CURSUS

Științele limbajului

Roxana Patraș
Loredana Cuzmici
(editors)

Tools, Methods, and Solutions
for the Exploration of Romanian Corpora

INSTITUTUL EUROPEAN

Roxana Patraş, Loredana Cuzmici (editors), *Tools, Methods, and Solutions for the Exploration of Romanian Corpora*

INSTITUTUL EUROPEAN
Iaşi, Str. Băluşescu nr. 2, et. I, O P 6, CP 1309
euroedit@hotmail.com.; www.euroinst.ro

Descrierea CIP a Bibliotecii Naţionale a României

Tools, methods, and solutions for the exploration of Romanian corpora /
ed.: Roxana Patraş, Loredana Cuzmici. - Iaşi : Institutul European, 2024
Conţine bibliografie
ISBN 978-606-24-0395-9

I. Patraş, Roxana (ed.)
II. Cuzmici, Loredana (ed.)

811.135.1

PRINTED IN ROMANIA

Roxana Patraș, Loredana Cuzmici
(editors)

TOOLS, METHODS, AND SOLUTIONS FOR THE EXPLORATION OF ROMANIAN CORPORA

INSTITUTUL EUROPEAN
2024

This work was supported by a grant of the Romanian Ministry of Education and Research, CNCS–UEFISCDI, project number PN-III-P1-1.1-TE-2019-0127, within PNCDI III.

The revision and editing of this collection was performed by BolderWords translation & revision and was supported by the Altissia Chair in Digital Cultures and Ethics held by Chris Tanasescu at Université catholique de Louvain.

Table of Contents

RELATE: a modern processing platform for Romanian language

Vasile PĂIȘ, Radu ION,
Andrei-Marius AVRAM,
Maria MITROFAN, Dan TUFIȘ

1. Introduction

Language Technology (LT) platforms provide services for the analysis and production of written or spoken language, making use of artificial intelligence (AI) methods to do so. Furthermore, these platforms have been developed either to showcase new technologies or to function as powerful tools for processing large amounts of data, in the form of offline corpora or online requests.

One of the principal goals of the 1st International Workshop on Language Technology Platforms (IWLTP 2020) was to address fragmentation in the language technology landscape. As the workshop organizers noted, instead of competing with one another, platforms should be designed to be interoperable and to interact with each other to create synergies towards a productive LT ecosystem (Rehm *et al.* 2020a). Interoperability is usually achieved by providing input and output in standardized formats, specific to the tasks at hand. Interaction between systems involves making the functionalities of one system available for use by another system.

In this paper, we present the RELATE platform (Păiș *et al.* 2020), a modular state-of-the-art platform that is used for processing the Romanian language developed at the Research Institute for Artificial Intelligence "Mihai Drăgănescu" of the Romanian Academy (for short RACAI in English, or ICIA in the Romanian language). Integrating resources and technologies developed in our institute as well as those developed by partner institutions, RELATE is actively being used in multiple national and international research projects. From the beginning, it was

designed to use standardized file formats, thus ensuring interoperability with other language processing systems. Internal functions are available as JSON REST web services, thus allowing for interaction with other systems. In the Representational State Transfer (REST) architectural style, data and functionality are considered resources and are accessed using Uniform Resource Identifiers (URIs).

This paper is organized as follows: section 2 presents related work in the field of LT platforms; section 3 presents the history of how the RELATE platform developed; section 4 introduces RELATE's architecture; section 5 describes the platform's components; section 6 presents several scenarios in which the platform might be further used; and we draw conclusions in section 7.

2. Related work

META-SHARE[1] (Federmann *et al.* 2012), CLARIN[2] and ELRC-Share[3] (European Language Resource Coordination Share) are publicly available European websites for research and development in the field of natural language processing, allowing access to language resources. Both ELRC-Share and META-SHARE offer advanced search facilities through which one can easily find various language tools and corpora (e.g. text corpora, annotated corpora, audio corpora) in any (European and other) language. Complex processing pipelines, like NLP-Cube (Boroș *et al.* 2018a) and TTL are able to perform tokenization, lemmatization, POS tagging, chunking and dependency parsing, but they require programming knowledge in order to integrate these functions into an application. A more recently created web service, TEPROLIN, integrates multiple tools — including NLP-Cube, TTL and solutions for named entity recognition (Păiș 2019) and biomedical named entity recognition — into an easy-to-use pipeline. However, this web service still requires users to have programming knowledge in order to engage in web service communication.

GATE (Cunningham *et al.* 2002) and TextFlows (Perovšek *et al.* 2016) aim to make the composition of the language processing chains more user-friendly. The

[1] http://www.meta-share.org.
[2] https://www.clarin.eu.
[3] https://elrc-share.eu/.

graphical interfaces thus allow text-processing widgets to be dragged and dropped into a graphical processing workflow. However, text processing output is not enhanced with specialized visualization tools that would allow access to the computational resources used for annotation. Moreover, they only focus on purely textual content, while multimodal content (such as a combination of text and audio) is not handled within the platforms.

CoBiLiRo (Cristea *et al.* 2020) is a storage platform for multimodal (text and audio) corpora. It was developed in the context of the RETEROM[4] project for storing Romanian speech corpora in a way that was suitable for the project's purposes. It allows a large number of metadata fields to be defined, but it does not enable any complex language processing tasks.

RELATE specifically serves to carry out automatic text processing, with annotations at multiple levels, as well as annotation visualization and expansion into the corresponding linguistic computational resources. The platform focuses on the interactive user experience, through web-based interfaces. Unlike other platforms, such as the one designed by Che *et al.* (2010), RELATE is not primarily focused on exposing APIs, even though API access to platform functionalities is available. Text-processing APIs are used internally and are accessible by platform components. Most of these APIs can also be invoked externally.

RELATE makes use of a common internal format, comprising multiple files (text, standoff metadata, CoNLL-U Plus[5] annotations). The processing workflow is built by adding individual tasks in the graphical user interface. Each task is able to both use and also produce data according to the common internal format. For this reason, no workflow editor, such as the one used in TextFlows (Perovšek *et al.* 2016), is currently available.

Coleman *et al.* (2020) describe a platform for integrating multiple Machine Translation (MT) models. In RELATE, we provide those MT capabilities that are related to the Romanian language (currently supporting Romanian-English and English-Romanian translation only). Furthermore, in RELATE, translated content

[4] https://racai.ro/p/reterom/.
[5] https://universaldependencies.org/ext-format.html.

can be used as input for further language processing tasks. Rebai *et al.* (2020) describe a platform that integrates a voice assistant for improving efficiency and productivity in business. In the case of RELATE, we only integrate existing Romanian (and English) Automatic Speech Recognition (ASR) and Text-to-Speech (TTS) systems while still focusing on the textual components of a corpus. Thus, ASR output can be used as input to further language processing tasks.

Rehm *et al.* (2020b) acknowledge that a large number of AI platforms are currently being developed, both nationally—supported through state funding programmes—and internationally, supported by the European Union. The authors further recognize the enormous fragmentation of the European AI and LT landscapes and believe that modern platforms should facilitate exchange information, data and services, in order to enable interoperability. We agree with this assessment and, in the RELATE platform, we therefore sought to make use of standardized formats as well as decoupling functionalities into components, which could be invoked externally if needed. In addition, similar to the AI4EU[6] and European Language Grid (ELG)[7] platforms, RELATE is able to integrate Docker containers for language processing tools. Such integration is not required, however, and only a small number of the available components are therefore integrated as containers.

3. Evolution of the RELATE platform

The RELATE platform was developed in the context of a number of national and international research projects and evolved according to the needs of the activities for which it was ultimately used. This section describes its evolution throughout these projects, from the first implementation to the current version.

