European Language Resource Coordination
Connecting Europe Facility

European Commission

TEHNOLOGII ALE LIMBAJULUI PENTRU LIMBA ROMÂNĂ - PLATFORMA RELATE

Vasile Păiș
Institutul de Cercetare pentru Inteligență Artificială ”Mihai Drăgănescu”,
Academia Română
vasile@racai.ro

29.09.2021

Cuprins

European Language Resource Coordination
Connecting Europe Facility

European Commission

RELATE

O PLATFORMĂ PENTRU TEHNOLOGII ALE LIMBAJULUI ÎN LIMBA ROMÂNĂ

Include tehnologii dezvoltate la ICIA și de PARTENERI în multiple proiecte:

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

Adaptată filosofiei **ELG**:

- web services, REST API, docker
- serviciile pot fi distribuite pe noduri de rețea multiple
- serviciile pot fi consumate direct de la parteneri

European Language Resource Coordination
Connecting Europe Facility

European Commission

RELATE

ARTICOLE ȘTIINȚIFICE

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. 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. https://profs.info.uaic.ro/~consilr/wp-content/uploads/2021/03/volum-ConsILR-v-4-final-revizuit.pdf#page=73

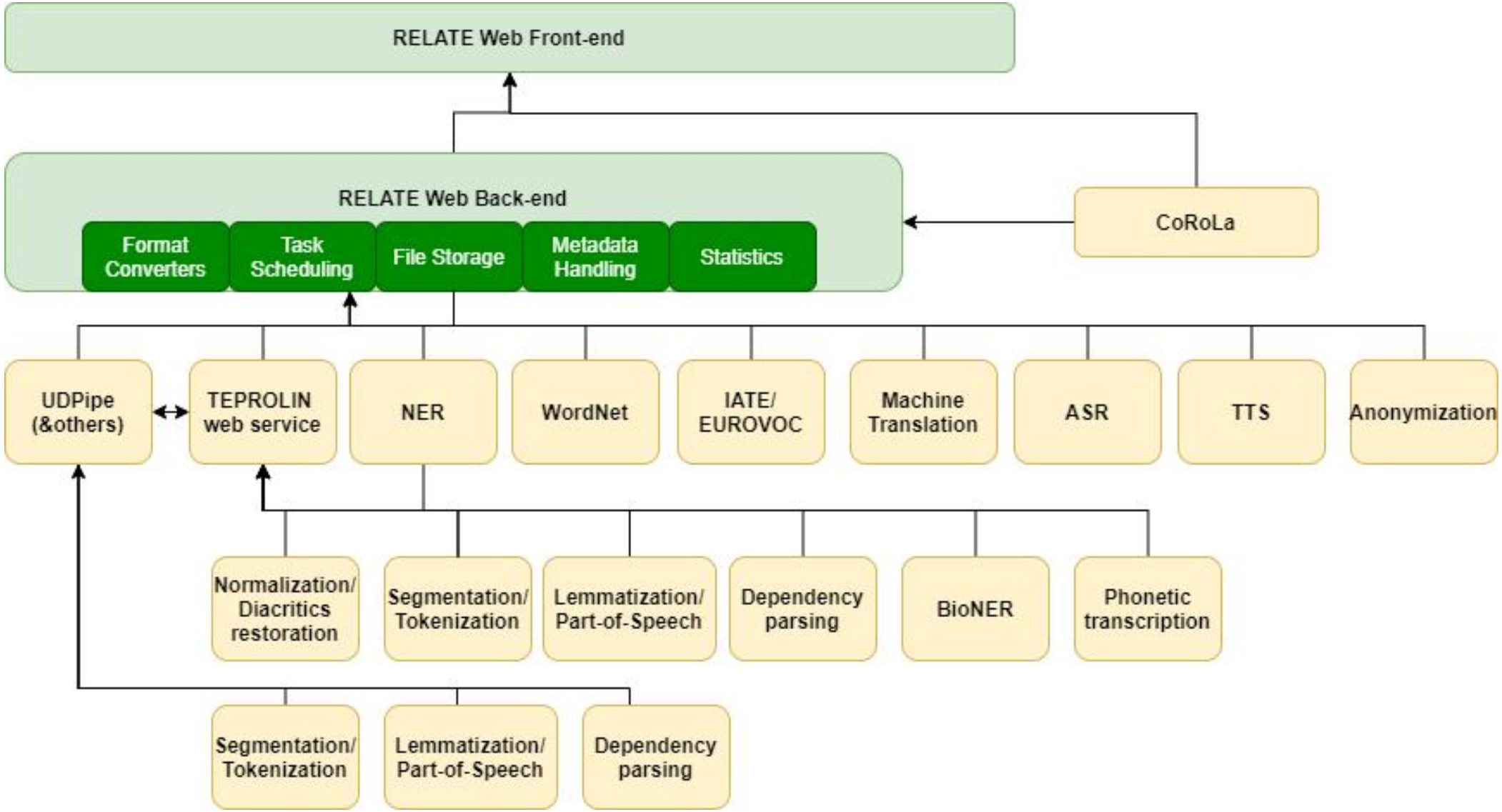
Vasile Păiș, Dan Tufiș, and Radu Ion. ***"Integration of Romanian NLP tools into the RELATE platform"***. In:International Conference on Linguistic Resources and Tools for Natural Language Processing. 2019, pp. 181–192. https://profs.info.uaic.ro/~consilr/2019/wp-content/uploads/2020/01/ConsILR2019_final_BTT-60-ex-B5.pdf#page=189

