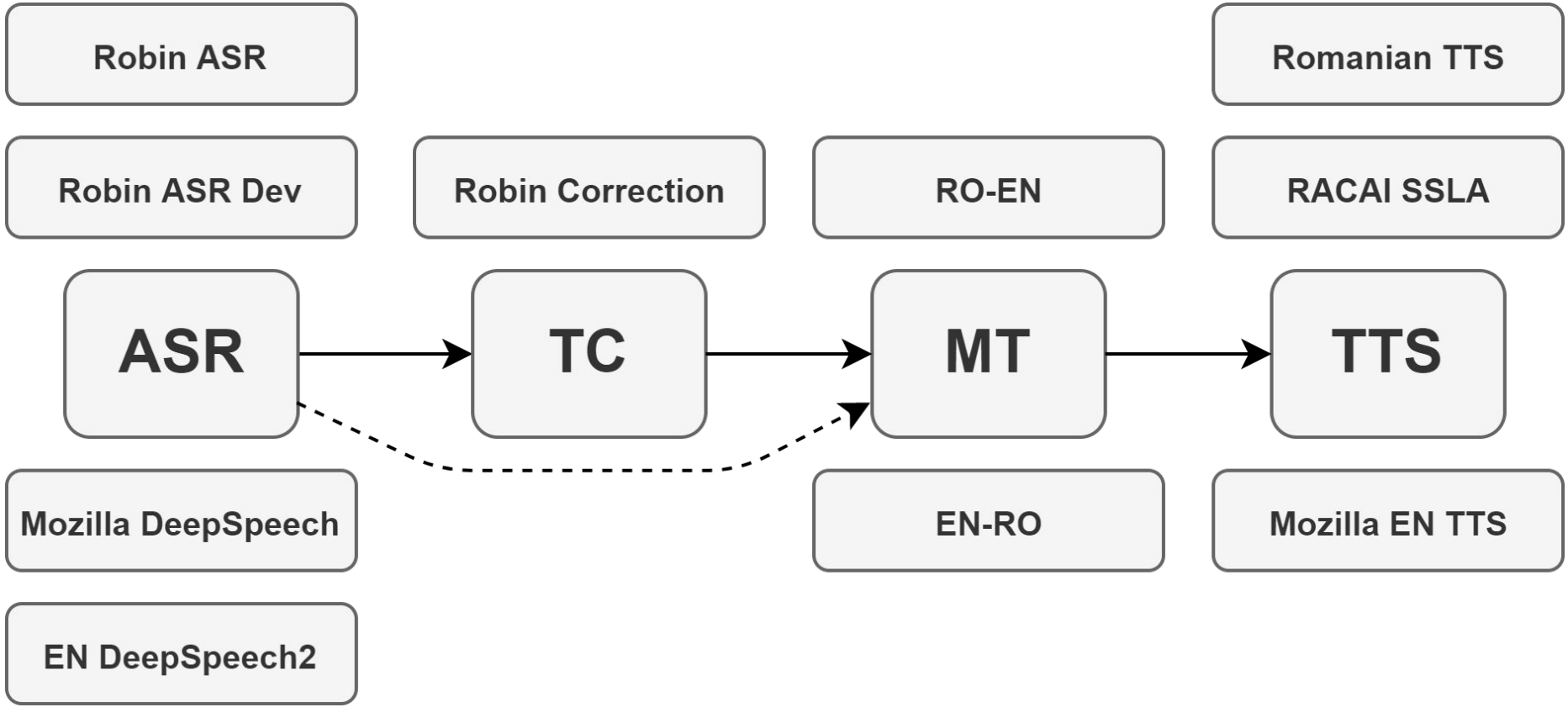
Boroș, Tiberiu and Ion, Radu and Dumitrescu, Ștefan Daniel (2013). ***The RACAI text-to-speech synthesis system***. In Blizzard Challenge Workshop 2013.

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

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.