RELATE platform development (Păiș *et al.* 2019) started in the context of the national RETEROM project, which began in 2018. One of the project's main goals, linked to the sub-project TEPROLIN, was to develop and integrate state-of-the-art technologies for Romanian natural language processing, such as those devoted to tokenization, part-of-speech tagging, dependency parsing, phonetic annotations, and

[6] https://www.ai4eu.eu/.
[7] https://www.european-language-grid.eu/.

named entity recognition. The primary result was the TEPROLIN web service (Ion 2018) which allows invocation that uses a raw text document and produces different levels of annotation (based on specified parameters) encoded using a custom JSON format.[8]

The first implementation of the RELATE platform allowed for the management (upload, download, editing) of large corpora and parallel processing using the TEPROLIN service. For processing purposes, we implemented a task management engine, which allowed documents to be distributed across any number of TEPROLIN processes that begin on the same server or over the network. As a result, the processing speed can be increased when needed, because the number of processing nodes can be adjusted dynamically. Furthermore, the platform is able to convert the custom JSON encoding associated with TEPROLIN into a more standard CoNLL-U format[9] (also used by the Universal Dependencies project[10]) or its extension CoNLL-U Plus,[11] when appropriate.

To allow human users to manually explore annotation results, we developed graphical representations, such as visualization in different column-based and JSON formats, tree-based representations for dependency parsing, and the highlighting of recognized named entities. In order to improve the user experience, we also included different interface elements that allow the Representative Corpus of Contemporary Romanian Language (CoRoLa) to be queried (Tufiș *et al.* 2019b). RELATE includes links to the main query interface of CoRoLa's text component, which are accessible through the KorAP corpus analysis platform (Banski *et al.* 2012), and to the speech component, which allows users to search audio files (Boroș *et al.* 2018b) and listen to words being pronounced by speakers of Romanian. Different visualizations also make use of word embeddings (Bojanowski *et al.* 2017) computed on the CoRoLa corpus (Păiș and Tufiș 2018) to suggest words that appear in similar contexts.

We integrated the Romanian WordNet (Tufiș and Mititelu 2015) into the RELATE platform to allow words to be searched online. Furthermore, we exploited

[8] http://relate.racai.ro/?path=teprolin/doc_dev.
[9] https://universaldependencies.org/format.html.
[10] https://universaldependencies.org/.
[11] https://universaldependencies.org/ext-format.html.

the alignment between the Romanian and English (Miller 1995) wordnets in order to provide aligned queries based on *synset* identifiers (in a WordNet, synsets represent sets of synonyms that share a common meaning).

A textual translation component—which was developed by TILDE with the involvement of ICIA, in the context of the project entitled "CEF Automated Translation toolkit for the Rotating Presidency of the Council of the EU"—was integrated into RELATE through the TILDE Machine Translation API.[12] This component allows users to translate documents directly within the platform. After a successful translation from English to Romanian, the resulting document can be analyzed using the platform's text analysis functionality.

As part of the "Multilingual Resources for CEF.AT in the legal domain" project (MARCELL),[13] a terminology annotation tool (Coman *et al.* 2019a) was developed, and we subsequently integrated it into the RELATE platform to facilitate the identification of terms available in the EuroVoc[14] and IATE[15] (Interactive Terminology for Europe) terminological databases. This annotation tool can be used following a lemmatization and part-of-speech tagging operation.

The ASR system (Avram *et al.* 2020b) that resulted from the national project ROBIN[16] and was initially developed for human-robot interaction (Ion *et al.* 2020; Tufiș *et al.* 2019a) was also integrated into the RELATE platform. This system allows text to be extracted from speech recorded directly in the platform or from uploaded audio files and then to be analyzed with the platform's processing tools. Several corrections on the recognized text (such as basic comma restoration and truecasing) can be performed before the actual analysis is carried out (Avram *et al.* 2020a).

We further exploited the availability of ASR and text translation features in order to provide speech-to-speech translation functionality for Romanian-English and English-Romanian. This brought together different platform modules and required the integration of additional text-to-speech components. To synthesize the

[12] https://www.tilde.com/developers/machine-translation-api.
[13] https://marcell-project.eu/.
[14] https://eur-lex.europa.eu/browse/eurovoc.html.
[15] https://iate.europa.eu/home.
[16] http://aimas.cs.pub.ro/robin/en/.

Romanian speech, we integrated two models into our pipeline: Romanian TTS (developed by Stan *et al.* (2011)) and RACAI SSLA (developed by Boroș *et al.* (2018b)), both of which are based on Hidden Markov Models (HMM) to compute the most probable sequence of spectrograms.

The "Curated Multilingual Language Resources for CEF.AT" (CURLICAT) project, launched in 2021, requires large-scale anonymization of the collected text corpus. This process was further integrated into RELATE, making use of the parallelization mechanism that was already available for different text-processing tasks.

One of our institute's research objectives, which aligns with the COST Action "Nexus Linguarum", is to create resources that are specific to Linguistic Linked Open Data specifications (Barbu Mititelu *et al.* 2020). Our involvement in the COST Action provided an opportunity to develop post-processing capabilities within RELATE, allowing RDF exports for different annotation levels.

The "European Language Equality" (ELE) project,[17] initiated in 2021, has as its primary goal to develop a sustainable evidence-based strategic research agenda and a roadmap that sets out actions, processes, tools, and actors that are integral to achieving—through the effective use of language technologies—digital equality among all of the languages (official or otherwise) used within the European Union. Additionally, the research agenda encompasses the assessment of the current state of language technologies and language equality within the EU. In turn, this led to an extension of RELATE, allowing a number of similar text-processing pipelines (Păiș 2020) to be executed on the same corpus in order to compare the results of each one (with or without the presence of a gold-standard annotated corpus).

4. Platform architecture

We adopted a modular approach to organize the RELATE platform. Each component was implemented independently of the others and provides a JSON descriptor that exposes the corresponding menu entries and functionalities at different user access levels. Moreover, components are service oriented. They can

[17] http://www.european-language-equality.eu.

consume web services (either SOAP or REST) and expose additional HTTP REST services. In turn, the exposed services can be consumed by the component's graphical user interface, through AJAX calls to the REST endpoints.

This service-oriented architecture allows for easy reuse of component-provided functionality in other parts of the platform or even from outside the platform. By consuming services exposed by processing components, RELATE can deploy these functions on different computing nodes, thus enabling high-speed parallel processing across interconnected servers. The processing parallelization was used for large corpora annotation in our institute, engaging multiple servers available in the same local area network (LAN). However, since a processing operation takes more time than network communication, it is possible to use remote hardware resources (across the internet) to further increase the processing capabilities available to the platform.

Figure 1 presents RELATE's general platform architecture, identifying its various components. The front-end part of the platform is responsible for user interaction. It provides visualizations of the different components. It is implemented using modern front-end technologies and languages such as HTML5, CSS, JavaScript. Communication with the platform's back end occurs through page loading or AJAX calls when possible. The use of AJAX enhances the user experience by reducing loading times. Furthermore, exposing different APIs in the form of HTTP REST services enables direct usage from outside the platform if needed.

The RELATE back end is responsible for communicating with different language-specific components. The back end itself is written in PHP, while the components can be written in any language, as long as a communication mechanism is available. The platform currently has components written in C++, Java and Python. The preferred means of communication is HTTP REST API calls. Nevertheless, the platform also integrated components by invoking new processes with command line arguments.

5. Primary platform components

This section will present the primary platform components (as indicated in *Figure 1*), addressing topics such as the platform's functions and implementation

details. Only components that have undergone recent developments are included. Older components, such as the WordNet and CoRoLa query interfaces are not discussed in detail. Additionally, some of the more minor components will be discussed in Section 6.

