Romanian language translation in the RELATE platform

Vasile Păiș and **Maria Mitrofan** and **Andrei-Marius Avram**
Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy
vasile,maria,andrei.avram@racai.ro

Abstract

This paper presents the usage of the RELATE platform[1] for translation tasks involving the Romanian language. Using this platform, it is possible to perform text and speech data translations, either for single documents or for entire corpora. Furthermore, the platform was successfully used in international projects to create new resources useful for Romanian language translation.

1 Introduction

Translation platforms represent a subset of Language Technology (LT) platforms, providing services and resources for written or spoken language translation. Artificial intelligence (AI) methods are used to implement the platform functionalities. These can be used either online, following a request-response model, or offline for processing large corpora, following an initial upload in the platform.

Rehm et al. (2020a) notes that instead of competing with one another, platforms should be constructed to be interoperable and interact with each other to create synergies toward a productive LT ecosystem. We agree with this observation and consider that one way to achieve interoperability is through standardized formats for both input and output, allowing data to be exchanged between platforms. Furthermore, web services can expose internal functionality, allowing for integration into other systems.

This paper provides a detailed presentation of the translation functionalities of the RELATE platform . RELATE was developed as a modular, state-of-the-art platform for processing the Romanian language. Available functions are provided by modules developed in multiple national and international projects, both in-house and by partner institutions. Since its inception, one of its main goals was using standardized and easy-to-use file formats, combined with web APIs, thus allowing integration with other systems (Păiș, 2020) , as needed. Component integration is performed directly by consuming the provided APIs from a partner's servers or utilizing Docker containers hosted on one or multiple servers associated with the platform. Thus, it follows the philosophy behind the European Language Grid[2].

RELATE contains multiple translation functions for both text and speech. Furthermore, it allows for development of translation related corpora. The paper is organized as follows: Section 2 presents related work, Section 3 describes the current architecture of the platform and its evolution. Text translation functions are presented in Section 4. Speech to speech translation is covered in Section 5. Examples of large corpora, useful for translation, created within the platform in the context of international projects are given in Section 6. Finally we conclude in Section 7.

2 Related work

Coleman et al. (2020) presents an architecture developed for a Machine Translation (MT) platform that uses specific components and pre-existing services of Amazon Web Services to assure the security, robustness and scalability of the platform. Its main functionality is to provide translation services for news using a single integration point. With the needed translation technology integrated into one place, this platform facilitates news publication in multiple languages and through different virtual environments.

Franceschini et al. (2020) presents ELITR (European Live Translator) project, that aims to combine different NLP technologies such as automatic speech recognition, machine translation, and spo-

[1] https://relate.racai.ro

[2] https://www.european-language-grid.eu/

Proceedings of the 29th International Conference on Computational Linguistic, pages 64–74
October 12–17, 2022.

ken language translation to create end-to-end systems mainly for face-to-face conferences (interpreting official speeches and workshop-style discussions) and for remote conferences (live video streaming for which the platform automatically transcribes and translate subtitles). For now, ELITR's ASR technology is available for 6 EU languages. Since ELITR services work in real-time, the translation of the conversations starts immediately as the ASR service has an output available. ELITR technology is based upon PerVoice Service Architecture, a proprietary software solution that enables the concatenation of different services.

Khanna et al. (2021) presents Apertium, a free open-source platform for rule-based machine translation (RBMT) for under-resourced languages. Apertium is a complex pipeline consisting of multiple modules such as deformatter, source language morphological analyzer, source language morphological disambiguator, source language retokenization, lexical transfer, lexical selection, source language anaphora resolution, shallow structural transfer, recursive structural transfer, target language retokenization, target language morphological generator, target language post-generator, reformatter. One of the platform's main advantages is that users can add or remove modules according to their needs. Currently, the platform offers translation for eleven of the forty-four languages considered vulnerable or endangered.