- 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

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

Choose File | No file chosen

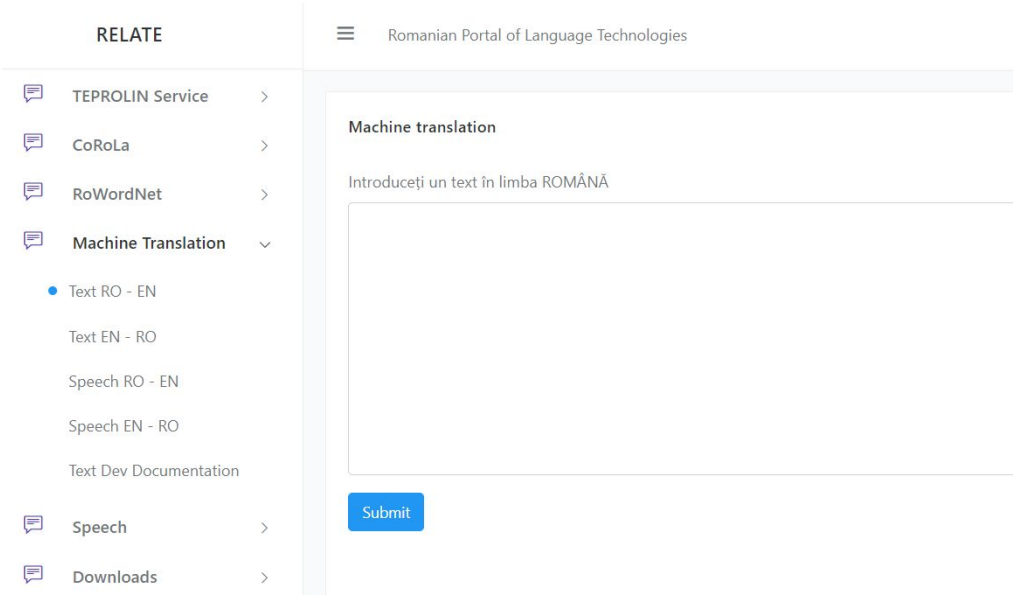
Lanț de prelucrare: ROBIN ASR | ROBIN Correction | RO Presidency | Mozilla EN TTS

