INTERNATIONAL JOURNAL OF COMPUTERS COMMUNICATIONS & CONTROL
Online ISSN 1841-9844, ISSN-L 1841-9836, Volume: 17, Issue: 1, Month: 2, Year: 2022
Article Number: 4641, https://doi.org/10.15837/ijccc.2022.1.4641

# Romanian Language Technology – a view from an academic perspective

Dan Tufiş

**Dan Tufiş**
Research Institute for Artificial Intelligence "Mihai Drăgănescu"
Romanian Academy
tufis@racai.ro

**Abstract**

The article reports on research and developments pursued by the Research Institute for Artificial Intelligence „Mihai Drăgănescu" of the Romanian Academy in order to narrow the gaps identified by the deep analysis on the European languages made by Meta-Net white papers and published by Springer in 2012. Except English, all the European languages needed significant research and development in order to reach an adequate technological level, in line with the expectations and requirements of the knowledge society.

**Keywords:** Romanian language resources and technologies, Meta-Net White papers, European Language Equality

## 1 Introduction

The technological evolution of the last decades, both in the field of hardware and communications and in the field of algorithms, representation models and automated processing flows has highlighted the exceptional value of data. The success of neural, sequence-by-sequence or transformer models, as well as the proven superiority of applications based on such models and machine learning, have placed the volume and quality of data at the forefront of the technological and scientific advancement of the information society, of the knowledge society. The awareness campaigns started by the European Commission (EC) on language technologies in the early 2000's continued with large projects (ELSNET[1] - http://www.elsnet.org, CLARIN- https://www.clarin.eu, META-NET- http://www.meta-net.eu/, and more recently ELG - https://www.european-language-grid.eu/, ELRC - https://www.lr-coordination.eu/index.php/, ELE - https://european-language-equality.eu/) involving all countries from European Union. The effectiveness of these actions, associated to more research and development projects funded by EC and the national authorities is apparent as the technological levels for the targeted languages are significantly improved, although many efforts are still needed to reach all the objectives set out in the European Language Equality manifesto.

[1] ELSNET (European Language and Speech Network) defined the BLARK concept - Basic Language and Resource Kit – as the minimal set of language technologies for judging the languages technological preparedness.

## 2 The Meta-Net report on European Languages

One of the most influential European initiatives in the area of Language Technologies (LT) was META-NET (http://www.meta-net.eu/mission), a Network of Excellence forging the Multilingual Europe Technology Alliance. Among other significant results, META-NET project made an impressive analysis of the state-of-play for European languages (White Papers) concerning their technological levels, published in 2012 by Springer. The language reports on 30 European languages (http://www.meta-net.eu/whitepapers/overview), Romanian included, outlined that there were tremendous deficits in technology support and significant research gaps for each language. The comparison among the languages was carried on along four domains, taking into account multiple criteria (availability, accessibility, quality, coverage, maturity, sustainability, adaptability and multilinguality):

- Speech processing: state of language technology support,
- Machine translation: state of language technology support,
- Text analysis: state of language technology support and
- Speech and text resources: state of support.

Among the main conclusions, the report highlighted the major gaps for each language:

- While some specific corpora of high quality exist, a very large syntactically annotated corpus is not available.
- Many of the resources lack standardization, i.e., even if they exist, sustainability is not given; concerted programs and initiatives are needed to standardize data and interchange formats.
- The more semantics a tool must deal with, the more difficult is to find the right data; more efforts for supporting deep processing are needed.
- Standards do exist for semantics in the sense of world knowledge (RDF, OWL, etc.); they are – however – not easily applicable in NLP tasks.
- Research was successful in designing particular high-quality software, but it is nearly impossible to come up with sustainable and standardized solutions given the current funding situations.
- For certain languages, certain technologies simply do not exist.

The report on Romanian (http://www.meta-net.eu/whitepapers/volumes/romanian) was the guide document for organizing and prioritizing research and development at ICIA and major LT centers in Romania (UAIC, IIT, UPB and others). For Romanian, the report showed the main shortcomings, many of them common for majority of the analyzed languages:

- Lack of large and high-quality corpora, deeply annotated (reference corpus)
- Limited open access to major resources and tools for processing Romanian Language
- Compliance with data standards and interchange formats
- Limited semantic and standardized language resources
- Few resources and tools for processing Romanian speech data
- Enhancing the resources and engines for Machine Translation
- Observing the latest trends in LT research and development.

In the period elapsed from the time of the META-NET reports, improvements were made for all languages. In the following, we will briefly comment on the major activities carried on by ICIA towards eliminating or narrowing the gaps revealed by the META-NET report.

# 3 The language resources program at ICIA

In line with the priorities of the European Commission's Artificial Intelligence programs focused on Language Technologies, ICIA has conducted advanced research and developed corpora, semantic resources, language models and applications for the Romanian language, and made them available to the repositories of the European Commission built through the European Language Grid, European Language Coordination, European Language Equality and Meta-Net projects. ICIA became the most important contributor in Romania with resources and technologies specific to the Romanian language to these repositories.

## 3.1 The lexical ontology for Romanian language

The most significant semantic resource, developed since 2002 and being continuously maintained is the Romanian WordNet (Tufiș et al., 2004), a lexical ontology containing 60,000 synsets. The lexical ontology RoWordNet was constructed and validated manually. This lexical resource, one of the largest in the world, is aligned with Princeton WordNet (Miller, 1995; Fellbaum, 1998) as majority of other language specific wordnets are. The alignment with Princeton WordNet makes the navigation possible between different languages, allowing for multilingual semantic experiments and applications (e.g., word sense disambiguation). RoWordNet is available on multiple sites out of which we mention:

- https://www.racai.ro/tools/text/rowordnet/
- https://github.com/dumitrescustefan/RoWordNet
- http://globalwordnet.org/resources/wordnets-in-the-world/
- http://compling.hss.ntu.edu.sg/omw/

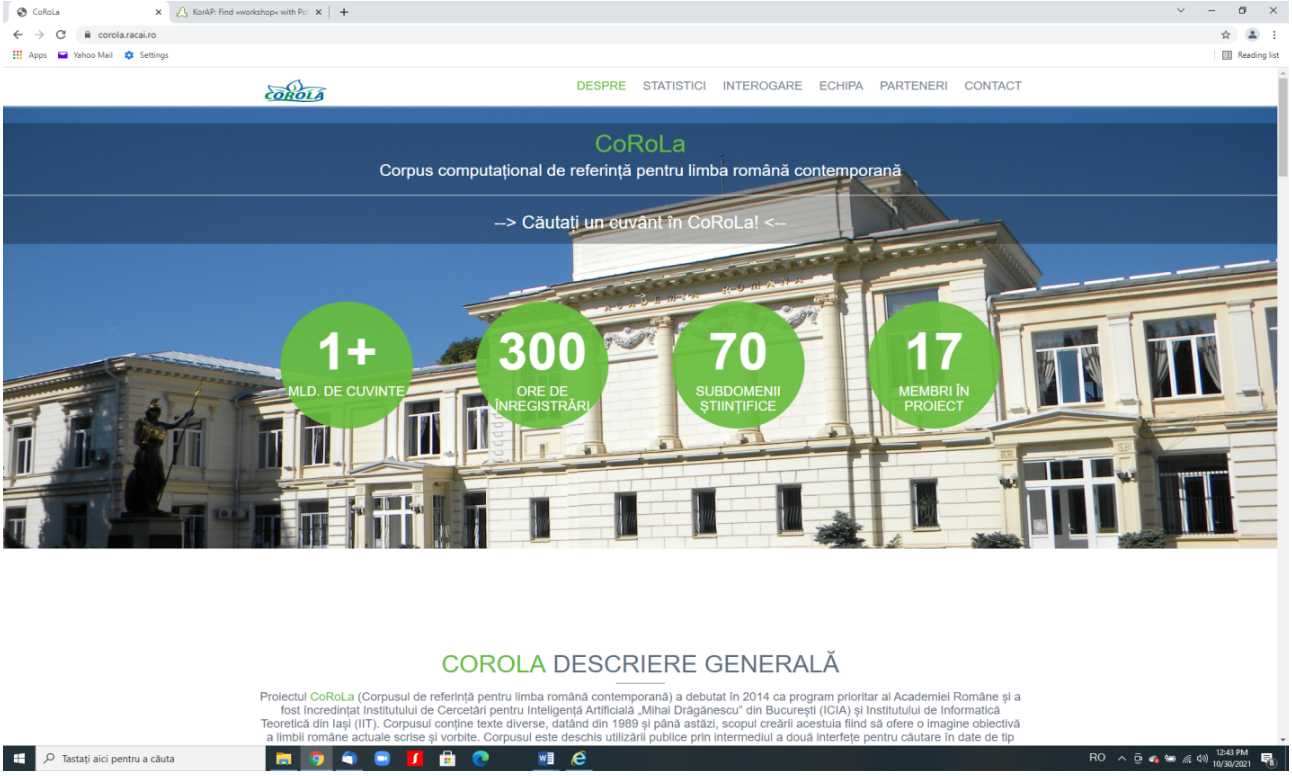
A parallel query of both Romanian WordNet and Princeton WordNet is available from the RELATE (Păiș et al., 2020) platform (https://relate.racai.ro/index.php?path=rown/queryp). RoWordNet was the main source for extracting verbal multiword expressions (VMWE) and annotating them in running texts, according to PARSEME annotation manual (https://parsemefr.lis-lab.fr/parseme-st-guidelines/1.1/index.php). After the hand validation of the first version of the list of verbal multiword expressions, the RoWN-VMWEs-v.2 has been publicly released (https://www.racai.ro/media/ro-data-mwe-v2.txt) together with the PARSEME_corpus.ro (https://gitlab.com/parseme/parseme_corpus_ro).