Juremy[3] is an intelligent concordance search tool available for all combinations of the 24 EU official languages. It can be used to search legal and technical terminology in documents. In order to display the results, Juremy uses EUR-Lex and IATE databases and to reference the source document, Juremy provides the user with a series of metadata such as document title, topic, IATE evaluation, or work date. A plus of this online service is that it allows the user a customized search, but the services of Juremy are only available after registration on the website.

Rehm et al. (2020b) emphasize that numerous AI domains are underdeveloped at the national and international levels. Even though AI technologies such as deep neural networks offer significant opportunities for many societal and economic challenges, there is still work to do until LT technologies can be considered viable solutions for all 24 EU official languages. Another essential aspect underlined by the authors is the enormous fragmentation of the European AI and LT landscape and consider that efforts should be made to ensure that all these platforms can exchange information, data and services to identify synergies in market capitalization. Furthermore, the authors propose implementing standardized ways of exchanging repository entries that enable multiplatform and multi-vendor service workflows. Ai4EU[4] and ELG[5] platforms are presented as large European ecosystems that can assure interoperability between language technologies in Europe.

3 Architecture of the RELATE platform

The RELATE platform was implemented primarily for processing large text corpora (Păiș et al., 2019). In addition, it also offers access to state-of-the-art (SOTA) tools for the Romanian language on a "per request" use case. Recent developments allow speech processing of the Romanian language by integrating tools for automatic speech recognition (ASR), text-to-speech (TTS) and speech-to-speech translation.

Modules available in the platform include: TEPROLIN (Ion, 2018), NLP-Cube (Boroș et al., 2018) , UDPipe (Straka et al., 2016), TTL (Ion, 2007), MLPLA (Boroș et al., 2018), RomanianTTS (Stan et al., 2011), Legal-domain NER (Păiş et al., 2021), Biomedical NER (Mitrofan and Păiș, 2022). Also, some of the older, existing tools were exposed as web services and integrated in the platform. These modules account for the following operations: text segmentation (paragraph, sentence, token), phonetic transcription, lemmatization, syllabification, dependency parsing, text classification, term extraction, named entity recognition, diacritic restoration, abbreviation and numeral expansion, speech recording, ASR, TTS, text translation, and speech translation. Additionally, web interfaces are available for querying the Representative Corpus of Contemporary Romanian Language (CoRoLa) (Tufiș et al., 2019) and the Romanian WordNet (Tufiș and Barbu Mititelu, 2015).

From a user perspective, the platform provides two interfaces: document-based and corpus-based. In the document-based interface, the user can work with a single file (text document or speech recording) and obtains the processing results in near real-

[3] https://juremy.com/)

[4] https://www.ai4europe.eu/

[5] https://www.european-language-grid.eu/

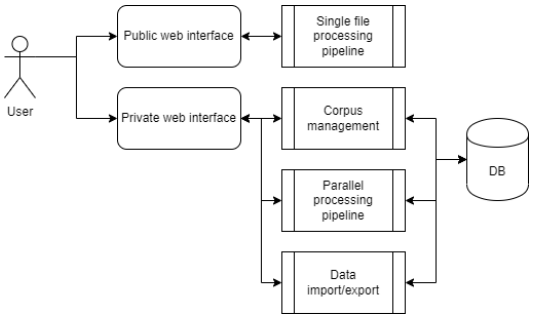

Figure 1: User perspective on the RELATE platform functionality

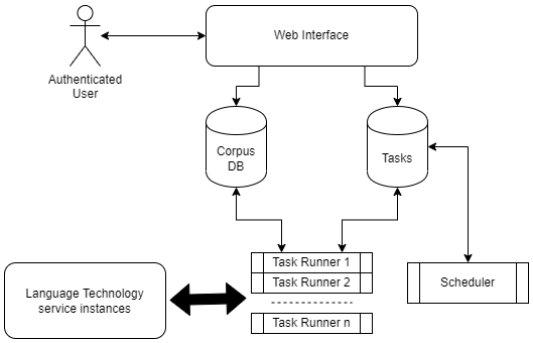

Figure 2: Task-based architecture in the RELATE platform

time. In the corpus-based interface, the user must first upload a corpus, then schedule processing tasks and finally obtain the results once all the tasks have been completed. Both interfaces are free for users, but the corpus-based interface requires the user to be registered (registration is provided free of charge for research purposes). This separation is presented in Figure 1.