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

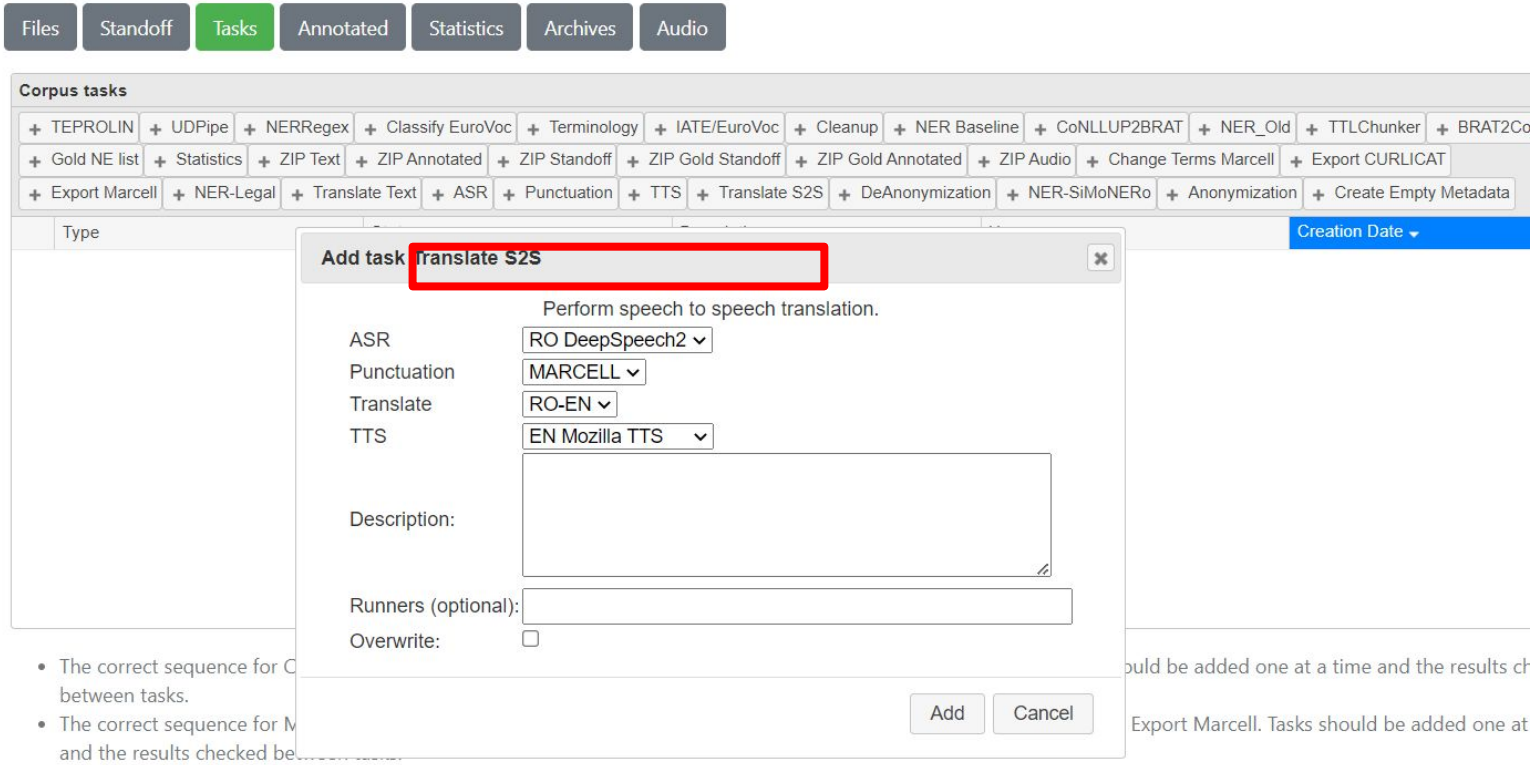


CEF Automated Translation toolkit for the Rotating Presidency of the Council of the EU

Developed by Tilde with support from ICIA during the Romanian presidency

The API is integrated in RELATE

Speech - Task-based processing



Păiș, Vasile and Mitrofan, Maria and Avram, Andrei-Marius. ***Romanian Language Translation in the RELATE Platform***. In Proceedings of LoResMT, pp. 64--74, https://aclanthology.org/2022.loresmt-1.9

Image processing

Specific interface elements activated for image datasets

Corpus: **NewsImages_BATCH1_Vasile** View Corpora

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

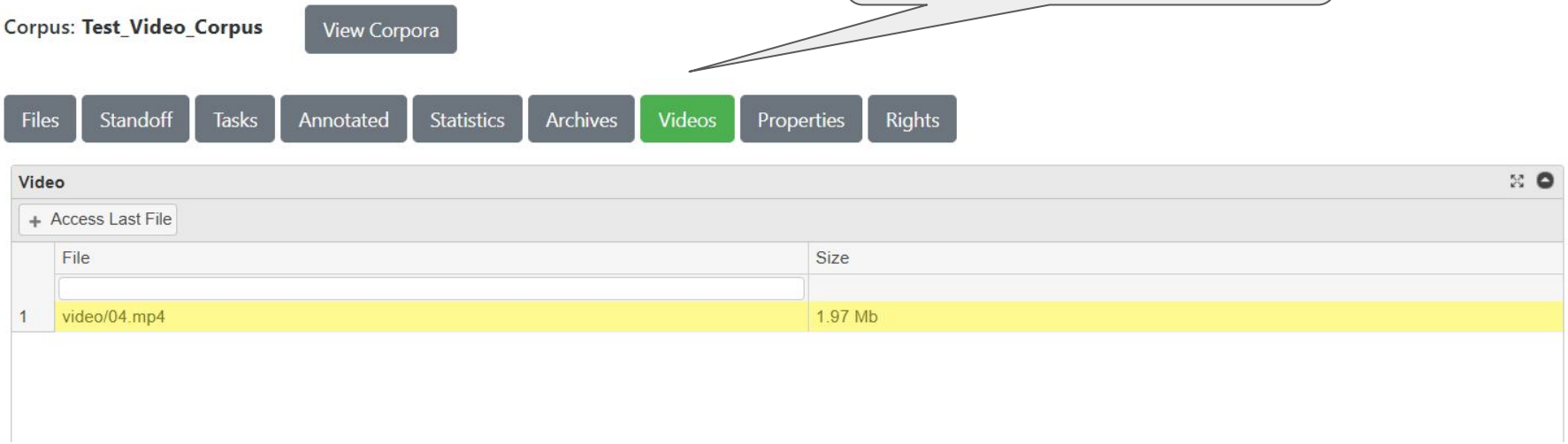
Corpus: **NewsImages_BATCH1_Vasile** File: **BATCH1_00005.jpg** View Corpora