5.1. Corpus annotation

TEPROLIN (Ion 2018) is a text pre-processing platform, implemented as a web service. It was developed as part of the ReTeRom project,[18] with the specific goal of making it easy to bring together diverse text pre-processing tools that are written in a host of programming languages. TEPROLIN is written in Python 3 and requires each participating text pre-processing application to implement the TEPROLIN API class.[19] Ensuring that the following conditions are met leads to best results in TEPROLIN:

- External resources should always be loaded at the same time as the object is created; this ensures that, with large resource files, the loading time does not interfere with the run time.
- Text pre-processing applications should be resident processes on the machine that is running TEPROLIN, and the I/O with TEPROLIN should be made with available IPC mechanisms, such as TCP sockets.
- If text pre-processing applications are designed as web services, no resource loading is necessary and the _runApp method is the only one that needs to be implemented.

TEPROLIN can be used either as a stand-alone Python 3 object or through the RELATE platform, in demo[20] or production mode (for annotating large corpora). Each processing operation (e.g. POS tagging, dependency parsing) can be configured individually. If desired, only a subset of operations can be requested to be performed. In this case, TEPROLIN will auto-configure to execute all the other required operations. For instance, if the user requires POS tagging, TEPROLIN will automatically conduct sentence splitting and tokenization beforehand.

[18] https://www.racai.ro/p/reterom/.
[19] https://github.com/racai-ai/TEPROLIN/blob/master/TeproApi.py.
[20] https://relate.racai.ro/index.php?path=teprolin/custom.

Currently, TEPROLIN includes the following text pre-processing applications:

- NLP-Cube (Boroș *et al.* 2018a), UDPipe (Straka *et al.* 2016) and TTL (Ion 2007)
  - All three carry out sentence splitting, POS tagging and lemmatization functions.
  - NLP-Cube and UDPipe perform dependency parsing.
  - TTL performs noun phrase, verb phrase, prepositional phrase, adverbial and adjectival phrase non-recursive chunking.
- MLPLA (Boroș *et al.* 2018b)
  - This application performs word hyphenation, stressed accented syllable detection, and phonetic transcription.
- Various in-house tools
  - These were developed for Romanian text normalization, diacritic restoration, abbreviation and numeral expansion (i.e. automatically derive the full form of the abbreviation or the literal transcription of the numeral).
- BioNER (Mitrofan *et al.* 2018)
  - This application is based on a previous version of NLP-Cube adapted for NER processing.
- NER (Păiș 2019)
  - This application detects names of persons, organizations, locations and time expressions.

UDPipe[21] (Straka *et al.* 2016) is a trainable application that is used for the tokenization, tagging, lemmatization and dependency parsing of CoNLL-U files. Sentence splitting and tokenization are performed by checking each character in the input—using a single-layer bidirectional GRU network—to determine whether the character is the end of a token, the end of a sentence, or both. POS tagging and lemmatization use an averaged perceptron with morphosyntactic features (e.g. word affixes paired with current POS tags), while the dependency parser uses a simple

[21] https://github.com/ufal/udpipe.

feed-forward neural network to select the optimal transition indicated by the chosen transition-based dependency parser. Instead of TEPROLIN, UDPipe can be used as a stand-alone text pre-processing pipeline in RELATE, within its production mode (i.e. using multiple processing threads to annotate large corpora).

BioNER annotation is obtained through TEPROLIN, through requests for the biomedical-named-entity-recognition operation. It produces inside-outside-beginning annotations of disorders (DISO), anatomical parts (ANAT), medical procedures (PROC), drugs and other chemicals (CHEM). Mitrofan *et al.* (2018) describe the corpus that was used to train a previous version of NLP-Cube to recognize these specialized text spans.

Another type of NER annotation involves using a modified version of an older Stanford Named Entity Recognizer (Stanford NER) (Finkel *et al.* 2005) system, making use of CoRoLa pre-trained word embeddings (Păiș and Tufiș 2018), and computing representations of unknown words based on sub-word information (Bojanowski *et al.* 2017). The annotation is further enhanced through a rule-based method. Finally, results are attached to a tokenized document using Inside-Outside (IO) notation, which is then converted to an Inside-Outside-Beginning (IOB) format for compatibility with other annotations.

Besides TEPROLIN and UDPipe processing, pre-trained language models for other basic language resource kits (BLARK) are available for download from within the RELATE platform.[22] These models are trained on the same corpus, namely version 2.7 of the RoRefTrees treebank (RRT) (Barbu Mititelu, 2018), which is accessible through the Universal Dependencies[23] project. Models are offered for Stanza[24] (Qi *et al.* 2020), RNNTagger[25] (Schmid 2019), NLP-Cube[26] (Boroș *et al.* 2018a), UDPipe (Straka *et al.* 2016) and TreeTagger[27] (Schmid 1994).

[22] https://relate.racai.ro/index.php?path=pretrainedlm/list.
[23] https://universaldependencies.org/.
[24] https://stanfordnlp.github.io/stanza/index.html.
[25] https://www.cis.uni-muenchen.de/~schmid/tools/RNNTagger/.
[26] https://github.com/adobe/NLP-Cube.
[27] https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.

5.2. Terminology annotation

The terminology annotation tool available in RELATE was initially created for the purposes of the international project "Multilingual Resources for CEF.AT in the legal domain" (MARCELL).[28] This project produced a large comparable corpus (Váradi *et al.* 2020) of national legislation from the seven partners' countries, including Romania, enriched with EuroVoc and IATE terminology annotations. Furthermore, a classification of these legislative documents based on the top-level domains of the EuroVoc multilingual thesaurus was carried out.

For categorizing documents belonging to the Romanian-language sub-corpus (Tufiș *et al.* 2020), we developed a classifier based on the method proposed by Joulin *et al.* (2017), using the average of word embeddings over n-grams as features to a linear classifier. The actual implementation used the FastText[29] tool. We trained several models, using different word embeddings (CoRoLa, Wikipedia and CommonCrawl). We wanted to compare these models with the well-known JEX tool,[30] developed by the European Commission's Joint Research Centre (EC-JRC). JEX (Steinberger *et al.* 2012) makes use of the classification algorithm described in Pouliquen *et al.* (2006), based on a list of lemma frequencies from normalized text, and their weights, which are statistically related to each descriptor, mentioned in the paper as associates or as topic signatures. For Romanian, JEX was trained on over 25,000 documents from two parallel corpora: JRC-Acquis (Steinberger *et al.* 2006) and OPOCE (Publications Office of the European Union). By considering only the first 6 most probable descriptors, it obtained a precision of 45.55%, a recall of 50.43% and a F1 score of 47.84%.

For training and evaluating our EuroVoc models, we used the same approach followed by JEX, employing the same datasets and cross-validation splits. Our best model was obtained through the use of CoRoLa word embeddings and achieved a precision of 50.93%, a recall of 56.41% and a F1 score of 53.53%, thus presenting an increase of over 5% with regard to F1 score as compared to JEX. We further

[28] https://marcell-project.eu/.
[29] https://fasttext.cc/.
[30] https://ec.europa.eu/jrc/en/language-technologies/jrc-eurovoc-indexer.

evaluated the model on top-level domains and obtained an F1 score of 70.80%. The model is available for download and interrogation from RELATE.[31] In order to integrate this model into the platform, we employed a modified version of the FastText application, including an embedded web server, which allowed the model to be kept in-memory. This server is open-sourced.[32] Besides document classification, term annotation was realized through the use of a NER-based approach, as described by Coman *et al.* (2019b).