From a technical point of view, the platform uses the same underlying LT services to operate in a single document and task-based modes. For performance purposes, different services can be instantiated multiple times on the same hardware nodes or servers. This methodology allows scaling the platform annotation capabilities with the size of the corpora to be processed. A scheduling component distributes the tasks across the available services. The number of instances associated with each LT service differs based on the service's speed. Faster services require fewer instances, while slower services benefit from more instances, thus allowing for increased parallelization of the processing queue. Figure 2 depicts the scheduling component with associated task runner processes.

The corpus management component provides basic functions such as file uploading (either a file-by-file process or an entire archive with multiple files), processing using the task-based system (task scheduling, task monitoring), visualization of both raw files and resulting annotations, and data export. Processed files can be exported in the internal format or converted to project-specific formats available within the platform. The internal platform format is based on the CoNLL-U Plus[6] specification, which in turn is derived from the basic CoNLL-U format[7], employed in the Universal Dependencies[8] project. This is a tabular format with an additional document or segment-specific metadata. The file starts with a metadata field ("global.columns") which describes the content associated with each column. For the RELATE platform, we keep the first ten columns corresponding to the basic CoNLL-U file and add additional columns, as needed, based on the tasks executed on each corpus. Therefore, the final annotated format may differ between corpora if different annotation tasks were executed. This can be further changed using format converters. Currently, converters are available for exporting in other CoNLL-U Plus structures or XML documents. Furthermore, due to our interest in Linguistic Linked Data, various Romanian language resources (Barbu Mititelu et al., 2020; Păiș and Barbu-Mititelu, 2022; Barbu Mitelu et al., 2022), some of which were created within the RELATE platform, were converted into RDF format. Examples of such resources are represented by the LegalNERo (Păiș et al., 2021) named entity corpus and the ROBIN Technical Acquisition Speech Corpus (RTASC) (Păiş et al., 2021).

Figure 3 presents the different components available in the RELATE platform. The web front-end is the graphical user interface employed to interact with the components. It handles unauthenticated interactions, user authentication and authenticated requests. The back-end exposes platform functionality as web APIs that can be consumed from the front end. This layer also allows for potential integration into other applications or platforms. At this level, corpus management functions are implemented, together with the task scheduling component. The other components, implementing specific processing functions, are called directly from the web back-end or task execution processes. Ro-

[6] https://universaldependencies.org/ext-format.html

[7] https://universaldependencies.org/format.html

[8] https://universaldependencies.org/

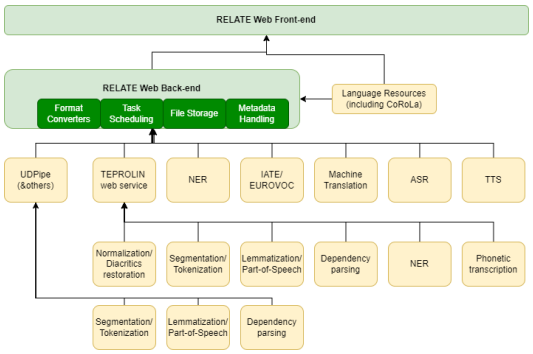

Figure 3: RELATE platform architecture

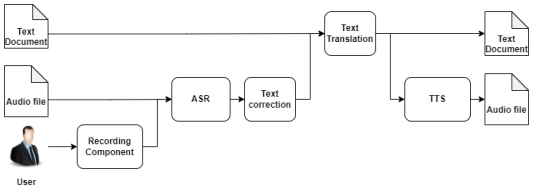

Figure 4: Translation flows in the RELATE platform

manian language resources can be downloaded or queried from the RELATE platform, such as pre-trained language models (for both word representations and annotation tasks) and gold annotated corpora. Considering the CoRoLa corpus, the user can directly access the main query interface of the text component, using the KorAP corpus analysis platform (Bański et al., 2012), and the speech component, allowing searching in audio files (Boroș et al., 2018) and listening to words being pronounced by Romanian speakers.