Back Download PREVIOUS NEXT

Realistic Image Generation 5-Excellent Save

Video processing

Specific interface elements activated for video datasets



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

Page 1 of 31 20 1 to 20 of 601

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

Page 1 of 31 20 1 to 20 of 601

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

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

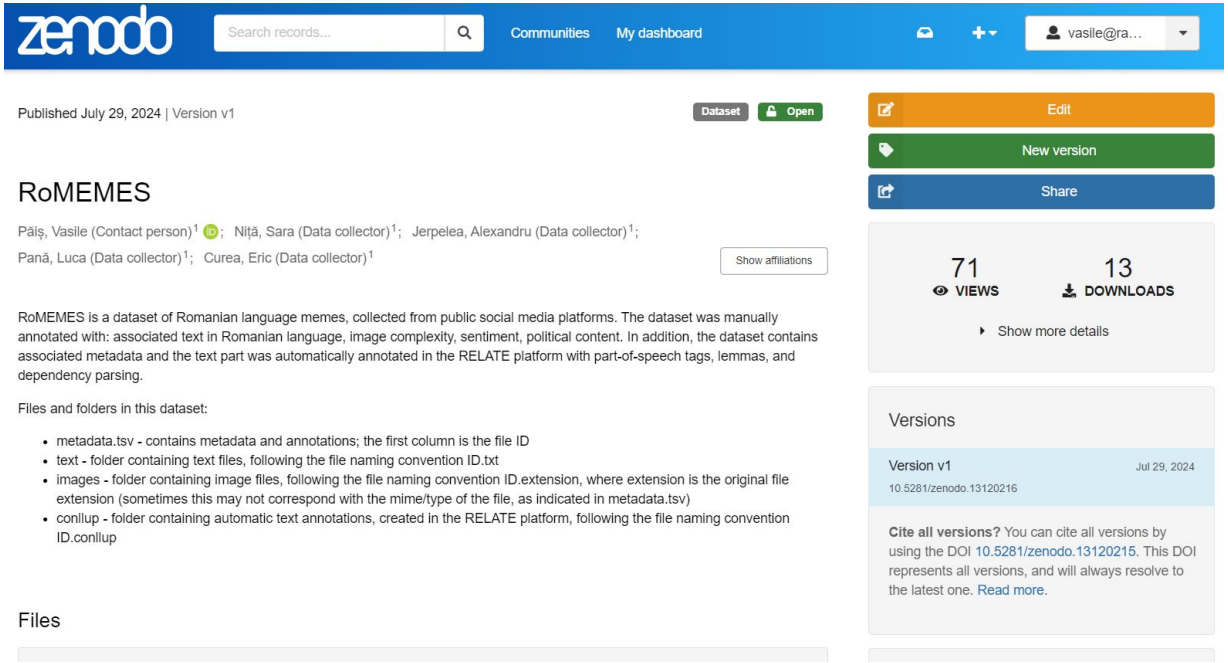
- usually funny
- tend to spread on social media
- image(s) + text
- may be targeted against individuals or groups