## 3.2 Multilingual and monolingual textual corpora

The most used parallel corpus, JRC-Acquis, in its version 2.2[2] (Steinberger et al., 2006) was released in 2006 and contained a part of the body of European Union (EU) laws applicable in the EU Member States. The documents included in the corpus were only those translated in at least 10 languages. The Acquis Communautaire is a collection of parallel texts in 22 languages: Bulgarian, Czech, Danish, German, Greek, English, Spanish, Estonian, Finnish, French, Hungarian, Italian, Lithuanian, Latvian, Maltese, Dutch, Polish, Portuguese, Romanian, Slovak, Slovenian and Swedish. ICIA was responsible for processing the Romanian and, partially, Bulgarian sub-corpora of the JRC-Acquis corpus. This parallel corpus was the major source of training machine translation systems all over the world. Since parallel texts for so many languages are difficult to find, during ACCURAT project, the ICIA team developed the innovative systems EMACC (Ion et al., 2011) and PEXACC (Ion, 2012) to extract parallel text fragments from comparable texts.

Research and development on parallel and comparable corpora gave a global impetus to development of statistical machine translation systems. In ICIA, we developed a state-of-the-art, Romanian to English statistical machine translation system (Tufiș and Dumitrescu, 2012), based on the Moses toolkit (Koehn, et al., 2007). The machine translation preoccupations continued, and, within the European project Presidency, a better and professional translation system (named MT Kit) was developed with the close cooperation with TILDE company (https://www.tilde.com/) and DGT-UE.

[2] Today, JRC-Acquis corpus is available in the version 3.0



CoRoLa

Corpus computațional de referință pentru limba română contemporană

--> Căutați un cuvânt în CoRoLa! <--

COROLA DESCRIERE GENERALĂ

Proiectul CoRoLa (Corpusul de referință pentru limba română contemporană) a debutat în 2014 ca program prioritar al Academiei Române și a fost încredințat Institutului de Cercetări pentru Inteligență Artificială „Mihai Drăgănescu" din București (ICIA) și Institutului de Informatică Teoretică din Iași (IIT). Corpusul conține texte diverse, datând din 1989 și până astăzi, scopul creării acestuia fiind să ofere o imagine obiectivă a limbii române actuale scrise și vorbite. Corpusul este deschis utilizării publice prin intermediul a două interfețe pentru căutare în date de tip

Figure 1: The welcome page of CoRoLa

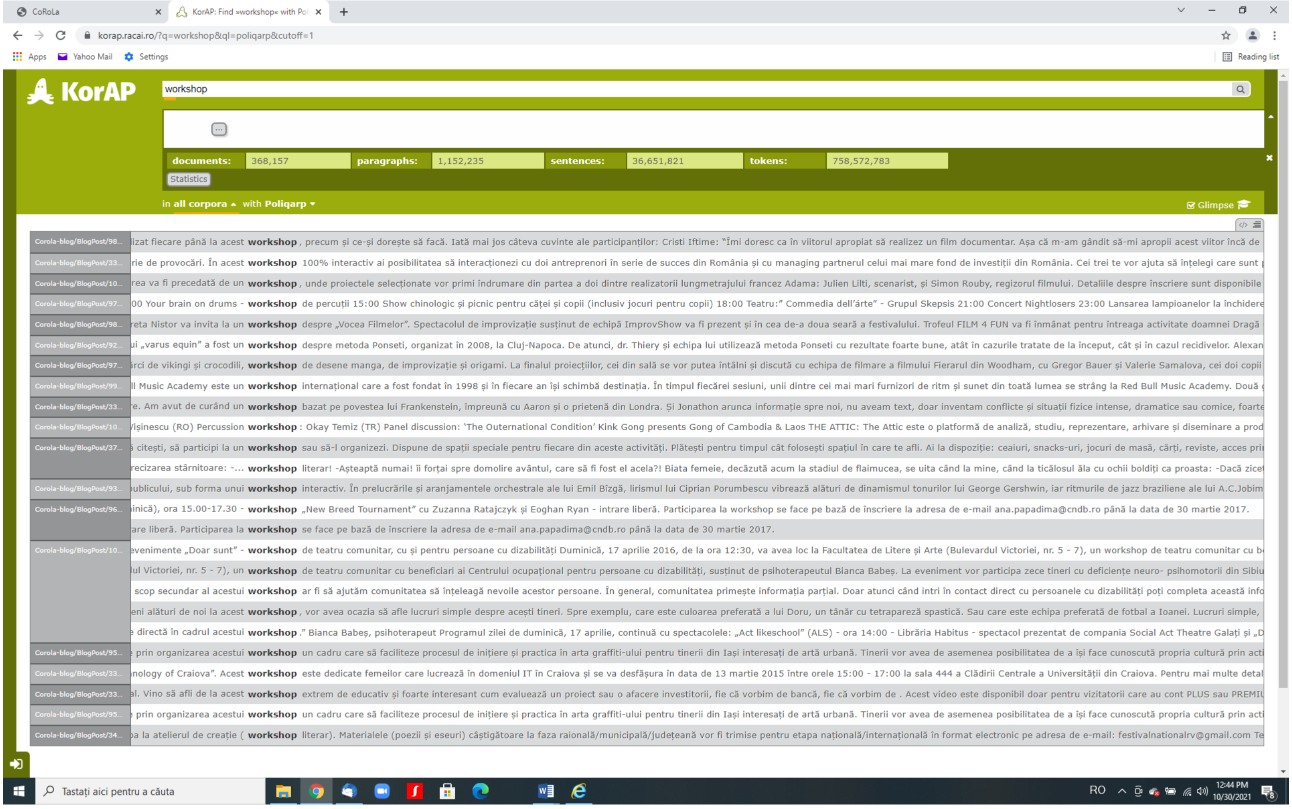
It was meant mainly for the vizitors of President of Romania's site for the presidency period of the Council of Europe. The result of this project was combined with the Automatic Speech Recognition and Text-To-Speech modules developed within the national projects ReTeRom and ROBIN thus producing the voice-to-voice translation system, first of the kind in Romania.

The first balanced large computational corpus for Romanian, ROMBAC (Ion et al., 2012), was created in 2012, in the context of the METANET4U EC project, containing about 36,000,000 words evenly distributed into five genres: journalistic (news and editorials), pharmaceutical and medical short texts, legalese, biographies of the major Romanian writers and critical reviews of their works, and fiction (both original and translated novels and poetry). The texts were tokenized, morpho-syntactically tagged, lemmatized, shallow-parsed (chunked) and XCES-compliant encoded.

ROMBAC was the predecessor of the much larger Reference Corpus for Contemporary Romanian the construction of which started in 2014 (Mititelu et al., 2014). The Reference Corpus for Contemporary Romanian Language, CoRoLa, is a bimodal corpus (https://corola.racai.ro/), built through a priority project of the Romanian Academy and continued today in the same regime. In the construction phase, the project enjoyed a close collaboration with the two institutes of the Romanian Academy (ICIA and IIT) but also with the University of Bucharest and the German Language Institute in Manheim. The bi-modal corpus (text and voice) CoRoLa is the most important public language resource in Romania (Figure 1), with more than one billion of words (Tufiș et al., 2019). There are several annotation levels which make CoRoLa corpus the queen of Romanian resources. The text data and associated metadata may be accessed by KorAP interface (Figure 2) developed by German Language Institute in Manheim (https://korap.ids-mannheim.de/), while the oral component is accessed by the OCQP interface (Figures 3 and 4) (http://89.38.230.23/corola_sound_search/index.php).

The OCQP interface also offers the possibility to listen to either the query word or to the entire sentence, spoken by native speakers. In the case of more results, the interface allows pagination and retrieval of 20 results for each page.

The corpus has undergone numerous revisions, eliminating the errors discovered by many users, adding new search facilities, being used in the construction of language models useful in developing applications for the Romanian language. Different variants of vector models (Păiș and Tufiș, 2018a; 2018b) (http://89.38.230.23/word_embeddings/), neural language models, and specific processing


CoRoLa
KorAP: Find »workshop« with Po
korap.racai.ro/?q=workshop&ql=poliqarp&cutoff=1
Apps
Yahoo Mail
Settings
Reading list
KorAP
Statistics
Glimpse
Tastați aici pentru a căuta
RO
12:44 PM
10/30/2021

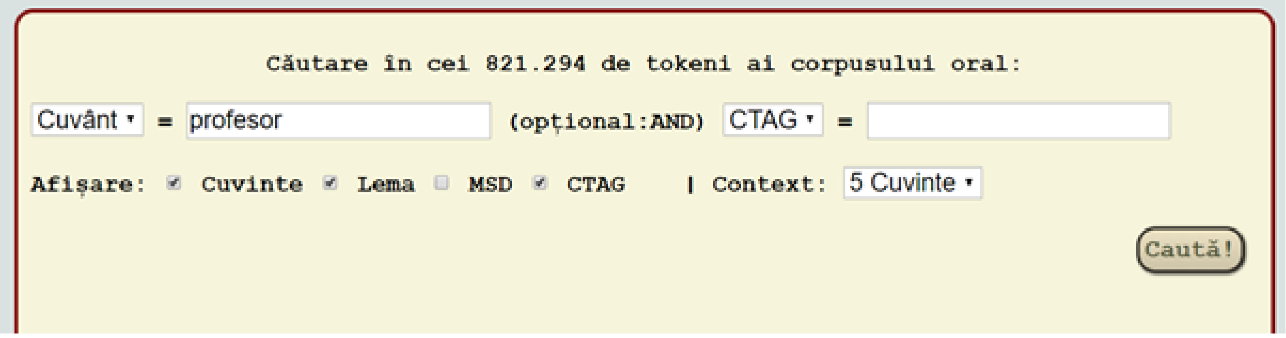
workshop

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| documents: | 368,157 | paragraphs: | 1,152,235 | sentences: | 36,651,821 | tokens: | 758,572,783 |