For the named entities belonging to the location class (LOC) of the LegalNERo corpus[33] (Păiș *et al.* 2021b), we carried out an enhancement that linked the named entities with the GeoNames database[34] by using the feature identifiers associated with each GeoNames feature. GeoNames is a geographical database that contains over 25 million geographical names, covering all countries, and consists of over 11 million unique placenames, 4.8 million populated places and 13 million alternate names. GeoNames is integrating geographical data such as elevation, population, names of places in various languages, and other data from various sources. To link the named entities with the GeoNames database, we created a new column in the conllup corpus files structure. This GeoNames column was placed last in the file and is, by default, filled with the "_" character. This column contains a value only when the named entity tag is of LOC type and when a perfect match is found within the GeoNames database. The actual match is done through a lookup algorithm written in the Python language. In this algorithm, the GeoNames database extract for the Romanian language is loaded into a *key:value* list, where the key is the standard location name and the value is the GeoNames location code (e.g. the code for the city of Suceava is 486885). In many cases, a given location can have several variations in the naming (whether on account of different spelling conventions or due to differences in how the places are known across and within different communities). GeoNames addresses this phenomenon by listing only one standard name (usually

[31] https://relate.racai.ro/index.php?path=eurovoc/classify.
[32] https://github.com/racai-ai/ServerFastText.
[33] https://doi.org/10.5281/zenodo.4772094.
[34] https://www.geonames.org/.

the location's official name) and including several secondary names alongside the GeoNames entry for that location. For each variation of this kind, the application has therefore created a different entry in the lookup dictionary, all of them pointing to the same location code value. The conllup file format is token centric, while certain locations have a multi-word format. Therefore, the first step in the lookup was actually to build the expressions based on the tokens (by finding consecutive LOC tokens). The actual lookup was performed for each named entity (a multi-token list). When a perfect match was found (partial sub-expression matches were ignored), all of the tokens involved in the match were flagged as a GeoNames entity. The corresponding location code was finally entered into the GeoNames column.

5.3 Automatic speech recognition

The RELATE platform contains two Romanian ASR models, which can be accessed using the graphical interface: RobinASR and RobinASRDev. The two models are based on the DeepSpeech2 architecture (Amodei *et al.* 2016) and, as compared with the original model, have a reduced number of parameters, which allows the smaller size of their training corpus to be accommodated. Each model is wrapped as a web service and deployed on a server with a GPU, allowing a low response time to the request on the front end to be achieved. Although the two variants are similar in architecture, however, they serve different purposes.

The former variant is a general ASR that was trained on a corpus containing 230 hours of transcribed audio and that obtained a 9.91 word-error rate (WER) when combined with a language model (Heafield 2011) used for textual correction. The latter variant is a specialized ASR. It fine-tunes the general model on the technical domain using the ROBIN Technical Acquisition Speech Corpus (ROBINTASC) corpus (Păiș *et al.* 2021a) and improving the model's performance by 16.4% on computer sales conversations. This variant also uses a language model that was trained on the general corpus together with the train ROBINTASC transcriptions, oversampled 10 times to increase n-grams frequencies.

5.4 Text and speech translation

The speech-to-speech translation (S2ST) system (Avram *et al.* 2021b) consists of four cascaded modules: (1) ASR, (2) Textual Correction (TC), (3) Machine Translation (MT) and (4) TTS. Each module contains one or several models that can be configured in a full Romanian-English or English-Romanian inference, as depicted in *Figure 2*.

The ASR module contains two models for Romanian—namely those presented in Subsection 5.3—and two models for English. The first variant for English (En DeepSpeech2) is also based on the DeepSpeech2 architecture and was trained on the LibriSpeech dataset (Panayotov *et al.* 2015), obtaining a 10.46 WER. The second variant for English (Mozilla DeepSpeech) is the 0.9.3 version of DeepSpeech (Hannun *et al.* 2014) offered by Mozilla Speech-To-Text engine,[35] and it obtained a 7.06% WER on LibriSpeech.

At this time, we offer only a single model in the TC module for Romanian: Robin Correction. Its role is (1) to truecase the transcribed speech using a list of predefined named entities, (2) to replace unknown words with words from an existing vocabulary and (3) to restore hyphens using uni-gram and bi-gram statistics.

The Romanian-English (Presidency Ro-En) and English-Romanian (Presidency En-Ro) models used by the MT module are those that were developed in the context of the "CEF Automated Translation toolkit for the Rotating Presidency of the Council of the EU" project, described in section 3.

The TTS module contains a speech synthesizer model for English and two for Romanian. The English version—Mozilla En TTS—is based on Tacotron2 with Dynamic Convolution Attention (Battenberg *et al.* 2020) offered by Mozilla Text-to-Speech.[36] Their evaluation outlined a median opinion score (MOS) of 4.31 ± 0.06 in a 95% confidence interval. The Romanian variants—Romanian TTS and RACAI SSLA—use HMM on their architecture, with the former variant being slower, but producing a higher quality speech and the latter being faster but with a lower quality of the synthesis.

[35] https://github.com/mozilla/DeepSpeech/releases/tag/v0.9.3.
[36] https://github.com/mozilla/TTS.

6. Usage scenarios

Due to its complex nature and its large number of integrated components, the RELATE platform has been used for a variety of purposes. However, the most common ones are described in this section. Moreover, the description of the usage scenarios provides insights into possible interactions between the platform's components.

6.1 Large corpus processing

Large corpora can be uploaded to RELATE as archived zip files. Once they are uploaded, the platform will automatically schedule a task for extracting the archive. Inside the archive, the platform accepts raw text files (with .txt extensions) and standoff metadata. Metadata are usually specific to different research projects and can be used by different processing components to embed them into annotated documents.

Once the documents are available in the platform, different annotation tasks can be scheduled and executed. The tasks interface is presented in *Figure 4*. Furthermore, statistics on both raw text and annotated documents can be computed. The resulting annotated corpus can be archived and downloaded as another zip file. Statistics can be either visualized within the platform or downloaded as csv files. The entire flow is depicted in *Figure 3*.

6.2. Creation of human-annotated gold-standard text corpora

Creating a gold-standard corpus involves human annotators identifying spans of text with certain properties. For the creation of the LegalNERo[37] corpus (a gold-standard corpus for named entity recognition in the Romanian legal domain), we used the services of five human annotators who worked with RELATE and annotated spans of text with the corresponding named entities: legal reference, person, location, organization and time (Păiș *et al.* 2021b).

[37] https://doi.org/10.5281/zenodo.4772094.

For the purpose of manual annotation of text corpora, RELATE integrates the BRAT[38] annotation tool (Stenetorp *et al.* 2012). This allows the user to view one document at a time inside the BRAT component, use their mouse to select the desired text span and then associate a corresponding annotation type with it. The text spans with an associated start and end index as well as the annotation type are saved as standoff metadata files.

Since different corpus processing applications may require token-based annotations instead of span-based annotations, RELATE offers dedicated tasks for parallel conversion of the metadata to token-based annotations. This process must be executed after an initial tokenization of the data, possibly including other automatic annotations such as part-of-speech tagging and dependency parsing. The final format for the data will be CoNLL-U Plus files with the gold-standard annotations stored in a dedicated column, named "RELATE:NE."

Once the corpus has been annotated (with or without conversion to the token-based format), it becomes possible to execute additional tasks within RELATE. A dedicated task allows extraction of gazetteer resources, allowing the produced file to be downloaded and used in other applications that require such resources. In addition, the statistics component can be used to extract information about the annotated corpus. The processing flow is depicted in *Figure 5*, while an example of the manual annotation component used with BioNER entities is presented in *Figure 6*.