Translation in the RELATE platform is performed around a text translation component. However, due to the integration of both ASR and TTS components (for Romanian and English), it is possible to translate also speech (by using pre-recorded audio files or by using the integrated speech recording functionality), resulting in new audio files (available for download or direct playback). Different translation scenarios are depicted in Figure 4. As described above, depending on the use case, the translation pipeline can be invoked on a request basis or for entire corpora. Details on the text translation component are given in Section 4 and the speech-to-speech translation component is further described in Section 5.

4 Text translation

The "CEF Automated Translation toolkit for the Rotating Presidency of the Council of the EU" Action aimed to make the European Commission's

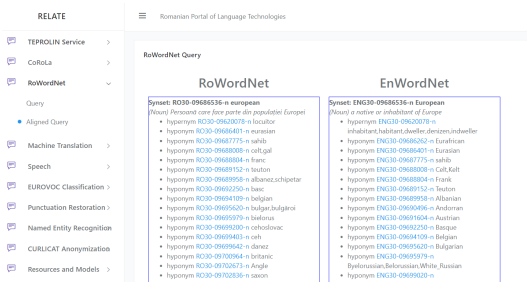

Figure 5: Aligned WordNet query using Romanian and English WordNets in the RELATE platform

eTranslation platform available to users from EU member states by extending the eTranslation platform with a set of custom MTs tailored for the EU Presidency domain. The Romanian-English and English-Romanian translation systems were improved by developing high-quality custom MT systems for the EU Presidency and DSI domains. For the Romanian language, the Research Institute for Artificial Intelligence "Mihai Drăgănescu" contributed to developing the translation system (Ro-En and En-Ro), a component of a wider system for the Presidency of the Council of the EU. The current MT platform[9] allows users to translate entire documents and local websites, including secure automated translation systems for all EU official languages.

Using the TILDE Machine Translation API[10], the textual translation component for Ro-En and En-Ro was integrated into the RELATE platform so that users can translate documents directly in the platform and also analyse the resulting document using the platform's functionalities.

In addition to the full-text translation, a version of the Romanian WordNet (Tufiș and Barbu Mititelu, 2015) aligned with the English WordNet (Miller, 1995) is available for querying. In this case, the user can look up a Romanian word and see the equivalent synset from the English WordNet. Figure 5 shows an example query for the word "european" (in English is written similarly, except with a capital letter "European").

5 Speech to speech translation

Automatic S2ST plays a core role in allowing people to communicate more naturally using spoken utterances when they do not share a common language, and nowadays, two methods are usually em-

[9] https://ro.presidencymt.eu

[10] https://www.tilde.com/developers/machine-translation-api

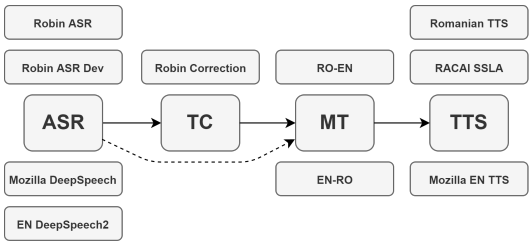

Figure 6: The proposed S2ST architecture with the four components: (1) automatic speech recognition (ASR), (2) textual correction (TC), (3) machine translation (MT) and (4) text-to-speech (TTS). The available models of each component are depicted in the upper part for Romanian and in the lower part for English (Avram et al., 2021b).

ployed for solving this problem: cascaded systems and end-to-end (E2E) models. Cascaded systems usually obtain better results compared to E2E models (Federico et al., 2020), but they have the drawback of propagating the error from one component to the next, making the overall system brittle. On the other hand, E2E models do not have this issue, and recent research has tried to minimize the gap between these two architectures (Jia et al., 2019).