in all corpora with Poliqarp

| | | | |
|---|---|---|---|
| Corola-blog/BlogPost/98... | lizat fiecare până la acest | **workshop** | , precum și ce-și dorește să facă. Iată mai jos câteva cuvinte ale participanților: Cristi Iftime: "Îmi doresc ca în viitorul apropiat să realizez un film documentar. Așa că m-am gândit să-mi apropii acest viitor încă de |
| Corola-blog/BlogPost/33... | rie de provocări. În acest | **workshop** | 100% interactiv ai posibilitatea să interacționezi cu doi antreprenori în serie de succes din România și cu managing partnerul celui mai mare fond de investiții din România. Cei trei te vor ajuta să înțelegi care sunt |
| Corola-blog/BlogPost/10... | rea va fi precedată de un | **workshop** | , unde proiectele selecționate vor primi îndrumare din partea a doi dintre realizatorii lungmetrajului francez Adama: Julien Lilti, scenarist, și Simon Rouby, regizorul filmului. Detaliile despre înscriere sunt disponibile |
| Corola-blog/BlogPost/97... | 00 Your brain on drums - | **workshop** | de percuții 15:00 Show chinologic și picnic pentru cățel și copii (inclusiv jocuri pentru copii) 18:00 Teatru:" Commedia dell'árte" - Grupul Skepsis 21:00 Concert Nightlosers 23:00 Lansarea lampioanelor la închidere |
| Corola-blog/BlogPost/98... | reta Nistor va invita la un | **workshop** | despre „Vocea Filmelor". Spectacolul de improvizație susținut de echipă ImprovShow va fi prezent și în cea de-a doua seară a festivalului. Trofeul FILM 4 FUN va fi înmânat pentru întreaga activitate doamnei Dragă |
| Corola-blog/BlogPost/92... | ui „varus equin" a fost un | **workshop** | despre metoda Ponseti, organizat în 2008, la Cluj-Napoca. De atunci, dr. Thiery și echipa lui utilizează metoda Ponseti cu rezultate foarte bune, atât în cazurile tratate de la început, cât și în cazul recidivelor. Alexan |
| Corola-blog/BlogPost/97... | ărci de vikingi și crocodili, | **workshop** | de desene manga, de improvizație și origami. La finalul proiecțiilor, cei din sală se vor putea întâlni și discută cu echipa de filmare a filmului Fierarul din Woodham, cu Gregor Bauer și Valerie Samalova, cei doi copii |
| Corola-blog/BlogPost/99... | ll Music Academy este un | **workshop** | internațional care a fost fondat în 1998 și în fiecare an își schimbă destinația. În timpul fiecărei sesiuni, unii dintre cei mai mari furnizori de ritm și sunet din toată lumea se strâng la Red Bull Music Academy. Două |
| Corola-blog/BlogPost/33... | re. Am avut de curând un | **workshop** | bazat pe povestea lui Frankenstein, împreună cu Aaron și o prietenă din Londra. Și Jonathon arunca informație spre noi, nu aveam text, doar inventam conflicte și situații fizice intense, dramatice sau comice, foarte |
| Corola-blog/BlogPost/10... | Vișinescu (RO) Percussion | **workshop** | : Okay Temiz (TR) Panel discussion: 'The Outernational Condition' Kink Gong presents Gong of Cambodia & Laos THE ATTIC: The Attic este o platformă de analiză, studiu, reprezentare, arhivare și diseminare a prod |
| Corola-blog/BlogPost/37... | ă citești, să participi la un | **workshop** | sau să-l organizezi. Dispune de spații speciale pentru fiecare din aceste activități. Plătești pentru timpul cât folosești spațiul în care te afli. Ai la dispoziție: ceaiuri, snacks-uri, jocuri de masă, cărți, reviste, acces pri |
| | recizarea stârnitoare: -... | **workshop** | literar! -Așteaptă numai! îi forțai spre domolire avântul, care să fi fost el acela?! Biata femeie, decăzută acum la stadiul de flaimucea, se uita când la mine, când la ticălosul ăla cu ochii bolditi ca proasta: -Dacă zice |
| Corola-blog/BlogPost/93... | publicului, sub forma unui | **workshop** | interactiv. În prelucrările și aranjamentele orchestrale ale lui Emil Bîzgă, lirismul lui Ciprian Porumbescu vibrează alături de dinamismul tonurilor lui George Gershwin, iar ritmurile de jazz braziliene ale lui A.C.Jobim |
| Corola-blog/BlogPost/96... | inică), ora 15.00-17.30 - | **workshop** | „New Breed Tournament" cu Zuzanna Ratajczyk și Eoghan Ryan - intrare liberă. Participarea la workshop se face pe bază de înscriere la adresa de e-mail ana.papadima@cndb.ro până la data de 30 martie 2017. |
| | are liberă. Participarea la | **workshop** | se face pe bază de înscriere la adresa de e-mail ana.papadima@cndb.ro până la data de 30 martie 2017. |
| Corola-blog/BlogPost/10... | evenimente „Doar sunt" - | **workshop** | de teatru comunitar, cu și pentru persoane cu dizabilități Duminică, 17 aprilie 2016, de la ora 12:30, va avea loc la Facultatea de Litere și Arte (Bulevardul Victoriei, nr. 5 - 7), un workshop de teatru comunitar cu b |
| | lul Victoriei, nr. 5 - 7), un | **workshop** | de teatru comunitar cu beneficiari ai Centrului ocupațional pentru persoane cu dizabilități, susținut de psihoterapeutul Bianca Babeș. La eveniment vor participa zece tineri cu deficiențe neuro- psihomotorii din Sibiu |
| | scop secundar al acestui | **workshop** | ar fi să ajutăm comunitatea să înțeleagă nevoile acestor persoane. În general, comunitatea primește informația parțial. Doar atunci când intri în contact direct cu persoanele cu dizabilități poți completa această info |
| | eni alături de noi la acest | **workshop** | , vor avea ocazia să afle lucruri simple despre acești tineri. Spre exemplu, care este culoarea preferată a lui Doru, un tânăr cu tetrapareză spastică. Sau care este echipa preferată de fotbal a Ioanei. Lucruri simple, |
| | e directă în cadrul acestui | **workshop** | ." Bianca Babeș, psihoterapeut Programul zilei de duminică, 17 aprilie, continuă cu spectacolele: „Act likeschool" (ALS) - ora 14:00 - Librăria Habitus - spectacol prezentat de compania Social Act Theatre Galați și „D |
| Corola-blog/BlogPost/95... | e prin organizarea acestui | **workshop** | un cadru care să faciliteze procesul de inițiere și practica în arta graffiti-ului pentru tinerii din Iași interesați de artă urbană. Tinerii vor avea de asemenea posibilitatea de a își face cunoscută propria cultură prin acti |
| Corola-blog/BlogPost/33... | nology of Craiova". Acest | **workshop** | este dedicate femeilor care lucrează în domeniul IT în Craiova și se va desfășura în data de 13 martie 2015 între orele 15:00 - 17:00 la sala 444 a Clădirii Centrale a Universității din Craiova. Pentru mai multe detal |
| Corola-blog/BlogPost/33... | al. Vino să afli de la acest | **workshop** | extrem de educativ și foarte interesant cum evaluează un proiect sau o afacere investitorii, fie că vorbim de bancă, fie că vorbim de . Acest video este disponibil doar pentru vizitatorii care au cont PLUS sau PREMIU |
| Corola-blog/BlogPost/95... | e prin organizarea acestui | **workshop** | un cadru care să faciliteze procesul de inițiere și practica în arta graffiti-ului pentru tinerii din Iași interesați de artă urbană. Tinerii vor avea de asemenea posibilitatea de a își face cunoscută propria cultură prin acti |
| Corola-blog/BlogPost/34... | pa la atelierul de creație ( | **workshop** | literar). Materialele (poezii și eseuri) câștigătoare la faza raională/municipală/județeană vor fi trimise pentru etapa națională/internațională în format electronic pe adresa de e-mail: festivalnationalrv@gmail.com Te |

Figure 2: KorAP results for a query on CoRoLa

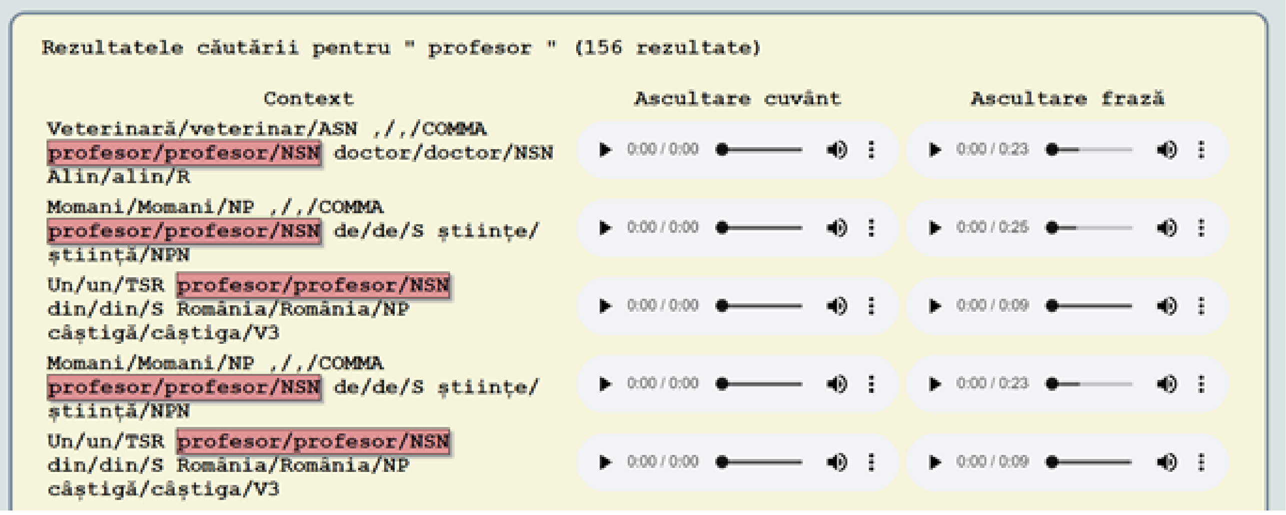


Căutare în cei 821.294 de tokeni ai corpusului oral:

Cuvânt ▾ = profesor (opțional:AND) CTAG ▾ =

Afișare: ☑ Cuvinte ☑ Lema □ MSD ☑ CTAG | Context: 5 Cuvinte ▾

Figure 3: OCQP - Query on the oral part of CoRoLa



Rezultatele căutării pentru " profesor " (156 rezultate)

| Context | Ascultare cuvânt | Ascultare frază |
|---|---|---|
| Veterinară/veterinar/ASN ,/,/COMMA profesor/profesor/NSN doctor/doctor/NSN Alin/alin/R | 0:00 / 0:00 | 0:00 / 0:23 |
| Momani/Momani/NP ,/,/COMMA profesor/profesor/NSN de/de/S științe/știință/NPN | 0:00 / 0:00 | 0:00 / 0:25 |
| Un/un/TSR profesor/profesor/NSN din/din/S România/România/NP câștigă/câștiga/V3 | 0:00 / 0:00 | 0:00 / 0:09 |
| Momani/Momani/NP ,/,/COMMA profesor/profesor/NSN de/de/S științe/știință/NPN | 0:00 / 0:00 | 0:00 / 0:23 |
| Un/un/TSR profesor/profesor/NSN din/din/S România/România/NP câștigă/câștiga/V3 | 0:00 / 0:00 | 0:00 / 0:09 |

Figure 4: OCQP – results for a query on the oral part of CoRoLa

tools were created based on it. All results were made public, by publishing at specialized conferences or in professional journals, as well as by providing public access to the corpus and associated processing facilities. The bi-modal corpus (text and voice) CoRoLa is the most important public language resource in Romania. The CoRoLa corpus has been and is being used intensively for the objectives of the European projects Presidency MT-Kit, CURLICAT, Nexus-Linguarum, ELRC, ELG, ELE etc.

Another significant corpus was created during the MARCELL Project (Váradi et al., 2020) (https://marcell-project.eu/). This specialized multilingual corpus contains legislative documents in 7 languages. The Romanian subcorpus (MARCELL-Ro) (Tufiș et al., 2020) includes over 163,000 documents (257.803.124 distinct words) and is morpho-lexically and terminologically processed (with the annotation of the terms in IATE and EUROVOC). It is distributed free of charge, via ELRC share platform (https://elrc-share.eu/) to all interested parties. In addition to the textual corpus, an automatic packed processing stream (docker) was created, also distributed free of charge to all those interested. Part of the MARCELL-Ro corpus (over 265,000 words) has been annotated and manually validated with tags for named entities (organizations, locations, people, time, and legal references) and in addition provides GEONAMES codes for named entities annotated as location. The LegalNERo dataset (Păiș et al., 2021) was used to train a neural system for recognizing named entities in the legal field for the Romanian language. The results obtained (F-score of 90.36) are competitive, at the level of the best world results.

Other datasets, MoNERo (Mitrofan et al., 2019) – publicly released (https://www.racai.ro/media/MoNERo_2019.7z) and its syntactically parsed version SiMoNERo (Barbu Mititelu, V., Mitrofan, M., 2020), are gold standards for biomedical domain, manually annotated with four types of domain-specific named entities. They were created for the purpose of training and evaluating NER systems in the medical domain. These datasets (LegalNERo, MoNERo and SiMoNERo) are multiple level annotated (lexical, morphologic, and syntactic), hand validated and corrected by experts, thus being extremely valuable for building, fine-tuning, and evaluating various applications (creating or enriching specific terminologies, machine translation systems, anonymization tools, etc.).

### 3.3 Speech resources and datasets for Romanian language

Besides the textual linguistic resources mentioned before, ICIA developed other language resources for supporting speech data processing. RoLEX is a phonological lexicon, developed in the ReTeRom project, providing morphosyntactic information, lemma, syllabification, stress, and phonetic transcription of words. Within the ReTeRom project, multiple technologies for Romanian language were developed. These include text-to-speech synthesis with expressivity, speech recognition models, technologies for processing written Romanian language and a thesaurus with audio and textual resources, annotated at different acoustic and linguistic levels. In this context, RoLEX was constructed based on the vocabulary extracted from the textual component of a speech corpus that contained data from the Romanian Wikipedia, news, interviews, talk-shows, spontaneous speech, fairy tales, novels. Syllabification, stress, and phonetic transcription information is based, partially, on previous resources like RoSyllabiDict (Barbu, 2008) and MaRePhor (Toma et al., 2017). The MaRePhor vocabulary contains a total of 72,375 entries from the Romanian Scrabble Association's official list of words and the entries from an additional 15,517 words dictionary, developed according to the SpeechDat (https://www.speechdat.org/) specifications. Following the automatic processing, and manual corrections on the RoLEX lexicon, the outcome was the largest validated resource of this type available for Romanian language (330,866 entries).

The ReTeRom project also created extensive bi-modal resources, managed by an innovative platform, COBILIRO (Cristea et al., 2020), developed by UAIC and on which partners from UPB, UTCN and ICIA uploaded bimodal resources which were further processed with TEPROLIN (http://relate.racai.ro/?path=teprolin/doc_dev), the processing flow implemented by ICIA.

ROBIN, another complex project in which ICIA coordinated one of the component projects (ROBIN-Dialog) provided the community with important linguistic resources for dialogues in Romanian with the robot PEPPER in specified discourse universes (a complex dictionary, a corpus of dialogues) and a general module for managing communication with the robot (Tufiș et al., 2019), (Ion et al., 2020). The ROBIN Technical Acquisition Speech Corpus (RTASC) (Păiș et al., 2021) was created as a read speech corpus in Romanian language to be used for the development of a speech-mediated dialog system with a personal robot. It was recorded by 6 native Romanian speakers of different genders (3 males and 3 females) and ages. The text component of the corpus was automatically annotated with information such as lemma, part-of-speech tags, and dependency parsing. In addition to text and audio data, metadata was stored, containing corpus and speaker characteristics, including number of sentences, total duration, speaker's gender and age, number of recorded files by each speaker, information about recording device used. To anonymize the speaker related data, the name is not given and the age is stored only as intervals (for example "40-50" years). All the new linguistic resources were further converted to linguistic linked data format, to open more ways for its exploitation as well as to allow linking with other corpora and facilitate queries that span multiple datasets. We adhered to the linguistic link open data in the context of the challenging Nexus Linguarum COST action (https://nexuslinguarum.eu/).

Participation in COST projects (LITHME, NEXUS-Linguarum, PAN) as well as in the dissemination of EC policies in the field of language technologies (European Language Grid, European Language Equality and European Language Resource Coordination) has placed ICIA among the European institutions of excellence in the field of artificial intelligence focused on natural language technologies.

### 3.4 Recent developments of language technologies for Romanian language

The research topics of the institute are aligned with the current priorities in the field of artificial intelligence, machine learning and natural language processing, capitalizing on the achievements of previous years:

- Processing texts in Romanian with the help of complex neural networks
- Linguistic resources for Romanian language processing in the current global context (Link Linguistic Open Data) in close collaboration with the Nexus-Linguarum project
- Intelligent systems for stylistic and pragmatic analysis, from a polyphonic perspective, of language on the social web
- Affective computing system for human voice analysis

In all the projects in which the researchers of the institute participated, remarkable results were obtained: linguistic resources (both textual and oral) of great size and special quality, new tools, superior to the previous ones and internationally competitive, publications at international reference conferences and associated workshops (LREC, EMNLP, NAACL, NLDB, Global WordNet Conference, RANLP, SpeD, etc.), in ISI journals (Artificial Intelligence Reviews, Proceedings of the Romanian Academy, Studies in Informatics and Control) as well as book chapters in volumes published by prestigious publishing houses (Springer, ACL, Hungarian Academy Publishing House).

The newest (neural) processing tools of the Romanian language have been made public on the git-hub site of the institute (https://github.com.racai.ai/repositories). This new architecture, called RODNA (Romanian Deep Neural Network Architectures) is a Python 3/TensorFlow /Keras project and includes high-performance, essential modules specifically targeted at Romanian text processing (sentence splitter, tokenizer, morphology analyzer, POS tagger, dependency parser).

A remarkable achievement, with great impact among the NLP community is the RELATE portal (https://relate.racai.ro/) which provides users with all the fundamental tools (BLARK) for processing the Romanian language, written or spoken (Păiș et al., 2019, 2020), (Păiș, 2020). It includes technologies and language resources developed by ICIA and its partners in several projects: COROLA, RETEROM, ROBIN, PRESIDENCY, MARCELL, CURLICAT. Additionally, it allows for creation of manually annotated gold corpora and was successfully used for creating the LegalNERo named entity corpus and the RTASC speech corpus.