6.3. Creation of read-speech corpora

Read speech corpora represent invaluable resources for the creation of ASR and TTS systems. These systems can then be used as the building blocks for more advanced processing, like human-machine interaction and speech-to-speech translation. In the context of the ROBIN project, we were concerned with the performance of a low-latency ASR system used for human-robot interaction, in a micro-world scenario. For this purpose, we created the ROBIN Technical

[38] https://brat.nlplab.org/index.html.

Acquisition Speech Corpus (RTASC)[39] (Păiș *et al.* 2021a), using the audio recording features of the RELATE platform.

Even though RELATE was initially constructed to process text corpora, the recent addition of speech processing features has allowed it to handle bimodal (text and audio) corpora. For recording purposes, a dedicated component was built. It makes use of JavaScript and runs on the user's browser. It displays each sentence from the text component of the corpus and allows the user to record the associated speech. To improve the user experience, RELATE displays the sentence in a larger font and provides additional information, such as the current sentence number and the total number of sentences. The interface also provides a playback function and, if necessary, allows the user to delete a certain recording in order to record it again. The recording interface is shown in *Figure 8*.

Since the text component still plays an important role in bimodal corpora, all the text-processing capabilities can be used to enhance the final corpus. In this case, the annotation pipelines and the statistics generation features are available. The final corpus, comprised of raw text, annotated text and audio files, can be downloaded in an archived format from the platform. The entire processing chain is presented in *Figure 7*.

7. Conclusions

In this article, we described the RELATE platform's development and architecture, and we considered several of the most common usage scenarios. Currently, the platform provides different levels of processing for the Romanian language. Even though RELATE was initially developed for handling large text corpora, recent work has integrated speech capabilities, including speech-to-speech translation for Romanian-English and English-Romanian language pairs.

The platform is designed to be highly customizable and easily extensible thanks to the creation of new components. The use of standardized file formats ensures interoperability with other systems. Also, the extensive use of REST APIs allows

[39] https://doi.org/10.5281/zenodo.4626539.

interoperability with external applications. Finally, the platform is open-sourced and available free of charge on GitHub.[40] The current running version of RELATE[41] spans its processing capabilities across four servers hosted at ICIA. The servers contain both classical CPUs and a GPU resource, allowing different types of language processing tools to be used and guaranteeing high performance.

In the future, we plan to continue integrating new language processing tools into the RELATE platform, as they become available, while aiming to boost interoperability and interactions leading to a productive LT ecosystem (Rehm *et al.* 2020a), both nationally and internationally. One possible extension in this direction is the replacement of the current EuroVoc annotator with the Romanian model of PyEuroVoc (Avram *et al.* 2021a), a new legal document classification tool that works with 22 languages that is based on a more modern architecture—Bidirectional Encoder Representations from Transformers (BERT) (Devlin *et al.* 2019)—and that significantly improved the results of JEX by an average of 28% F1 score for all languages and 33.8% for Romanian.

[40] https://github.com/racai-ai/RELATE.
[41] https://relate.racai.ro.

Annex 1: Figures

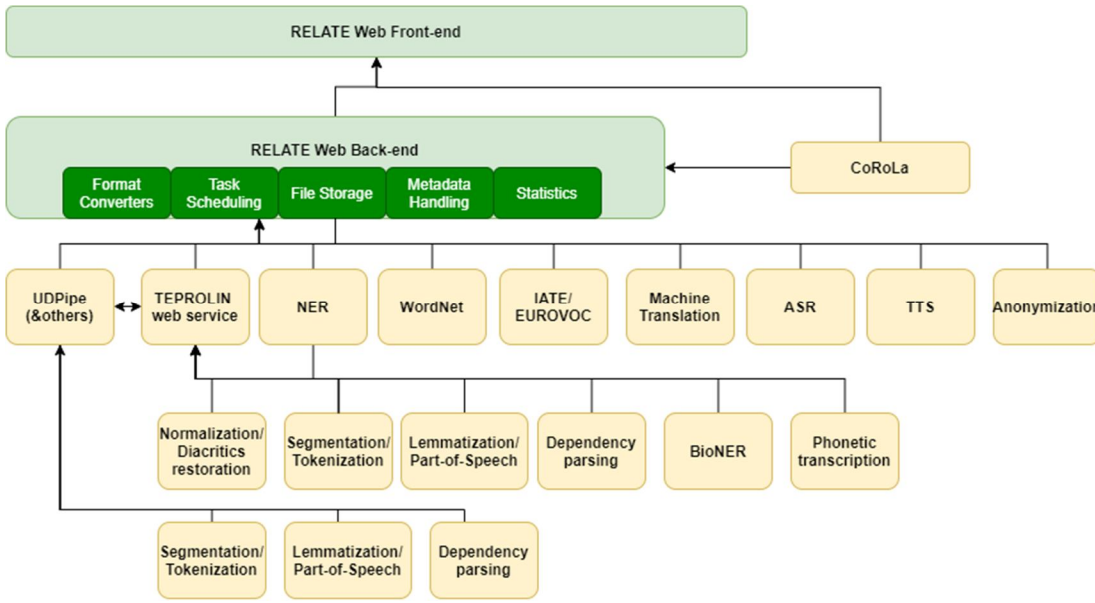

Figure 1: RELATE platform architecture

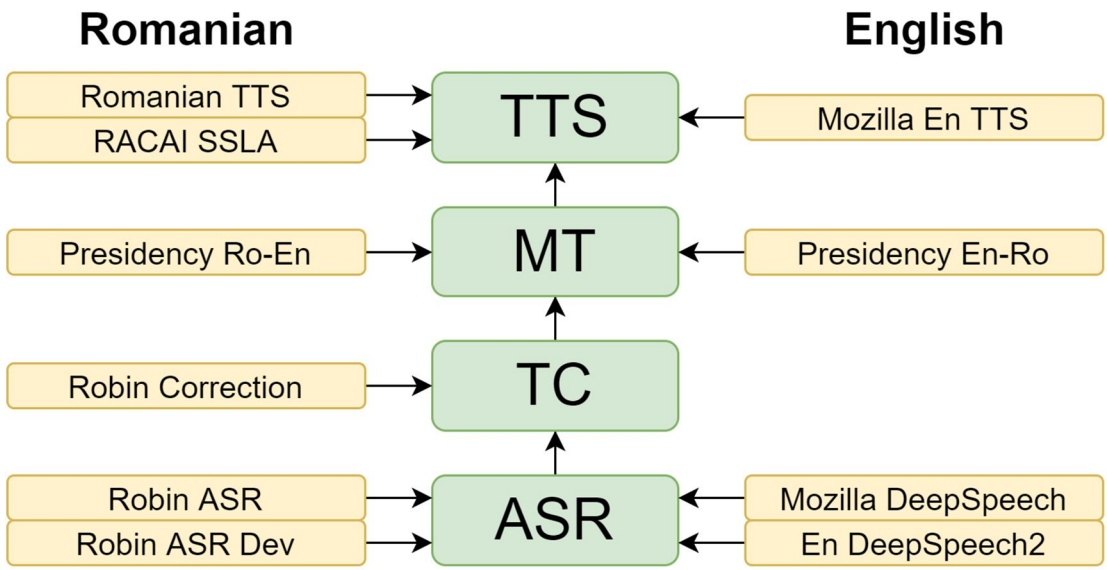

Figure 2: The S2ST cascaded architecture with the four modules and their models (Romanian in the left, English in the right).