Due to the limited amount of Romanian resources that are available for directly training an E2E model on this task, it was created a cascaded S2ST service for both Romanian to English[11] and English to Romanian[12] speech translation that contains four components in their respective pipeline (Avram et al., 2021b), as depicted in Figure 6. Each component incorporates at least one configurable model for the Romanian and English languages, enabling simple integration of new models into the system and improved flexibility in choosing a specific configuration given a potential requirement.

The first component of the cascaded S2ST system is the ASR to transcribe the audio input. The Romanian version has two models: Robin ASR and Robin ASR Dev. The former used the DeepSpeech2 architecture (Amodei et al., 2016) and was trained on approximately 230 hours of public speech data, obtaining a 9.91% word-error-rate (WER) on a customized dataset that was created by randomly extracting 5,000 samples from the training set (Avram et al., 2020). The Robin ASR Dev model is a specialized speech recognition system developed to better recognize utterances from the technical domain, specific to the ROBIN project[13]. ROBIN was a user-centred project designing software systems and services to use robots in an interconnected digital society. It also included a component for human-machine dialogue in specific micro-world scenarios (Ion et al., 2020). The Robin ASR Dev model (Avram et al., 2022) was trained only on the RTASC corpus and, in order to leverage the benefits of transfer learning on small datasets, we started from a Wav2Vec2 (Baevski et al., 2020) model that was pre-trained on the whole unlabeled audio data from VoxPopuli[14] (Wang et al., 2021) that is publicly available on HuggingFace[15]. Robin ASR Dev achieved 13.93% WER on the RTASC test set.

The English version of the ASR component also contains two models: Mozilla DeepSpeech, which is based on the DeepSpeech (Hannun et al., 2014) architecture and that contains the latest speech-to-text system offered by Mozilla [16], and EN DeepSpeech2 which, as Robin ASR, is based on the DeepSpeech2 model and was trained only on LibriSpeech (Panayotov et al., 2015). Both models were evaluated on the clean test set of LibriSpeech and obtained 7.06% WER and 9.19% WER, respectively. This difference in performance comes from the training set used for each model, Mozilla DeepSpeech being trained on more data than EN DeepSpeech2 which is not available for public usage.

The RELATE platform offers, at the time of writing this paper, a single model for textual correction on Romanian - Robin Correction that applies two postprocessing algorithms to the incoming transcriptions. Firstly, the component capitalizes the first character of the words that are found in a list of known named entities and then, in the second part, it replaces the unknown words from the transcription with known words from a vocabulary. This component is optional and can be removed from the cascaded system if needed (e.g. when working with uncased text or with an open vocabulary). A new neural punctuation restoration component for the Romanian language is still under active development and will become available in the future(Păiș and Tufiș, 2022; Păiș, 2022). A prototype is avail-

[11] https://relate.racai.ro/index.php?path=translate/speech_ro_en

[12] https://relate.racai.ro/index.php?path=translate/speech_en_ro

[13] https://aimas.cs.pub.ro/robin/en/

[14] https://huggingface.co/facebook/wav2vec2-large-100k-voxpopuli

[15] https://huggingface.co/facebook/wav2vec2-base-100k-voxpopuli

[16] https://github.com/mozilla/DeepSpeech

able for testing[17], but is not yet fully integrated and is not available in the speech translation pipeline. The Romanian to English and English to Romanian machine translation components use the API introduced in Section 4.

The English language comes with one model for the TTS component in our cascaded system - Mozilla EN TTS, a pretrained Tacotron2 with Dynamic Convolution Attention (Battenberg et al., 2020), that was developed by Mozilla[18] using the LJSpeech dataset[19]. The model obtained a median opinion score (MOC) of 4.31±0.06 using a 95% confidence interval. For the Romanian language, we offer two models that are not based on deep neural networks but use the classical Hidden Markov Models (HMM) to generate the audio output for a given sentence: Romanian TTS (Stan et al., 2011) and RACAI SSLA (Boroș et al., 2018). The difference between these two versions is that the former model outputs a synthesis of higher quality than the former model but at a slower rate, thus allowing a user to trade off computational speed for a better synthesis, or vice-versa, depending on the requirements.