European Language Resource Coordination
Connecting Europe Facility

European Commission

RELATE

ARHITECTURA PLATFORMEI


RELATE Web Front-end
RELATE Web Back-end
Format Converters
Task Scheduling
File Storage
Metadata Handling
Statistics
CoRoLa
UDPipe (&others)
TEPROLIN web service
NER
WordNet
IATE/ EUROVOC
Machine Translation
ASR
TTS
Anonymization
Normalization/ Diacritics restoration
Segmentation/ Tokenization
Lemmatization/ Part-of-Speech
Dependency parsing
BioNER
Phonetic transcription
Segmentation/ Tokenization
Lemmatization/ Part-of-Speech
Dependency parsing

European Language Resource Coordination
Connecting Europe Facility

European Commission

RELATE

CARACTERISTICI PENTRU VOLUME MARI DE DATE

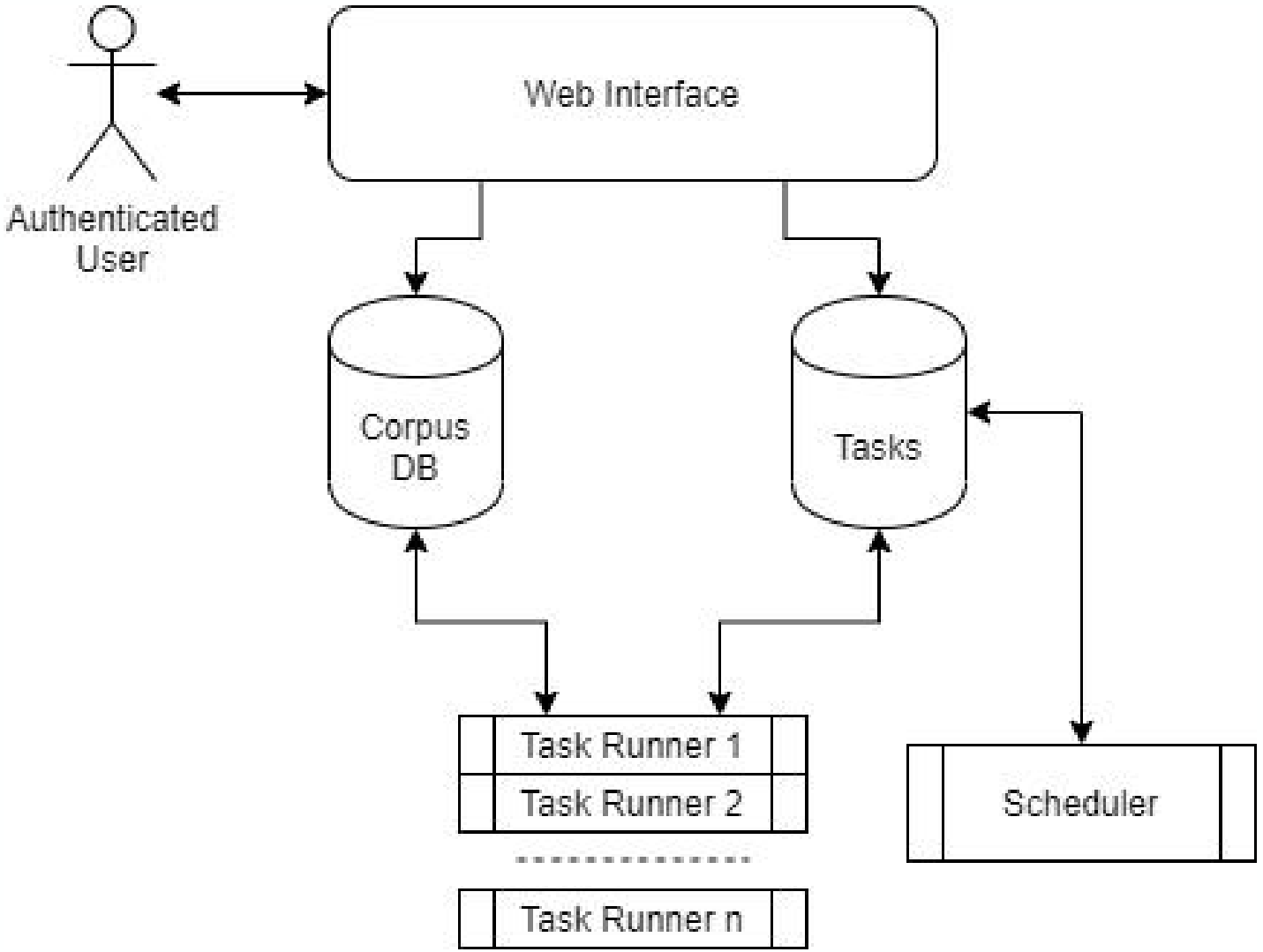
Stocare și Procesare pentru corpora de dimensiuni mari

- text
- text + audio

Procesare paralelă:

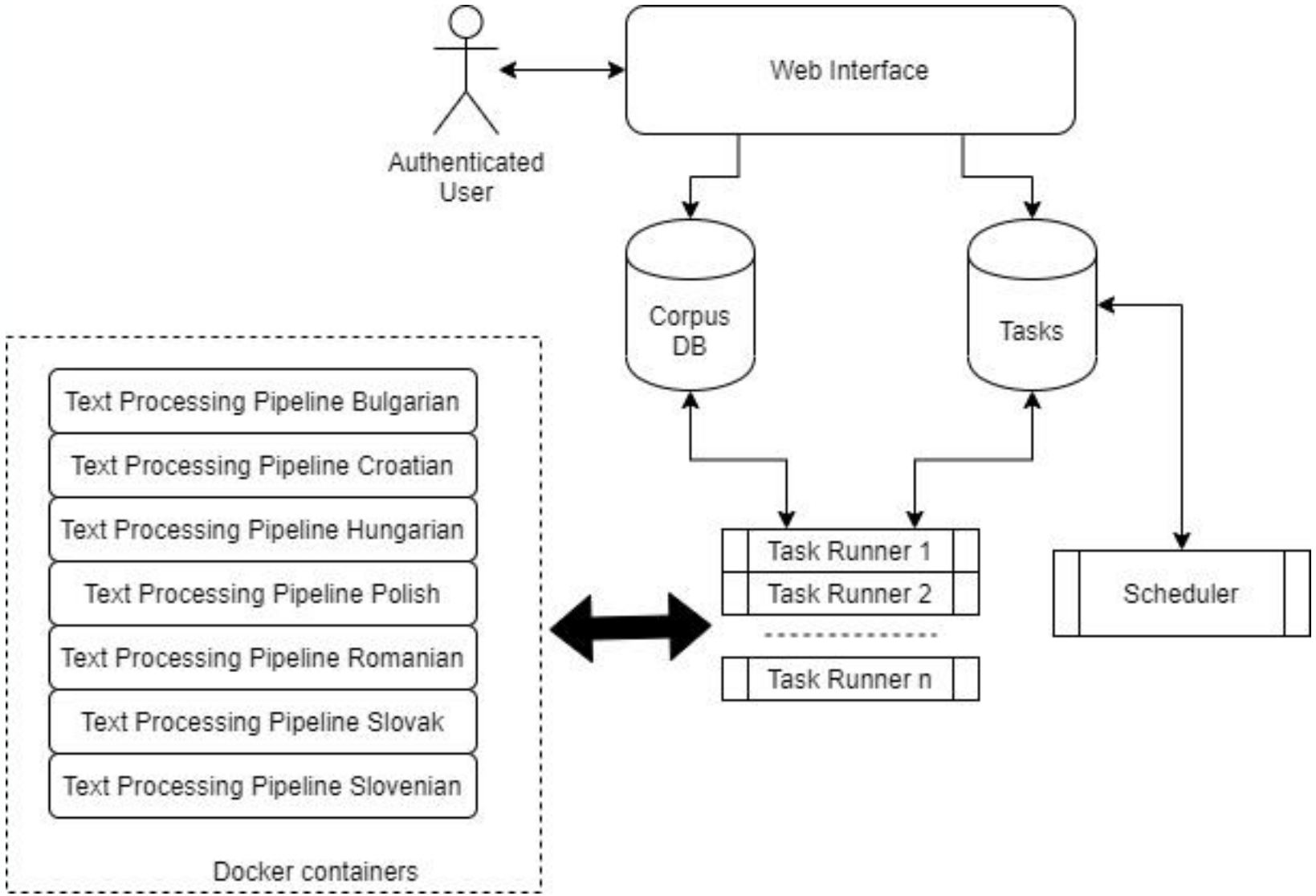
- sistem de task-uri
- consum de servicii web externe platformei
    - noduri de rețea diferite
    - direct de la parteneri

TASK MANAGEMENT



ADAPTABILITATE LA NOI LANȚURI DE PRELUCRARE

MARCELL SUSTAINABILITY PLATFORM



European Language Resource Coordination
Connecting Europe Facility

European Commission

RELATE

CARACTERISTICI GENERALE

CORPUS MANAGEMENT: CREARE, ÎNCĂRCARE, DESCĂRCARE, ARHIVARE, ADNOTARE, STATISTICI, VIZUALIZARE

CREARE CORPUS "GOLD": INTEGREAZĂ BRAT PENTRU NER, ÎNREGISTRATOR PENTRU CORPUS VOCE-TEXT

ADNOTARE: LEMMA, PART-OF-SPEECH, DEPENDENCY PARSING, SYLLABIFICATION, DIACRITICS RESTORATION, NER, BIONER, IATE, EUROVOC

TRADUCERE: TEXT-TEXT, VOCE-VOCE

VOCE: ASR, TTS, TRADUCERE

WORDNET: ROWORDNET, INTEROGARE ALINIATĂ ENWORDNET

COROLA: INTEGRARE KORAP, COMPONENTA AUDIO, WORD EMBEDDINGS

ANONIMIZARE: APLICAȚIE DIN PROIECTUL CURLICAT (ÎN DESFĂȘURARE)