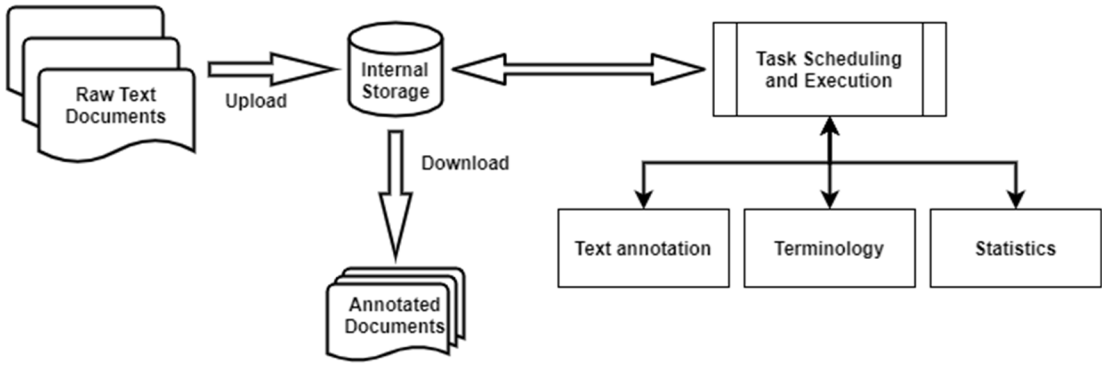

Figure 3: Corpus processing flow

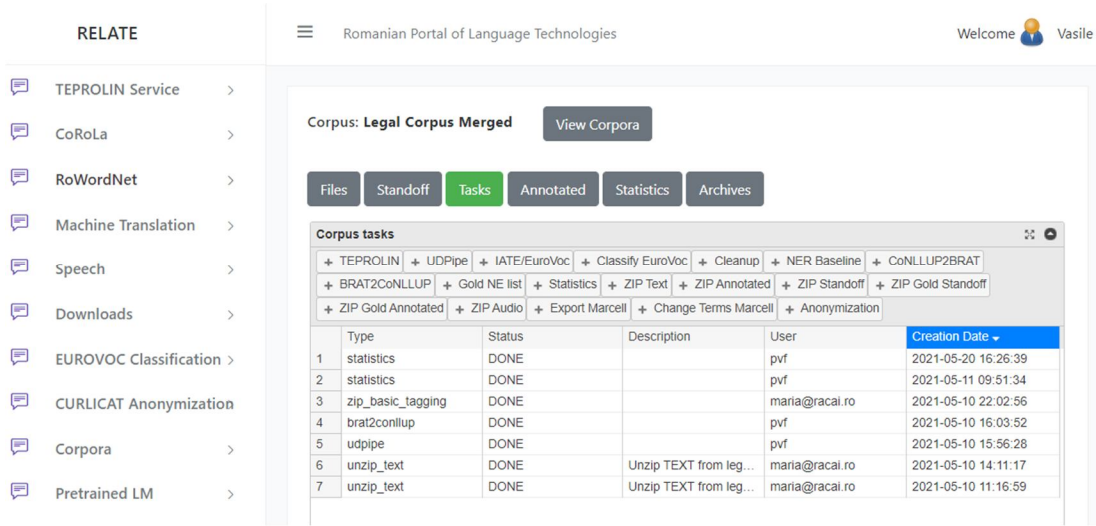

Figure 4: Tasks interface.