6 Creating corpora relevant for machine translation

6.1 The MARCELL legislative corpus

The CEF Telecom project Multilingual Resources for CEF.AT in the legal domain (MARCELL) [20] had as the primary goal the enhancement of the eTranslation system developed by the European Commission. Within this project, seven legislative corpora have been created that contain the total body of national legislative documents in effect for seven countries included in the consortium: Bulgaria, Croatia, Hungary, Poland, Romania, Slovakia and Slovenia. All the corpora were tokenized, lemmatized and morphologically annotated, dependency parsed, named entities were also added, nominal phrases were identified together with IATE [21] terms and EuroVoc [22] descriptors. Interactive Terminology for Europe (IATE) has been the EU's terminology database since 2004. The primary purpose of this resource is to help translators working for the European Commission; that is why it is used in EU institutions and agencies for the collection, dissemination and management of terminology. It contains over 8 million terms in 24 official languages of the EU.

EuroVoc is a multilingual thesaurus developed and maintained by the Publications Office of the European Union. The main purpose for which it was built is to help process the information contained in documents issued by the EU institutions. The current version, EuroVoc 4.4, was released in 2012 and includes 6,883 unique IDs for thesaurus concepts, organized in 21 top-level domains, which are further refined in 127 micro-tesauri. It serves as the basis for the main domains of the IATE database. All the corpora are in CoNLL-U Plus format with fourteen columns in each file, the first ten columns keep the standard CONLL-U values (ID, FORM, LEMMA, UDPOS, XPOS, FEASTS, HEAD, DEPREL, DEPS and MISC), while the following four columns (NER, NP, IATE and EUROVOC) are specific to the MARCELL project.

Since texts from each corpus came from different sources, metadata harmonization was necessary to create a homogeneous resource in file format. Therefore, many fields were established, some mandatory for each language and others optional. The obligatory keys that assure the harmonization of the data are: id - unique identifier of the document, date - date of the document in ISO 8601 format, title - the title of the document in the original language, type - the legal type of the document in the original language, entype - the legal type of the document in English. The optional keys are: url - the address of each document, keywords - several keywords in the original language, and topic - the human-readable topic of the document in the original language. In (Váradi et al., 2020) are presented all the available metadata keys and attributes in source archives for each language.

The MARCELL Romanian language corpus contains approximately 144k processed legislative documents that can be classified into five main categories: governmental decisions (25%), ministerial orders (18%), decisions (16%), decrees (16%) and laws (6%). In terms of document length, most of them contain more than 1,000 words per document, and only 6,000 can be considered short documents because they contain less than 100 words per docu-

[17] https://relate.racai.ro/index.php?path=punctuation_restoration/demo

[18] https://github.com/mozilla/TTS

[19] https://keithito.com/LJ-Speech-Dataset/

[20] https://marcell-project.eu/

[21] https://iate.europa.eu/home

[22] https://eur-lex.europa.eu/browse/eurovoc.html?locale=ro

ment. A general overview of the Romanian legislative corpus can be seen in Table 1.

No. of raw documents	144,131
No. of sentences	4,300,131
No. of tokens	66,918,022
No. of unique lemmas	200,888
No. of unique tokens	281,532