LEGAL NER: PERSOANE, ORGANIZAȚII, LOCAȚII, TIMP, REFERINȚE LEGISLATIVE
- PE BAZA UNUI SUB-CORPUS EXTRAS DIN MARCELL

MODELE DE LIMBĂ: POT FI DESCĂRCATE DIN PLATFORMĂ

FORMATE STANDARD: CONLL-U, CONLL-U PLUS, XML, JSON, RDF

ARHITECTURĂ MODULARĂ: PLATFORMA SE EXTINDE CU FIECARE NOU PROIECT

European Language Resource Coordination
Connecting Europe Facility

European Commission

RELATE

VIZUALIZĂRI

VIZUALIZARE ADAPTATĂ CONTEXTULUI

- **TABEL** - CORPUS MANAGEMENT
- **ARBORE** - GRAF DEPENDEȚE
- **TEXT** - DOCUMENTE, ALTERNATIVĂ LA TABEL
- **ADNOTĂRI FRAGMENTE TEXT** - NER
- **BUTOANE, FORMULARE** - TASK-URI
- **PLAYER AUDIO** - FIȘIERE AUDIO
- **RECORDER AUDIO** - ÎNREGISTRĂRI, TRADUCERE VOCE

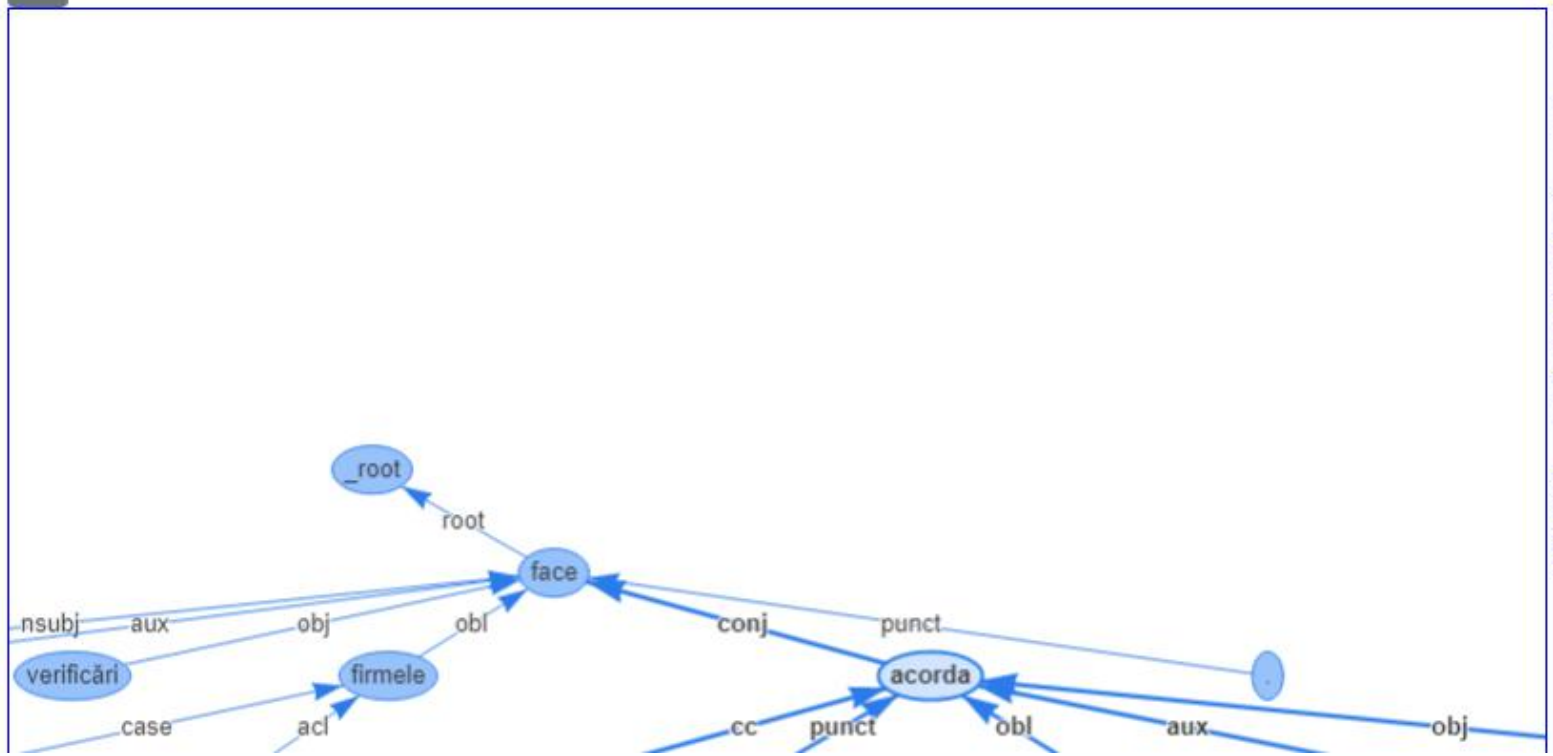
Complete run

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

Complete run

JSON | CoNLL-U | CoNLL-X | XML | Text | Chunks | Tree | Entities

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

2 Diabetul (DISO) zaharat este un sindrom (DISO) caracterizat prin valori crescute ale concentrației glucozei (CHEM) în sânge (ANAT) (hiperglicemie (DISO)) și dezechilibrarea metabolismului .

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

Corpus: **legalnero** View Corpora

Files | Standoff | Tasks | Annotated | Statistics | Archives

Corpus tasks

+ TEPROLIN + UDPipe + IATE/EuroVoc + Classify EuroVoc + Cleanup + NER Baseline + CoNLLUP2BRAT + BRAT2CoNLLUP + Gold NE list + Statistics + ZIP Text + ZIP Annotated + ZIP Standoff