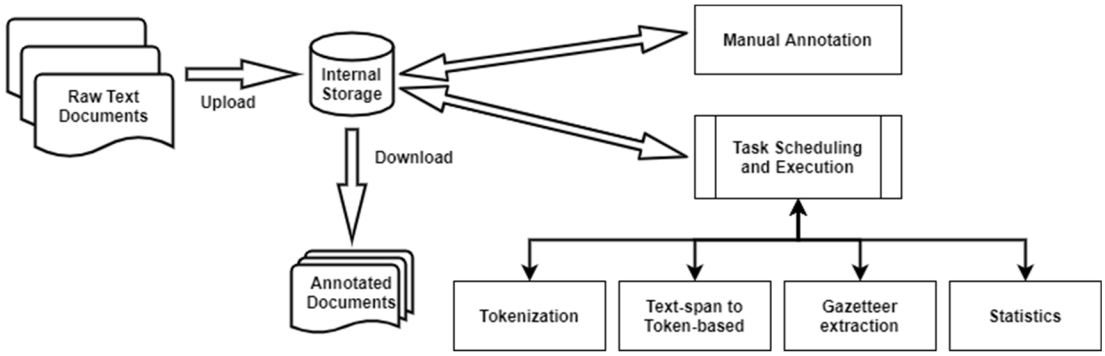

Figure 5: Manual corpus annotation processing flow

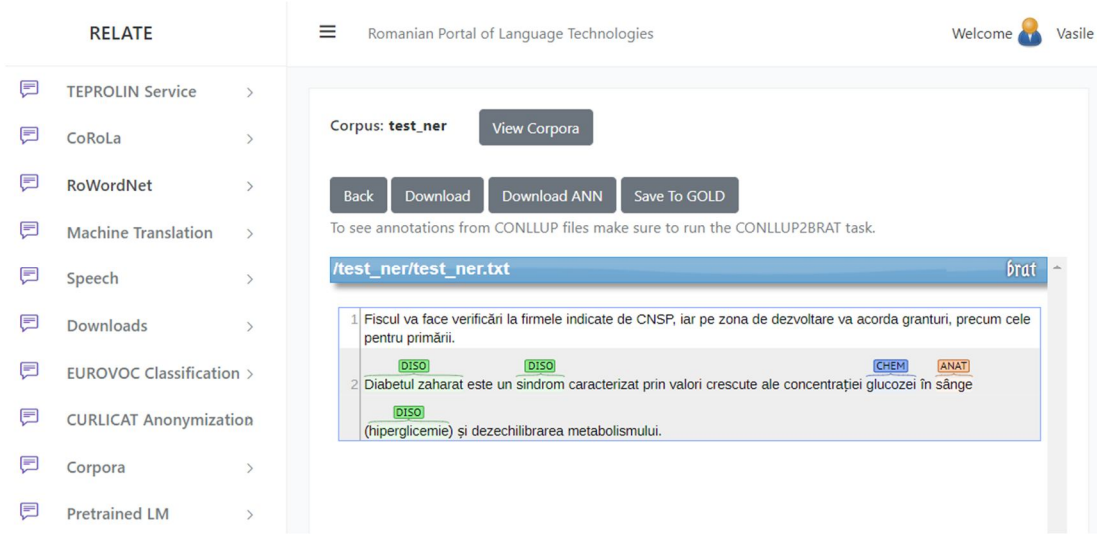

Figure 6: Manual corpus annotation GUI

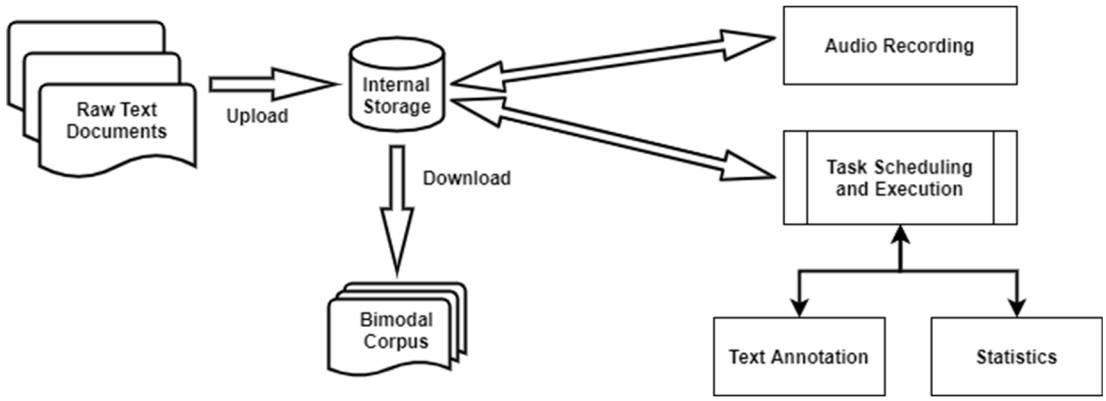

Figure 7: Creation of a bimodal corpus

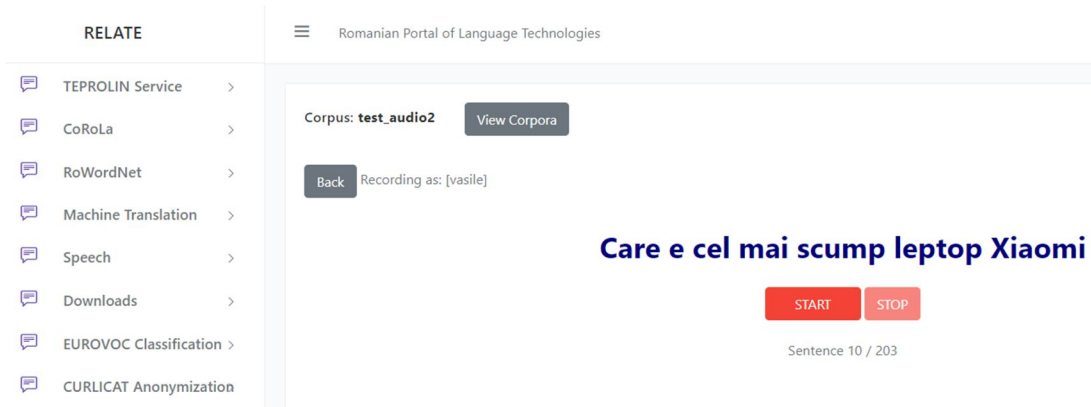

Figure 8: The interface used for recording read speech

References

Amodei, Dario, Sundaram Ananthanarayanan, Rishita Anubhai, Jingliang Bai, Eric Battenberg, Carl Case, Jared Casper, Bryan Catanzaro, Qiang Cheng, Guoliang Chen, *et al*. 2016. "Deep speech 2: End-to-end speech recognition in English and Mandarin." In *International Conference on Machine Learning*, 173–182. PMLR.

Avram, Andrei-Marius, Vasile Păiș, și Dan Tufiș. 2020a. "Romanian speech recognition experiments from the ROBIN project." In *The 15th International Conference on Linguistic Resources and Tools for Natural Language Processing*, 103–114.

Avram, Andrei-Marius, Vasile Păiș, and Dan Tufiș. 2020b. "Towards a Romanian end-to-end automatic speech recognition based on DeepSpeech2." *Proceedings of the Romanian Academy Series A* 21:395–402.

Avram, Andrei-Marius, Vasile Păiș, and Dan Tufiș. 2021a. "PyEuroVoc: A tool for multilingual legal document classification with EuroVoc descriptors." arXiv preprint arXiv:2108.01139.

Avram, Andrei-Marius, Vasile Păiș, and Dan Tufiș. 2021b. "A modular approach for Romanian-English speech translation." In *International Conference on Applications of Natural Language to Information Systems*, 57–63. Springer.

Banski, Piotr, Peter M. Fischer, Elena Frick, Erik Ketzan, Marc Kupietz, Carsten Schnober, Oliver Schonefeld, and Andreas Witt. 2012. "The new IDS corpus analysis platform: Challenges and prospects." In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 2905–2911. Istanbul, Turkey: European Language Resources Association (ELRA).

Battenberg, Eric, RJ Skerry-Ryan, Soroosh Mariooryad, Daisy Stanton, David Kao, Matt Shannon, and Tom Bagby. 2020. "Location-relative attention mechanisms for robust long-form speech synthesis." In *ICASSP 2020-2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6194–6198. IEEE.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. "Enriching word vectors with subword information." *Transactions of the Association for Computational Linguistics* 5:135–146.

Boroș, Tiberiu, Ștefan Daniel Dumitrescu, and Ruxandra Burtică. 2018a. "NLP-cube: End-to-end raw text processing with neural networks." In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 171–179. Brussels, Belgium: Association for Computational Linguistics.

Boroș, Tiberiu, Ștefan Daniel Dumitrescu, and Vasile Păiș. 2018b. "Tools and resources for Romanian text-to-speech and speech-to-text applications." In *Proceedings of the International Conference on Human-Computer Interaction (RoCHI)*, 46–53.

Che, Wanxiang, Zhenghua Li, and Ting Liu. 2010. "LTP: A Chinese language technology platform." In *Coling 2010: Demonstrations*, 13–16. Beijing, China: Coling 2010 Organizing Committee.

Coleman, Susie, Andrew Secker, Rachel Bawden, Barry Haddow, and Alexandra Birch. 2020. "Architecture of a scalable, secure and resilient translation platform for multilingual news media." In *Proceedings of the 1st International Workshop on Language Technology Platforms*, 16–21. Marseille, France: European Language Resources Association.

Coman, Andrei, Maria Mitrofan, and Dan Tufiș. 2019a. "Automatic identification and classification of legal terms in Romanian law texts." In *The 14th International Conference on Linguistic Resources and Tools for Natural Language Processing*, 39–49.

Coman, Andrei, Maria Mitrofan, and Dan Tufiș. 2019b. "Automatic identification and classification of legal terms in Romanian law texts." In *International Conference on Linguistic Resources and Tools for Natural Language Processing*.

Cristea, Dan, Ionuț Pistol, Șerban Boghiu, Anca-Diana Bibiri, Daniela Gîfu, Andrei Scutelnicu, Mihaela Onofrei, Diana Trandabăț, and George Bugeag. 2020. "CoBiLiRo: A research platform for bimodal corpora." In *Proceedings of the 1st International Workshop on Language Technology Platforms*, 22–27. Marseille, France: European Language Resources Association.

Cunningham, Hamish, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. "GATE: An architecture for development of robust HLT applications." In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 168–175. Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-training of deep bidirectional transformers for language understanding." In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 4171–4186. Minneapolis, Minnesota: Association for Computational Linguistics.

Federmann, Christian, Ioanna Giannopoulou, Christian Girardi, Olivier Hamon, Dimitris Mavroeidis, Salvatore Minutoli, and Marc Schroder. 2012. "META-SHARE v2: An open network of repositories for language resources including data and tools." In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, 3300–3303. Istanbul, Turkey: European Language Resources Association (ELRA).

Finkel, Jenny Rose, Trond Grenager, and Christopher Manning. 2005. "Incorporating non-local information into information extraction systems by Gibbs sampling." In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics (ACL'05)*, 363–370. Ann Arbor, Michigan: Association for Computational Linguistics.

Hannun, Awni, Carl Case, Jared Casper, Bryan Catanzaro, Greg Diamos, Erich Elsen, Ryan Prenger, Sanjeev Satheesh, Shubho Sengupta, Adam Coates, *et al*. 2014. "Deep speech: Scaling up end-to-end speech recognition." arXiv preprint arXiv:1412.5567.

Heafield, Kenneth. 2011. "KenLM: Faster and smaller language model queries." In *Proceedings of the Sixth Workshop on Statistical Machine Translation*, 187–197.

Ion, Radu. 2007. "Word sense disambiguation methods applied to English and Romanian." PhD thesis, Romanian Academy in Romanian.

Ion, Radu. 2018. "TEPROLIN: An extensible, online text preprocessing platform for Romanian." In *The 13th International Conference on Linguistic Resources and Tools for Natural Language Processing*, 69–76.

Ion, Radu, Valentin Gabriel Badea, George Cioroiu, Verginica Barbu Mititelu, Elena Irimia, Maria Mitrofan, and Dan Tufiș. 2020. "A dialog manager for micro-worlds." *Studies in Informatics and Control* 29(4):411–420.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. 2017. "Bag of tricks for efficient text classification." In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, 427–431. Valencia, Spain: Association for Computational Linguistics.