Table 1: General statistics of the Romanian legal corpus

The Romanian legal corpus (Tufiş et al., 2020) was processed in the RELATE platform, using the integrated TEPROLIN web service (Ion, 2018). In terms of dependency parsing annotation, NLP-Cube (Boroș et al., 2018) was used, which according to the evaluation made by Păiș et al. (2021a), has a labelled attachment score (LAS) of 85.87 for Romanian. One of the objectives of the MARCELL project was the classification into EuroVoc topics and enrichment with EuroVoc and IATE terms identified in each of the seven monolingual corpora. The algorithm we employed for EuroVoc classification was based on static word embeddings representations (Păiș and Tufiș, 2018), trained on the CoRoLa corpus (Tufiș et al., 2019). These were used to train a classifier utilizing the FastText tool (Joulin et al., 2017). Currently, the RELATE platform also offers a transformer-based classification with EuroVoc descriptors that were developed later, using the PyEuroVoc toolkit (Avram et al., 2021a). After this step, all the corpora were compiled into a comparable corpus of seven languages aligned at the topic level domains identified by EuroVoc descriptors. This project activity has positively impacted both MT systems in the seven languages concerned and the improvement of both the e-justice and the Online Dispute Resolution Digital Service infrastructures. Regarding the identification of IATE and EuroVoc terms, the Romanian team used a custom algorithm similar to the Aho-Corasick algorithm (Aho and Corasick, 1975), that uses a language-specific compression function (Coman et al., 2019) and which has a term matching rate of approximately 98%. All these services were integrated into the RELATE platform (Păiș et al., 2019) so that its output is as visually descriptive as possible and can configure each processing step according to different algorithms integrated into the platform. As a result, the user only needs to specify the type of annotation desired to build the processing chain.

6.2 The CURLICAT corpus

Curated Multilingual Resources for CEF.AT (CURLICAT) [23] is an ongoing project that, similar to the MARCELL project, aims to deliver language resources, in particular monolingual corpora, in the EU/CEF languages: Bulgarian, Croatian, Hungarian, Polish, Romanian, Slovak and Slovenian. Unlike the MARCELL project, in which legislative documents belonging to each country in the consortium were collected, the CURLICAT project's main aim is to create seven monolingual corpora containing texts from the following fields: culture, education, economy, health, nature, politics, science. Creating these resources for the under-resourced languages will contribute to breaking down linguistic barriers to the creation of the Digital Single Market in Europe and implicitly will lead to improvements in automatic translations between EU's languages.

In order to assure a harmonized structure of the resulting resources, all the corpora use the CONLL-U Plus format; each language-specific sub-corpus has the same format. The first ten columns have the standard CONLL-U values, and the last three are specific to the CURLICAT project. Regarding the metadata for each document, the principles used in the MARCELL project have been adopted. After the harmonization of the metadata phase ends, all the metadata information will be classified as: obligatory - information that all partners have to provide, optional - information that can be missing or that contains an empty value in some language corpora, and local - information specific to a given language corpus. At the end of the project, each consortium partner will provide a corpus of at least 2 million sentences containing at least 20 million words. Each corpus will consist of at least 500k sentences (5 million words) for the five main domains: culture, economy, finances, health and science.

Regarding the Romanian component of the CURLICAT corpus, most texts were extracted from The Reference Corpus of the Contemporary Romanian Language (CoRoLa) (Tufiș et al., 2019). The documents were selected based on different metadata attributes present in CoRoLa metadata scheme. After selecting the texts according to the established criteria, they went through an automatic cleaning phase. Next, the texts were processed with the RELATE platform, so each corpus was tokenized, lemmatized, annotated with part of speech

[23] https://curlicat-project.eu/

tags, and dependency parsed. In Table 2 are given relevant statistics for each domain of the current version of the Romanian sub-corpus created for the CURLICAT project.

Domain	No. of sentences
culture	577,307
education	320,484
economy	311,721
health	417,681
nature	338,953
politics	379,188
science	2,113,454
TOTAL	4,458,788

Table 2: Current statistics of the CURLICAT-RO corpus

Since "the protection of natural persons in relation to the processing of personal data is a fundamental right" (Spiekermann, 2012), text anonymization is one of the natural processing phases that all the corpora need to go through. Therefore, for the Romanian language, an anonymization solution was implemented (Păiș et al., 2021b) and the "local" pseudonymization approach was considered. Since most anonymization requirements appear in relation to news and other blog posts, to allow NLP algorithms to use this resource better, it was decided to keep suffixes specific to Romanian named entities as part of the pseudonym being used. Experiments [24] have shown that this is a viable solution for anonymizing texts for the Romanian language. It was integrated into the RELATE platform to automatically allow the entire corpus to be anonymized. Upon completion, this project will make a significant contribution to different kinds of linguistic research, such as neural machine translation training, cross-lingual legal terminology extraction, or cross-lingual entity mapping.