+ ZIP Gold Standoff + ZIP Gold Annotated + ZIP Audio + Export Marcell + Change Terms Marcell + Anonymization

	Type	Status	Description	User	Creation Date
1	statistics	DONE		maria@racai.ro	2021-05-24 19:14:35
2	unzip_text	DONE	Unzip TEXT from legalnero.zip	maria@racai.ro	2021-05-24 19:09:59

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

European Language Resource Coordination
Connecting Europe Facility

European Commission

RELATE

CORPORA DEZVOLTATE CU AJUTORUL PLATFORMEI

MARCELL-RO

Dan Tufiș, Maria Mitrofan, Vasile Păiș, Radu Ion, and Andrei Coman. ***"Collection and Annotation of the Romanian Legal Corpus"***. In:Proceedings of The 12th Language Resources and Evaluation Conference. Marseille, France: European Language Resources Association, 2020, pp. 2773–2777. URL:https://www.aclweb.org/anthology/2020.lrec-1.337

ADNOTARE

STATISTICI

EXPORT ÎN ARHIVE CU FORMAT DIFERIT (TEXT, CONLLU PLUS AND XML)

LEGALNERO

PĂIȘ, VASILE, MITROFAN, MARIA, GASAN, CAROL LUCA, IANOV, ALEXANDRU, GHIȚĂ, CORVIN, CONESCHI, VLAD SILVIU, & ONUȚ, ANDREI. (2021). ***ROMANIAN NAMED ENTITY RECOGNITION IN THE LEGAL DOMAIN (LEGALNERO)*** [DATA SET]. ZENODO. HTTP://DOI.ORG/10.5281/ZENODO.4772094

ADNOTARE MANUALĂ NER

ADNOTARE AUTOMATĂ

STATISTICI

FORMAT RDF

LEGARE ENTITĂȚI GEONAMES

ROBIN TECHNICAL ACQUISITION SPEECH CORPUS

PĂIȘ, VASILE, ION, RADU, BARBU MITITELU, VERGINICA, IRIMIA, ELENA, MITROFAN, MARIA, & AVRAM, ANDREI. (2021). ***ROBIN TECHNICAL ACQUISITION SPEECH CORPUS*** [DATA SET]. ZENODO.
HTTP://DOI.ORG/10.5281/ZENODO.4626539