Miller, George A. 1995. "WordNet: A Lexical Database for English." *Commun. ACM* 38 (11): 39–41.

Mitrofan, Maria, Verginica Barbu Mititelu, and Grigorina Mitrofan. 2018. "Towards the Construction of a Gold Standard Biomedical Corpus for the Romanian Language." *Data* 3 (4): 12.

Păiș, Vasile. 2019. "Contributions to Semantic Processing of Texts; Identification of Entities and Relations Between Textual Units; Case Study on Romanian Language." PhD thesis, Romanian Academy.

Păiș, Vasile. 2020. "Multiple Annotation Pipelines Inside the RELATE Platform." In *The 15th International Conference on Linguistic Resources and Tools for Natural Language Processing*, 65–75.

Păiș, Vasile, Radu Ion, Verginica Barbu Mititelu, Elena Irimia, Maria Mitrofan, and Andrei-Marius Avram. 2021a. "ROBIN Technical Acquisition Speech Corpus."

Păiș, Vasile, Maria Mitrofan, Carol Luca Gasan, Alexandru Ianov, Corvin Ghiță, Vlad Silviu Coneschi, and Andrei Onuț. 2021b. "Romanian Named Entity Recognition in the Legal Domain (LegalNERo)."

Păiș, Vasile, Dan Tufiș. 2018. "Computing Distributed Representations of Words Using the CoRoLa Corpus." In *Proceedings of the Romanian Academy, Series A* 19 (2): 403–409.

Păiș, Vasile, Dan Tufiș, Radu Ion. 2019. "Integration of Romanian NLP Tools into the RELATE Platform." In *The 14th International Conference on Linguistic Resources and Tools for Natural Language Processing*, 181–192.

Păiș, Vasile, Dan Tufiș, Radu Ion. 2020. "A Processing Platform Relating Data and Tools for Romanian Language." In *Proceedings of the 12th Language Resources and Evaluation Conference*, 81–88. Marseille, France: European Language Resources Association.

Panayotov, Vassil, Guoguo Chen, Daniel Povey, and Sanjeev Khudanpur. 2015. "LibriSpeech: An ASR Corpus Based on Public Domain Audio Books." In *2015 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 5206–5210. IEEE.

Perovšek, Matic, Janez Kranjc, Tomaž Erjavec, Bojan Cestnik, and Nada Lavrač. 2016. "TextFlows: A Visual Programming Platform for Text Mining and Natural Language Processing." *Science of Computer Programming* 121: 128–152. Special Issue on Knowledge-based Software Engineering.

Pouliquen, B., R. Steinberger, and C. Ignat. 2006. "Automatic Annotation of Multilingual Text Collections with a Conceptual Thesaurus." *ArXiv*, abs/cs/0609059.

Qi, Peng, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. "Stanza: A Python Natural Language Processing Toolkit for Many Human Languages." In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 101–108. Online: Association for Computational Linguistics.

Rebai, Ilyes, Sami Benhamiche, Kate Thompson, Zied Sellami, Damien Laine, and Jean-Pierre Lorre. 2020. "LinTO Platform: A Smart Open Voice Assistant for Business Environments." In *Proceedings of the 1st International Workshop on Language Technology Platforms*, 89–95. Marseille, France: European Language Resources Association.

Rehm, Georg, Kalina Bontcheva, Khalid Choukri, Jan Hajic, Stelios Piperidis, and Andrejs Vasiljevs, eds. 2020a. *Proceedings of the 1st International Workshop on Language Technology Platforms*. Marseille, France: European Language Resources Association.

Rehm, Georg, Dimitris Galanis, Penny Labropoulou, Stelios Piperidis, Martin Wels, Ricardo Usbeck, Joachim Kohler, Miltos Deligiannis, Katerina Gkirtzou, Johannes Fischer, Christian Chiarcos, Nils Feldhus, Julian Moreno-Schneider, Florian Kintzel, Elena Montiel, Victor Rodriguez Doncel, John Philip McCrae, David Laqua, Irina Patricia Theile, et al. 2020b. "Towards an Interoperable Ecosystem of AI and LT Platforms: A Roadmap for the Implementation of Different Levels of Interoperability." In *Proceedings of the 1st International Workshop on Language Technology Platforms*, 96–107. Marseille, France: European Language Resources Association.

Schmid, Helmut. 1994. "Probabilistic Part-of-Speech Tagging Using Decision Trees." In *Proceedings of International Conference on New Methods in Language Processing*.

Schmid, Helmut. 2019. "Deep Learning-Based Morphological Taggers and Lemmatizers for Annotating Historical Texts." In *New York, NY, USA*. Association for Computing Machinery, 133–137.

Stan, Adriana, Junichi Yamagishi, Simon King, and Matthew Aylett. 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.

Steinberger, Ralf, Mohamed Ebrahim, and Marco Turchi. 2012. "JRC EuroVoc Indexer JEX - A Freely Available Multi-label Categorisation Tool." In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey: European Language Resources Association (ELRA), 798–805.

Steinberger, Ralf, Bruno Pouliquen, Anna Widiger, Camelia Ignat, Tomaž Erjavec, Dan Tufiș, and Dániel Varga. 2006. "The JRC-Acquis: A Multilingual Aligned Parallel Corpus with 20+ Languages." In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC'06)*, Genoa, Italy: European Language Resources Association (ELRA).

Stenetorp, Pontus, Sampo Pyysalo, Goran Topic, Tomoko Ohta, Sophia Ananiadou, and Junichi Tsujii. 2012. "BRAT: A Web-based Tool for NLP-assisted Text Annotation." In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, Avignon, France: Association for Computational Linguistics, 102–107.

Straka, Milan, Jan Hajič, and Jana Straková. 2016. "UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing." In *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia: European Language Resources Association.

Tufiș, Dan, Verginica Barbu Mititelu, Elena Irimia, Maria Mitrofan, Radu Ion, and George Cioroiu. 2019a. "Making Pepper Understand and Respond in Romanian." In *22nd International Conference on Control Systems and Computer Science (CSCS)*, 682–688.

Tufiș, Dan, and Verginica Barbu Mititelu. 2015. "The Lexical Ontology for Romanian." In *Language Production, Cognition, and the Lexicon*, edited by Núria Gala, Reinhard Rapp, and Gemma Bel-Enguix, 491–504. Berlin and New York: Springer International Publishing.

Tufiș, Dan, Verginica Barbu Mititelu, Elena Irimia, Vasile Păiș, Radu Ion, Nils Diewald, Maria Mitrofan, and Mihaela Onofrei. 2019b. "Little Strokes Fell Great Oaks. Creating CoRoLa, the Reference Corpus of Contemporary Romanian." *Revue Roumaine de Linguistique* 64 (3): 227–240.

Tufiș, Dan, Maria Mitrofan, Vasile Păiș, Radu Ion, and Andrei Coman. 2020. "Collection and Annotation of the Romanian Legal Corpus." In *Proceedings of the 12th Language Resources and Evaluation Conference*, 2773–2777. Marseille, France: European Language Resources Association.

Váradi, Tamás, Svetla Koeva, Martin Yamalov, Marko Tadić, Bálint Sass, Bartłomiej Nitoń, Maciej Ogrodniczuk, Piotr Pęzik, Verginica Barbu Mititelu, Radu Ion, Elena Irimia, Maria Mitrofan, Vasile Păiș, Dan Tufiș, Radovan Garabík, Simon Krek, Andraz Repar, Matjaž Rihtar, and Janez Brank. 2020. "The MARCELL Legislative Corpus." In *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille, France: European Language Resources Association, 3761–3768.