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



Paper presented at CONSILR 2024

RoMEMEs v2 - at SPED 2025

Cluj 19-22 October with LDS Workshop

Linked Data

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

- 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

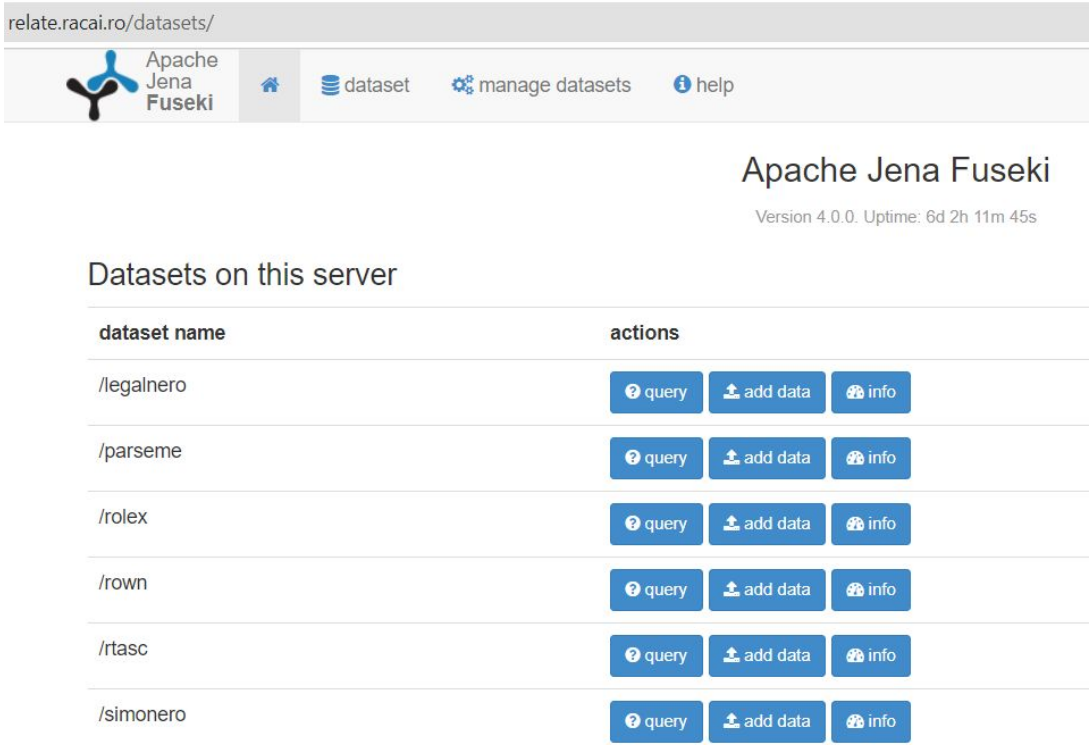
Back

```
@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



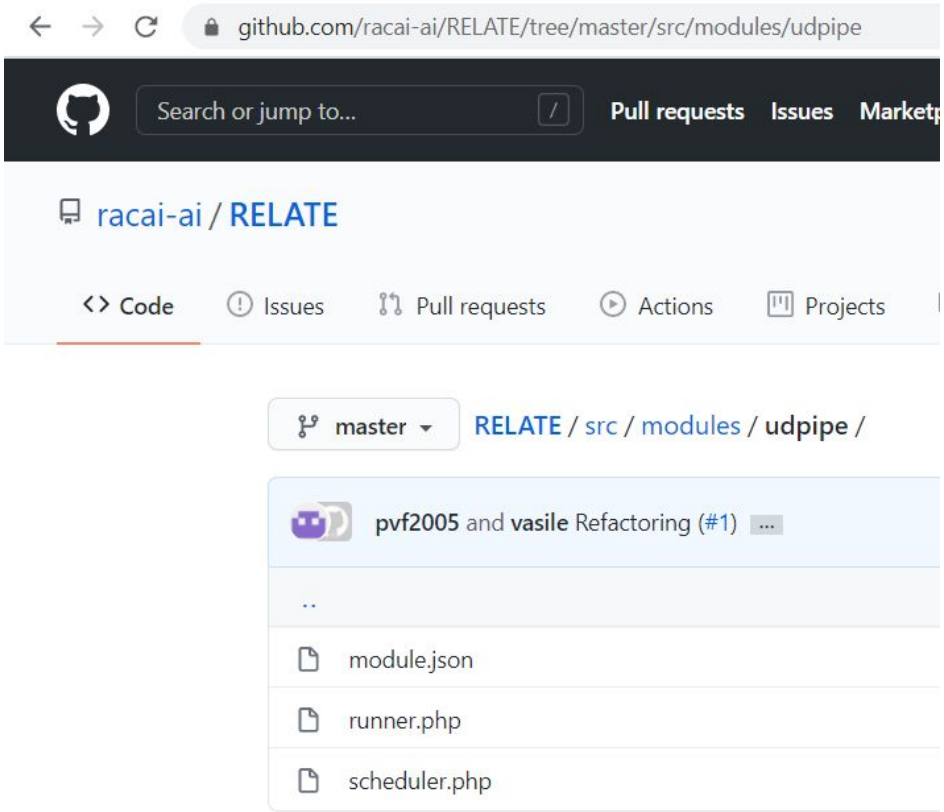
Păiș, Vasile and Barbu-Mititelu, Verginica (2022). ***Linguistic Linked Open Data for Speech Processing***. Chapter in Speech Recognition Technology and Applications (ed. Păiș, Vasile). Nova Science Publishers, pp. 155--188, https://doi.org/10.52305/BKWM899

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