ÎNREGISTRARE

ADNOTARE AUTOMATĂ

STATISTICI

FOLOSIT PENTRU **ADAPTAREA ASR**-ULUI LA CERINȚELE PROIECTULUI

European Language Resource Coordination
Connecting Europe Facility

European Commission

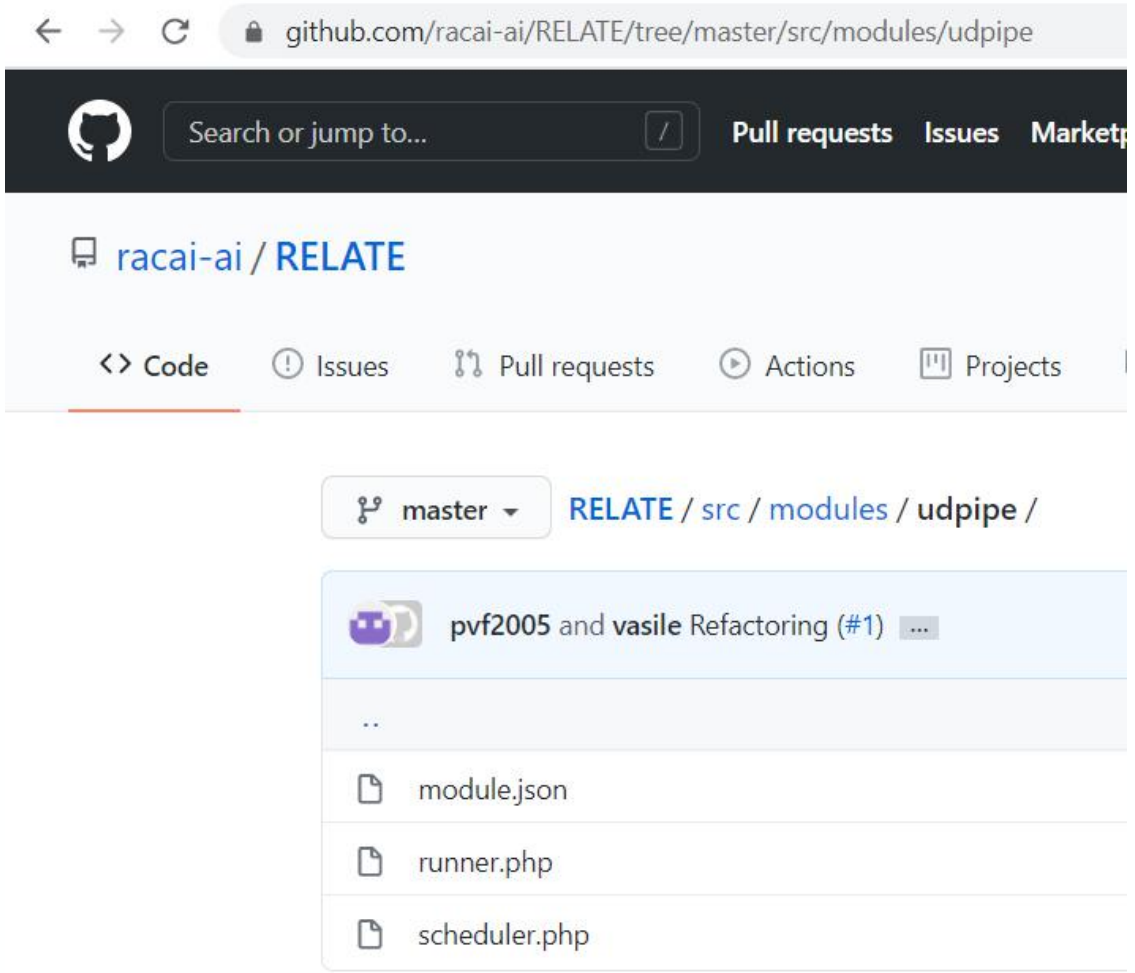
RELATE

DEZVOLTARE

Open Source

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

Bazat pe Componente



European Language Resource Coordination
Connecting Europe Facility

European Commission

READY? LET'S GIVE IT A TRY!

HTTPS://RELATE.RACAI.RO

European Language Resource Coordination
Connecting Europe Facility

European Commission

MULȚUMESC PENTRU ATENȚIE!

https://relate.racai.ro