The Basic Language Advanced Research Kit for Romanian (BLARK-Ro) is available as a configurable processing chain, ensuring text normalization, diacritical marks restoration, punctuation restoration, phrase and lexical segmentation, syllable division, accent identification, phonetic transcription, expanded abbreviations, morpho-lexical annotation, lemmatization, recognition of named

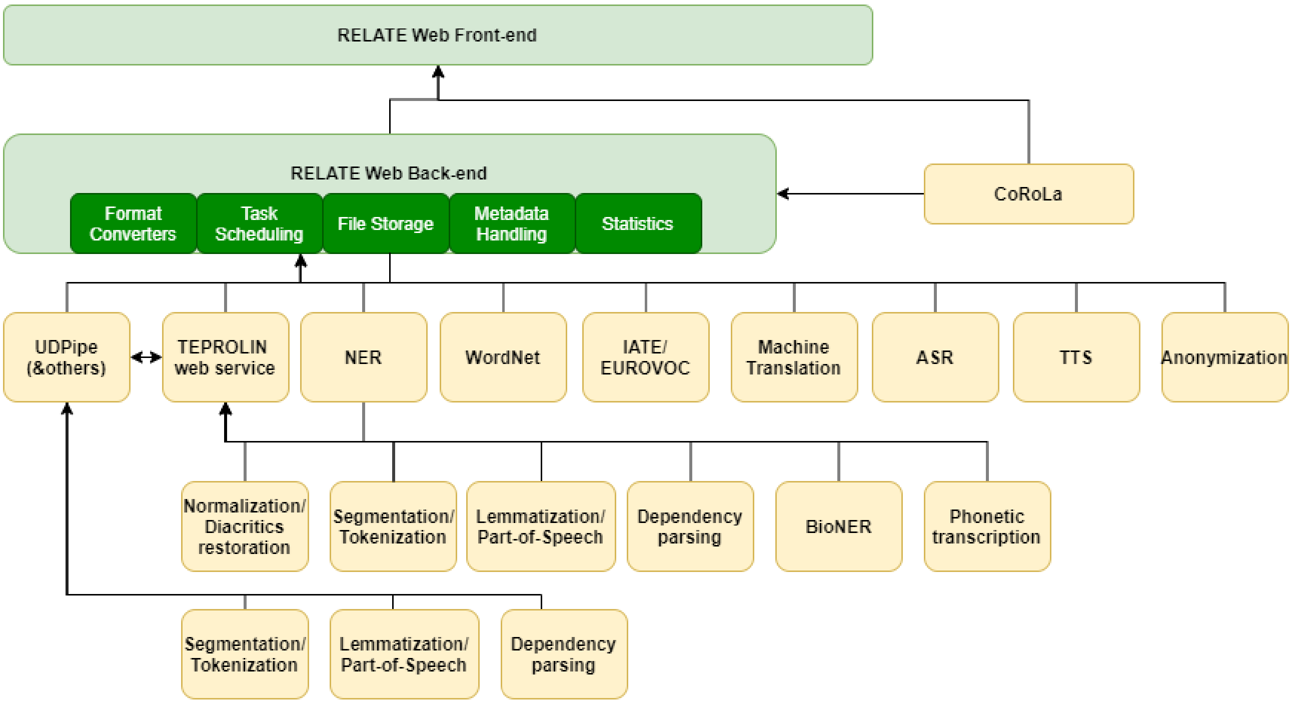


Figure 5: The architecture of RELATE portal

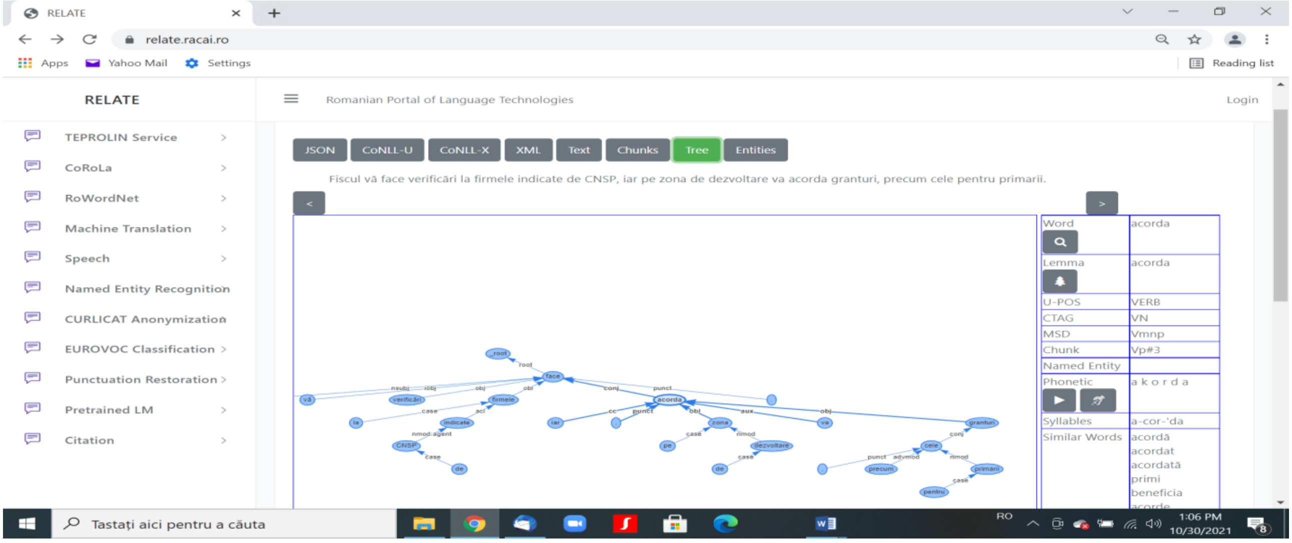
entities, syntactic analysis). Besides basic processing modules, the RELATE portal integrates the interfaces to CoRoLa corpus query and RoWordNet lexicons, EUROVOC classification of documents, anonymization of texts and automatic text and voice translation (ro-en). The RELATE portal is open source (https://github.com/racai-ai/RELATE) and offers users a range of (neural) language models and pre-trained semantic vectors, ready to be incorporated into practical applications. The implementation of the RELATE portal is aligned with the development philosophy of European Language Grid (https://www.european-language-grid.eu/), relying on WEB services, REST APIs and DOCKERs packaging of processing flows. The services may be distributed on multiple network nodes and may be consumed directly from the partners. Being heavily used for deep processing of large and very large corpora (more than 200,000 documents) in our international running projects, we may say that RELATE riched a maturity implementation level and it is a robust infrastructure, allowing for both CPU and GPU processing.

Architecturally, RELATE has two main components (see figures 5 and 6): a Web Front-end (freely accessible) and a Web back-end (accessible after log-in with free access credentials). Both interfaces offers the full set of 18 (for now) processing modules for text and speech data and various visualization modes. While the Web Front-end allows processing of a single document, the Web back-end facilitates processing of mass textual data by parallel processing (task scheduling, services from multiple nodes). Additionaly, the back-end offers services for corpora management (create, upload, download, archive, annotate, statistics, visualize, converting among formats CoNLL-U, CoNLL-U Plus, XML, JSON, RDF), metadata management, statistics, creation of gold corpora: integrates BRAT for NER, speech recorder for speech-text aligned corpora.

The RELATE portal offers its users a series of pretrained neural language models and semantic vectors (word embeddings): ROBERT (available in two versions, cased and uncased), word embeddings from CoRoLa corpus, annotation models for lemma, UPOS/XPOS tagging, classification and dependency parsing models.

### 3.5 Open access to resources and technologies for Romanian language

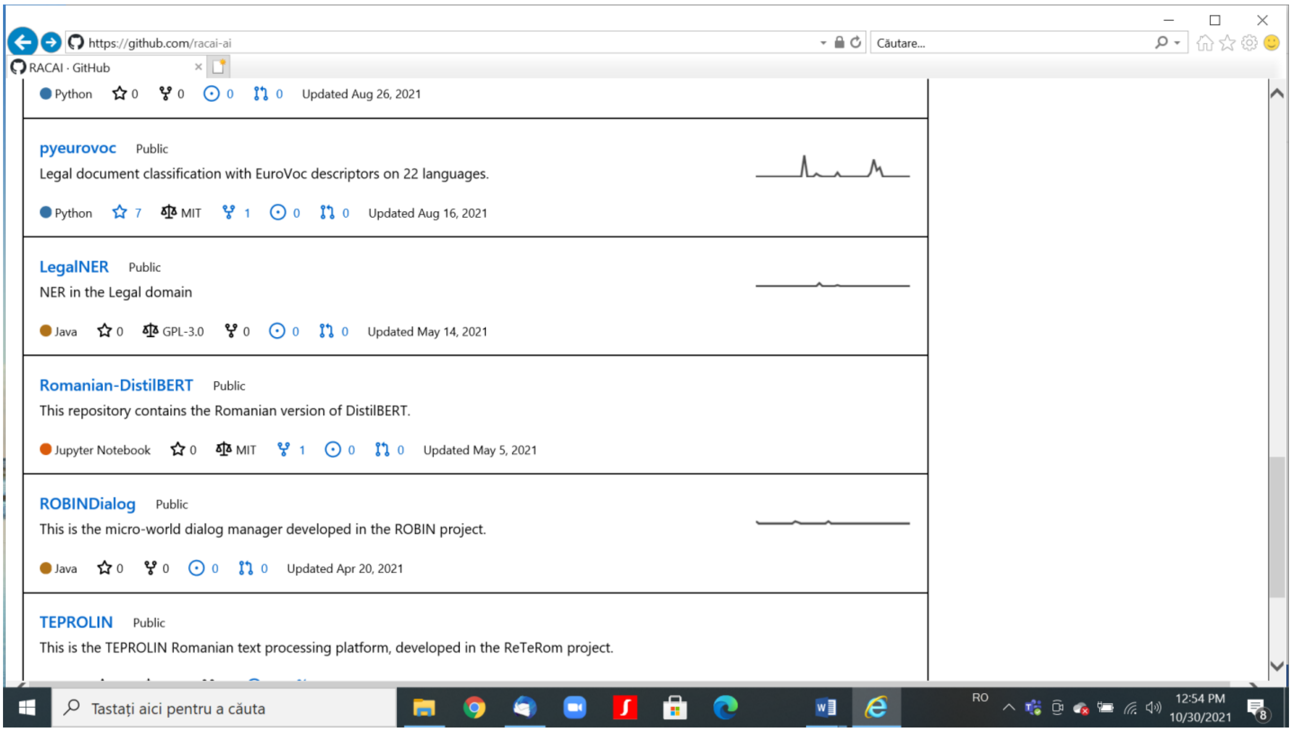
As mentioned several times, almost all the language resources and processing tools are open source, available in the github of the institute (https://github.com/racai-ai). The content of this github complements the resources and tools for Romanian available on the ELRC-share platform (Figure 9).



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

| | |
|---|---|
| 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-'da |
| Similar Words | acordă<br>acordat<br>acordată<br>primi<br>beneficia<br>acorde |

Figure 6: The visualization of a parse tree



Python ☆ 0 0 0 0 Updated Aug 26, 2021

**pyeurovoc** Public

Legal document classification with EuroVoc descriptors on 22 languages.

Python ☆ 7 MIT 1 0 0 Updated Aug 16, 2021

**LegalNER** Public

NER in the Legal domain

Java ☆ 0 GPL-3.0 0 0 0 Updated May 14, 2021

**Romanian-DistilBERT** Public

This repository contains the Romanian version of DistilBERT.

Jupyter Notebook ☆ 0 MIT 1 0 0 Updated May 5, 2021

**ROBINDialog** Public

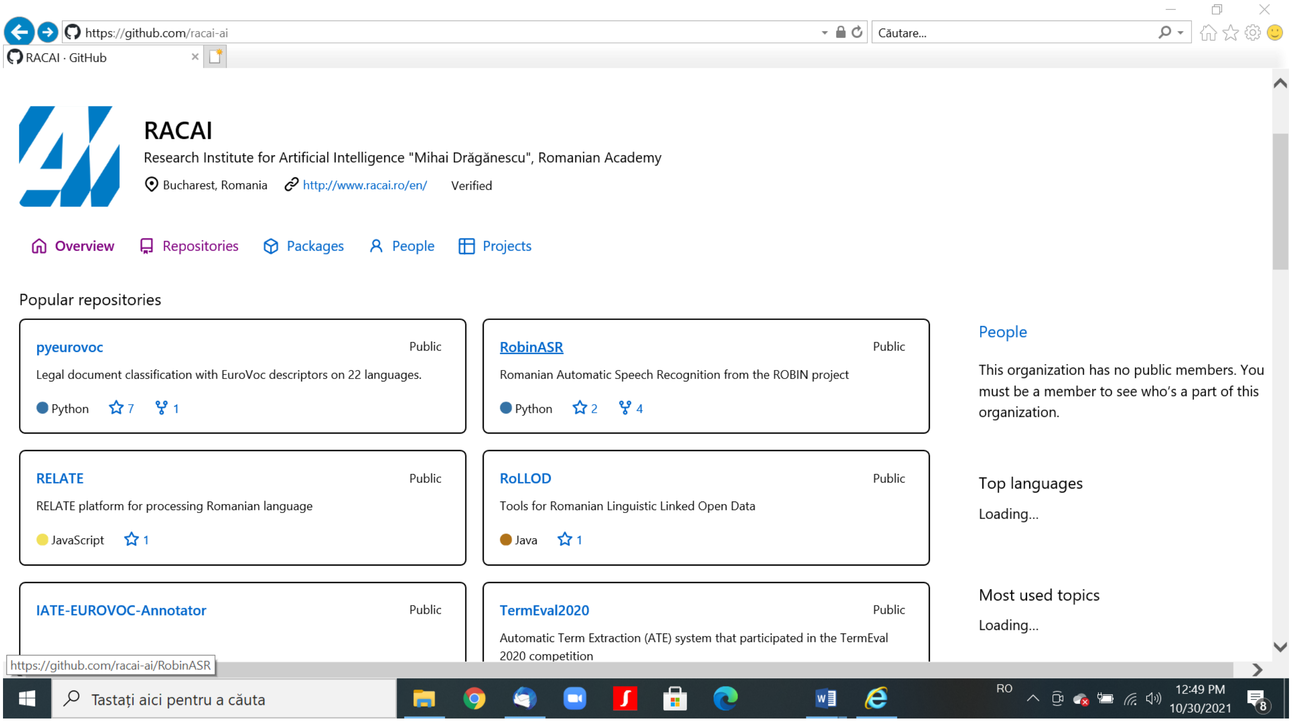
This is the micro-world dialog manager developed in the ROBIN project.

Java ☆ 0 0 0 0 Updated Apr 20, 2021

**TEPROLIN** Public

This is the TEPROLIN Romanian text processing platform, developed in the ReTeRom project.

Figure 7: Some open-source tools and language resources available in the github of the institute


https://github.com/racai-ai
Căutare...
RACAI · GitHub
https://github.com/racai-ai/RobinASR
Tastați aici pentru a căuta
RO
12:49 PM
10/30/2021

RACAI

Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy

Bucharest, Romania http://www.racai.ro/en/ Verified

Overview Repositories Packages People Projects

Popular repositories

| Repository | Visibility | Description | Language | Stars | Forks |
|---|---|---|---|---|---|
| pyeurovoc | Public | Legal document classification with EuroVoc descriptors on 22 languages. | Python | 7 | 1 |
| RobinASR | Public | Romanian Automatic Speech Recognition from the ROBIN project | Python | 2 | 4 |
| RELATE | Public | RELATE platform for processing Romanian language | JavaScript | 1 | |
| RoLLOD | Public | Tools for Romanian Linguistic Linked Open Data | Java | 1 | |
| IATE-EUROVOC-Annotator | Public | | | | |
| TermEval2020 | Public | Automatic Term Extraction (ATE) system that participated in the TermEval 2020 competition | | | |

People

This organization has no public members. You must be a member to see who's a part of this organization.

Top languages

Loading...

Most used topics

Loading...

Figure 8: Other open-source tools and language resources available in the github of the institute

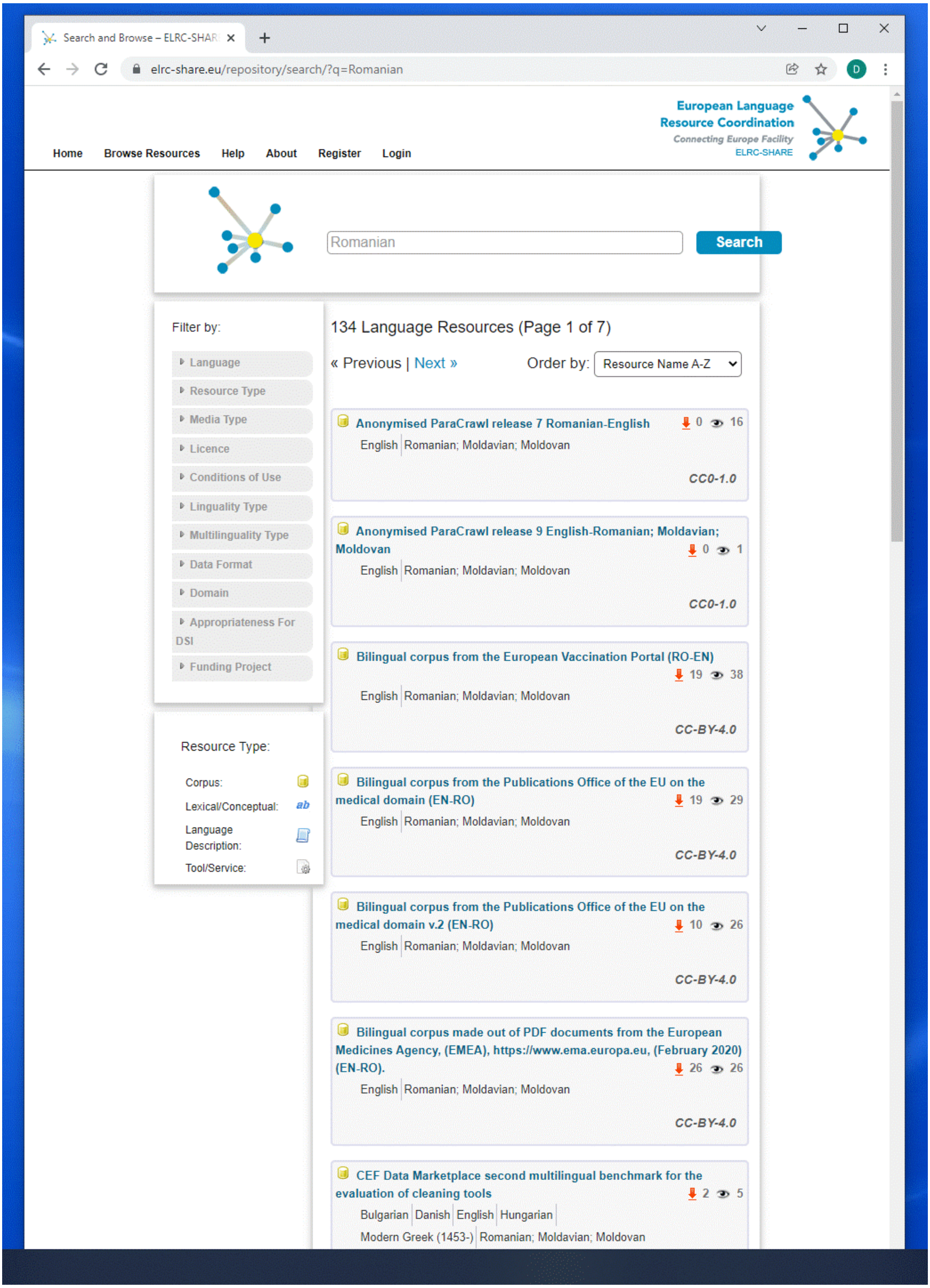
# 4 Conclusions and further work

The resources and technologies available for Romanian, developed since the publishing of META-NET report, filled most of the identified gaps. The running ELE (European Language Equality) and ELG (European Language Grid) projects will publish the Strategic Agenda and Roadmap for reaching the goal of technological European language equality by 2030. The language reports will be updated to reflect the current state of play in European language technology. The ICIA actions reported here answered the commandments defined 10 years ago. The advancement of language science and technology, certainly, will call for new priorities and more efforts to meet the Agenda and the roadmap for reaching the digital language equality by 2030.

## Aknowledgements

The results reviewed in this article have been obtained by a team of dedicated researchers: Radu Ion, Vasile Păiș, Elena Irimia, Verginica Barbu Mititelu, Maria Carp, Marius Andrei Avram and Eric Curea. Former colleagues essentially contributed the results reviewed here: Alexandru Ceaușu, Dan Ștefănescu, Tiberiu Boroș, Daniel Ștefan Dumitrescu. The partnerships with University A.I. Cuza of Iași, Institute of Computer Science of Iași, Technical University of Cluj-Napoca, Faculty of Electronics, Telecommunications and Information Technology, UPB Romania, University of Bucharest and IDS-Mannheim, and with colleagues in all the projects facilitated gluing creative energies from different competence areas, building highly professional research teams. To all of them we express our gratitude.

This article is a tribute to Prof. Ioan Dziţac, to whom this special issue of the journal is dedicated. He was not only the heart of this journal, but also the champion of Romanian information and communication science promotion. As one of the managers of the International Journal of Computers, Communication and Control, for which he was one of the founders, Ioan managed, in less than one year, to include the journal in the mainstream publications. The regular editions (biennial) of the Conference on Computers Communications and Control (ICCCC) he founded, were masterly managed and being among the highest quality scientific events in Romania got quickly international reputation and thus, attracted top lecturing scholars. One of these, was Prof. Lotfi Zadeh, a giant of the information science, who honoured Prof. Dizţac, Acad. Fl. Filip and myself in co-organizing the reference workshop From



134 Language Resources (Page 1 of 7)

« Previous | Next »

Order by: Resource Name A-Z

Anonymised ParaCrawl release 7 Romanian-English ⬇ 0 👁 16
English | Romanian; Moldavian; Moldovan
CC0-1.0

Anonymised ParaCrawl release 9 English-Romanian; Moldavian; Moldovan ⬇ 0 👁 1
English | Romanian; Moldavian; Moldovan
CC0-1.0

Bilingual corpus from the European Vaccination Portal (RO-EN) ⬇ 19 👁 38
English | Romanian; Moldavian; Moldovan
CC-BY-4.0

Bilingual corpus from the Publications Office of the EU on the medical domain (EN-RO) ⬇ 19 👁 29
English | Romanian; Moldavian; Moldovan
CC-BY-4.0

Bilingual corpus from the Publications Office of the EU on the medical domain v.2 (EN-RO) ⬇ 10 👁 26
English | Romanian; Moldavian; Moldovan
CC-BY-4.0

Bilingual corpus made out of PDF documents from the European Medicines Agency, (EMEA), https://www.ema.europa.eu, (February 2020) (EN-RO). ⬇ 26 👁 26
English | Romanian; Moldavian; Moldovan
CC-BY-4.0

CEF Data Marketplace second multilingual benchmark for the evaluation of cleaning tools ⬇ 2 👁 5
Bulgarian | Danish | English | Hungarian |
Modern Greek (1453-) | Romanian; Moldavian; Moldovan

Figure 9: ELRC-Share lists 134 language resources and tools for Romanian language

Natural Language to Soft Computing: New Paradigms in Artificial Intelligence (Zadeh et al., 2008). This year, ICCCC will be the 8th edition and everybody will miss Ioan Dziţac but will feel him around with his always warm and carrying presence.

## References

[1] Avram, A.-M., Păiș, V., and Tufiș, D. (2020a). Towards a romanian end-to-end automatic speech recognition based on deepspeech2. Proceedings of the Romanian Academy Series A, 21:395–402.

[2] Avram, A.-M., Păiș, V., and Tufiș, D. (2020b). Romanian speech recognition experiments from the robin project. In The 15th International Conference on Linguistic Re-sources and Tools for Natural Language Processing, pages 103–114.

[3] Barbu, A.-M. (2008). Romanian lexical data bases: Inflected and syllabic forms dictionaries. In LREC.

[4] Barbu Mititelu, V., Irimia, E., Tufiș, D. (2014). CoRoLa — The Reference Corpus of Contemporary Romanian Language. In Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14), May 26-31, 2014, Reykjavik, Iceland, ISBN 978-2-9517408-8-4, pages 1235-1239.

[5] Barbu Mititelu, V., Tufiș , D., and Irimia, E. (2018). The reference corpus of the contemporary romanian language (corola). In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018), pages 1178–1185.

[6] Barbu Mititelu, V., Tufiș, D., Irimia, E., Păiș, , V., Ion, R., Diewald, N., Mitrofan, M., and, Onofrei, M. (2019). Little strokes fell great oaks. creating corola, the reference corpus of contemporary romanian. In Revue Roumaine de Linguistique, No./Issue 3.

[7] Boros, , T., Dumitrescu, C. D., and Păiș, V. (2018). Tools and resources for romanian text-to- speech and speech-to-text applications. In Proceedings of the International Conference on Human-Computer Interaction (RoCHI), pages 46–53.

[8] Ceaușu, A., Tufiş, D. (2011) Addressing SMT Data Sparseness when Translating into Morphologically-Rich Languages. In Bernadette Sharp, Michael Zock, Michael Carl, and Arnt Lykke Jakobsen (eds.) Proceedings of the 8th international NLPCS workshop. Special theme: Human-machine interaction in translation, pp. 57-68, Copenhagen Business School, 20-21 August 2011.

[9] Cimiano, P., Chiarcos, C., McCrae, J. P., and Gracia, J. (2020). Linguistic Linked Data. Representation, Generation and Applications. Springer.

[10] Cristea, D., Diewald, N., Haja, G., Mărănduc, C., Barbu Mititelu, V., and Onofrei, M. (2019). How to find a shining needle in the haystack. querying corola: solutions and perspectives. RRL, (3):279–292.Fazekas, G. and Sandler, M. B. (2011). The studio ontology framework. In 12th International Society for Music Information Retrieval Conference (ISMIR).

[11] Cristea, D., Pistol, I., Boghiu, Ș., Bibiri, A., D., Gîfu, D., Onofrei, M., Trandabăț, D., Bugeag, G. (2020). CoBiLiRo: A Research Platform for Bimodal Corpora. Proceedings of the 1st International Workshop on Language Technology Platforms, LREC 2020, Marseille, pages 22-27,

[12] Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. Bradford Books. Hauck, E., Ewert, D., Gramatke, A., and Henning, K. (2011). Software architecture, knowledge compiler and ontology design for cognitive technical systems suitable for controlling assembly tasks. In Jeschke, S., Isenhardt, I., and Henning, K., editors, Automation, Communication and Cybernetics in Science and Engineering 2009/2010, pages 383–391, Berlin, Heidelberg. Springer Berlin Heidelberg.

[13] Ide, N. and Pustejovsky, J. (2010). What does interoperability mean , anyway ? toward an operational definition of interoperability for language technology. In Proceedings of the 2nd International Conference on Global Interoperability for Language Resources (ICGL 2010).

[14] Ion, R. (2012) Graphic Comparability Levels for Comparable Corpora. In Mihai Alex Moruz, Dan Cristea, Dan Tufiş, Adrian Iftene, Horia-Nicolai Teodorescu (eds.) Proceedings of the 8th International Conference "Linguistic Resources and Tools for Processing of the Romanian Language", pp. 127-133, April 26-27, 2012

[15] Ion, R., Tufiş, D., Boroş, T., Ceauşu, A., Ştefănescu D. (2010). On-Line Compilation of Comparable Corpora and their Evaluation. In Marko Tadić, Mila Dimitrova-Vulchanova, and Svetla Koeva (eds.), Proceedings of The 7th International Conference Formal Approaches to South Slavic and Balkan Languages (FASSBL-7), pp. 29—34, Croatian Language Technologies Society – Faculty of Humanities and Social Sciences, Zagreb, Croatia, October 2010. ISBN: 978-953-55375-2-6.

[16] Ion, R., Ceaușu Al., Irimia, E. (2011): An Expectation Maximization Algorithm for Textual Unit Alignment. In Proceedings of the 4th Workshop on Building and Using Comparable Corpora, pages 128-135, The 49th Annual Meeting of the Association for Computational Linguistics, Portland, Oregon, 2011.

[17] Ion, R. (2012) PEXACC: A Parallel Sentence Mining Algorithm from Comparable Corpora. Proceedings of the Eigth International Conference on Language Resources and Evaluation (LREC'2012), pages 2181-2188.

[18] Ion, R., Irimia, E., Ștefănescu, D., Tufiș, D. (2012): ROMBAC: The Romanian Balanced Annotated Corpus. In Proceedings of the 8th LREC Conference, Istanbul, Turkey, 21-27 May, 2012, pp.339-344, ISBN 978-2-9517408-7-7.

[19] Ion, R. (2018). TEPROLIN: An extensible, online text preprocessing platform for Romanian. In The 13th International Conference on Linguistic Resources and Tools for Natural Language Processing - CONSILR.

[20] Ion, R., Badea, V. G., Cioroiu, G., Barbu Mititelu, V., Irimia, E., Mitrofan, M., and Tufiș, D. (2020). A dialog manager for micro-worlds. Studies in Informatics and Control, 29(4):411–420.

[21] Irimia, E. (2012). Experimenting with Extracting Lexical Dictionaries from Comparable Corpora for English-Romanian language pair, In Proceedings of The Fifth Workshop on Building and Using Comparable Corpora (5th BUCC), LREC 2012, Istanbul Turkey.

[22] Irimia, E. (2011). DEACC – Lexical Dictionary Extractor from Comparable Corpora, In Proceedings of the 8th International Conference "Linguistic Resources and Tools for Processing of the Romanian Language", December 8-9, 2011 and April 26-27, 2012, Bucharest, Romania, Eds. Moruz, Mihai Alex; Cristea, Dan; Tufis, Dan; Iftene, Adrian; Teodorescu, Horia-Nicolai, "Alexandru Ioan Cuza" University Publishing House, Iași, pp. 173-180.

[23] Klyne, G., Carroll, J., and McBride, B. (2004). Resource Description Framework (RDF): Concepts and Abstract Syntax.

[24] Koehn, P., Hoang, H., Birch, A., Burch, C., C., Federico, M., Bertoldi, N., Cowan, B., Shen, W., Moran, C, Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E. (2007). Moses: Open Source Toolkit for Statistical Machine Translation. Proceedings of the ACL 2007, Prague, pages 177-180.

[25] Kumar, K., Haider, M. T. U., and Ahsan, S. S. (2021). Ontology-based full-text searching using named entity recognition. In Hura, G. S., Singh, A. K., and Siong Hoe, L., editors, Advances in Communication and Computational Technology, pages 211–222, Singapore. Springer Singapore.

[26] Manzoor, S., Rocha, Y. G., Joo, S.-H., Bae, S.-H., Kim, E.-J., Joo, K.-J., and Kuc, T.-Y. (2021). Ontology-based knowledge representation in robotic systems: A survey oriented toward applications. Applied Sciences, 11(10).

[27] Miller, George A. (1995). WordNet: A Lexical Database for English. Communications of the ACM Vol. 38, No. 11: 39-41

[28] Mititelu Barbu, V., Irimia, E., Tufiș, D. (2014). CoRoLa – The Reference Corpus of Contemporary Romanian Language.

[29] Mitrofan, M., Barbu Mititelu, V., Mitrofan G. (2019). MoNERo: a Biomedical Gold Standard Corpus for the Romanian Language. In Proceedings of the BioNLP workshop. Association for Computational Linguistics, Florence, Italy, pp. 71-79, aug 2019

[30] Oltramari, A. and Lebiere, C. (2013). Knowledge in Action: Integrating Cognitive Architectures and Ontologies, pages 135–154. Springer Berlin Heidelberg, Berlin, Heidelberg.

[31] Păiș, V. and Tufiș, D. (2018a) Computing distributed representations of words using the CoRoLa corpus. In Proceedings of the Romanian Academy Series A - Mathematics Physics Technical Sciences Information Science. vol. 19, no. 2, pp. 185–191.

[32] Păiș, V. and Tufiș, D. (2018b). More Romanian word embeddings from the RETEROM project. In Proceedings of the International Conference on Linguistic Resources and Tools for Processing Romanian Language - CONSILR. pp. 91-100

[33] Păiș, V., Mitrofan, M., Gasan, C. L., Coneschi, Vl., and Ianov, A. (2021). Named Entity Recognition in the Romanian Legal Domain. In Proceedings of the Natural Legal Language Processing Workshop 2021. Association for Computational Linguistics, Punta Cana, Dominican Republic, pp. 9–18, Nov. 2021.

[34] Păiș, V., Mitrofan, M. (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. Skopje, North Macedonia, pp. 16–17, Sept. 2021

[35] Păiș, V., Ion, R., Barbu Mititelu, V., Irimia, E., Mitrofan, M., and Avram, A. (2021). Robin technical acquisition speech corpus. Zenodo, March 2021, 10.5281/zenodo.4626539

[36] Păiș, V. (2020). Multiple annotation pipelines inside the relate platform. In The 15th International Conference on Linguistic Resources and Tools for Natural Language Pro-cessing, pages 65–75.

[37] Păiș, V., Ion, R., and Tufiș, D. (2020). A processing platform relating data and tools for Romanian language. In Proceedings of the 1st International Workshop on Language Technology Platforms, pages 81–88, Marseille, France. European Language Resources Association.

[38] Păiș, V., Tufiș, D., and Ion, R. (2019). Integration of romanian nlp tools into the relate platform. In International Conference on Linguistic Resources and Tools for Natural Language Processing.

[39] Păiș V., Ion, R., Avram, A.-M., Irimia, E., Mititelu, V. B., and Mitrofan, M. (2021). Human-machine interaction speech corpus from the robin project. In 2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), pages 91–96.

[40] Pinnis, M., Ion, R., Ştefănescu, D., Su, F., Skadiņa, I., Vasiļjevs, A. Babych, B.(2012). Toolkit for Multi-Level Alignment and Information Extraction from Comparable Corpora. In Proceedings of ACL 2012, System Demonstrations Track, Jeju Island, Republic of Korea, July 8-14, 2012

[41] Pinnis, M., Ljubešić, N., Ştefănescu, D., Skadiņa, I., Tadić, M., Gornostay, T. (2012). Terminology Extraction and Mapping Tools for Under-Resourced Languages, in Proceedings of the 10th Terminology and Knowledge Engineering Conference (TKE 2012), Madrid, Spain.

[42] Skadiņa, I., Vasiļjevs, A., Skadiņš, R., Gaizauskas, R., Tufiş, D. , Gornostay, T. (2010). Analysis and Evaluation of Comparable Corpora for Under Resourced Areas of Machine Translation. In Proceedings of the 3rd Workshop on Building and Using Comparable Corpora" (BUCC10) at the 7th Language Resources and Evaluation Conference (LREC 2010), pp. 6-14, Valletta, Malta, May 2010.

[43] Skadiņa, I., Aker, A., Giouli, V., Tufiş, D., Gaizauskas, R., Mieiriņa, M., Mastropavlos, N. (2010). A Collection of Comparable Corpora for Under-resourced Languages. In Inguna Skadiņa and Andrejs Vasiļjevs (eds.), Frontiers in Artificial Intelligence and Applications, volume 219: Human Language Technologies – The Baltic Perspective – Proceedings of the Fourth International Conference Baltic (HLT 2010), pp. 161-168, IOS Press, Riga, Latvia, October 2010. ISBN: 978-1-60750-640-9.

[44] Skadiņa, I., Aker, A., Mastropavlos, N., Su, F., Tufiș, D., Verlic, M., Vasiļjevs, A., Babych, B., Clough, P., Gaizauskas, R., Glaros, N., Lestari Paramita, M. Pinnis, M.(2012). Collecting and Using Comparable Corpora for Statistical Machine Translation. In Nicoletta Calzolari, Khalid Choukri, Thierry Declerck, Mehmet Uğur Doğan, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis (eds.), Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), pp. 438-445, May 23-25, 2012, Istanbul, Turkey. ISBN: 978-2-9517408-7-7

[45] Stan, A., Yamagishi, J., King, S., and Aylett, M. (2011). The romanian speech synthesis (rss) corpus: Building a high quality hmm-based speech synthesis system using a high sampling rate. Speech Communication, 53(3):442–450.

[46] Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufis, D., & Varga, D. (2006). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In Proceedings of the 5th LREC Conference, Genoa, Italy, 22-28 May, 2006, pp.2142-2147, ISBN 2-9517408-2-4, EAN 9782951740822, arXiv preprint cs/0609058.

[47] Ştefănescu, D. (2012). Extracting Parallel Terminology from Comparable Corpora, In Mihai Alex Moruz, Dan Cristea, Dan Tufiş, Adrian Iftene, Horia-Nicolai Teodorescu (eds.) Proceedings of the 8th International Conference "Linguistic Resources and Tools for Processing of The Romanian Language", pp. 181-188, April 26-27, 2012.

[48] Ştefănescu, D., Ion, R., Hunsicker, S. (2012). Hybrid Parallel Sentence Mining from Comparable Corpora. In Proceedings of the 16th Conference of the European Association for Machine Translation (EAMT 2012), pp. 137—144, Trento, Italy, May 28-30, 2012

[49] Ştefănescu, D. (2012). Mining for Term Translations in Comparable Corpora, in Proceedings of the 5th Workshop on Building and Using Comparable Corpora (BUCC 2012), Istanbul, Turkey

[50] Toma, S.-A., Stan, A., Pura, M.-L., and Bârsan, T. (2017). MaRePhoR- An open access machine-readable phonetic dictionary for romanian. In 2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD), pages 1–6. IEEE.

[51] Tufiș, D. and Cristea, D. and Stamou, S. (2004). BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. In Romanian Journal on Information Science and Technology, Special Issue on BalkaNet. (ed. Tufiș, Dan). vol. 7, no. 2-3, pp. 9-34, 2004

[52] Tufiș, D. (2012). Finding Translation Examples for Under-Resourced Language Pairs or for Narrow Domains; the Case for Machine Translation. In Computer Science Journal of Moldova, Academy of Sciences of Moldova, Institute of Mathematics and Computer Science, ISSN 1561-4042, vol.20, no.2(59), 2012, pp. 1-19.

[53] Tufiș, D., Dumitrescu, D., Ș. (2012). Cascaded Phrase-Based Statistical Machine Translation Systems. In Proceedings of the 16th EAMT Conference, Trento, Italy, pages 129-136

[54] Tufiș, D. and Cristea, D. (2017). An outlook over corola: The reference corpus of contemporary written and spoken corpus. In Proceedings of SpeD conference, (invited talk), Bucharest, Romania.

[55] Tufiș, D., Barbu Mititelu, V., Irimia, E., Mitrofan, M., Ion, R., and George, C. (2019). Making pepper understand and respond in romanian. In the 22nd International Conference on Control Systems and Computer Science.

[56] Tufiș, D., Barbu Mititelu, V., Irimia, E., Păiș, V., Ion, R., Diewald, N., Mitrofan, M., Onofrei, M. (2019). Little Strokes Fell Great Oaks. Creating Corola, The Reference Corpus of Contemporary Romanian. Revue roumaine de linguistique, In Revue roumaine de linguistique, No./Issue 3, 2019, pages 227-240.

[57] Tufiș, D. Mitrofan, M., Păiș, V., Ion, R., Coman, A. (2020) Collection and Annotation of the Romanian Legal Corpus. In Proceedings of The 12th Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, pp. 2766-2770, May 2020

[58] Váradi, T., Koeva, S., Yamalov, M., Tadić, M., Sass, B., Nitoń, B., Ogrodniczuk, M., Pęzik, P., Barbu Mititelu, V., Ion, R., Irimia, E., Mitrofan, M., Păiș, V., Tufiș, D., Garabík, R., Krek, S., Repar, A., Rihtar, M., and Brank, J. (2020). The MARCELL Legislative Corpus. In Proceedings of The 12th Language Resources and Evaluation Conference. European Language Resources Association, Marseille, France, pp. 3754-3761, May 2020

[59] L. A. Zadeh, D. Tufiş, F. Filip, I. Dziţac (eds), From Natural Language to Soft Computing: New Paradigms in Artificial Intelligence, Ed. Acad. Române, 2008, 226 pages, ISBN: 978-973-27-1678-6;

*Cite this paper as:*
Tufiş D. (2022). Romanian Language Technology – a view from an academic perspective, *International Journal of Computers Communications & Control*, 17(1), 4641, 2022.
https://doi.org/10.15837/ijccc.2022.1.4641