7 Conclusion

In this paper, we described the usability of the RELATE platform in the context of machine translation of the Romanian language. It provides options for text and speech translation using a modular architecture. Additionally, the platform successfully created considerable language resources relevant for machine translation. Sections 6.1 and 6.2 described the creation of two large comparable corpora in 7 official EU languages, including the Romanian language.

The platform is designed to be highly customizable and easily extensible, while the standardized file formats ensure interoperability with other systems. Furthermore, the service-oriented architecture (SOA), based on REST APIs, allows for additional integration options with external applications or other language platforms. The RELATE platform is available open source on GitHub[25]. Its current form resulted from integration of components developed in many research projects. We aim to continue the development, both in terms of translation capabilities and more generally with regard to language technology for Romanian language.

Acknowledgements

Part of this research was conducted in the context of the "Curated Multilingual Language Resources for CEF.AT" (CURLICAT) project, CEF-TC-2019-1 – Automated Translation grant agreement number INEA/CEF/ICT/A2019/1926831. Part of this research was conducted in the context of the European Language Equality (ELE) project, action ELE/101018166, work programme PPPA-LANGEQ2020.

References

Alfred V Aho and Margaret J Corasick. 1975. Efficient string matching: an aid to bibliographic search. *Communications of the ACM*, 18(6):333–340.

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

Andrei-Marius Avram, Vasile Păiș, and Dan Tufiș. 2021a. PyEuroVoc: A tool for multilingual legal document classification with EuroVoc descriptors. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 92–101.

Andrei-Marius Avram, Vasile Păiș, and Dan Tufiș. 2020. Towards a Romanian end-to-end automatic speech recognition based on Deepspeech2. *Proceedings of the Romanian Academy Series A*, 21:395–402.

[24] https://github.com/racai-ai/ROAnonymization_CURLICAT

[25] https://github.com/racai-ai/RELATE

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

Andrei-Marius Avram, Vasile Păiș, and Dan Tufiș. 2022. Self-supervised pre-training in speech recognition systems. In Vasile Păiș, editor, *Speech Recognition Technology and Applications*, pages 27–56. Nova Science Publishers.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. Wav2Vec 2.0: A framework for self-supervised learning of speech representations. *Advances in Neural Information Processing Systems*, 33:12449–12460.

Piotr Bański, 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)*, pages 2905–2911, Istanbul, Turkey. European Language Resources Association (ELRA).

Verginica Barbu Mititelu, Elena Irimia, Vasile Păiș, Andrei-Marius Avram, Maria Mitrofan, and Eric Curea. 2020. Romanian resources in LLOD format. In *Proceedings of the 15th International Conference Linguistic Resources and Tools for Natural Language Processing*, pages 29–40, online.

Verginica Barbu Mititelu, Elena Irimia, Vasile Păiș, Andrei-Marius Avram, and Maria Mitrofan. 2022. Use case: Romanian language resources in the lod paradigm. In *Proceedings of the Linked Data in Linguistics Workshop @ LREC2022*, pages 35–44. European Language Resources Association (ELRA).

Eric Battenberg, 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)*, pages 6194–6198. IEEE.

Tiberiu Boroș, Stefan Daniel Dumitrescu, and Ruxandra Burtica. 2018. 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*, pages 171–179, Brussels, Belgium. Association for Computational Linguistics.

Tiberiu Boroș, Ștefan Dumitrescu, and Vasile Păiș. 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.

Susie Coleman, 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*, pages 16–21, Marseille, France. European Language Resources Association.

Andrei Coman, Maria Mitrofan, and Dan Tufiș. 2019. 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*, pages 39–49.

Marcello Federico, Alex Waibel, Kevin Knight, Satoshi Nakamura, Hermann Ney, Jan Niehues, Sebastian Stüker, Dekai Wu, Joseph Mariani, and Francois Yvon, editors. 2020. *Proceedings of the 17th International Conference on Spoken Language Translation*. Association for Computational Linguistics, Online.

Dario Franceschini, Chiara Canton, Ivan Simonini, Armin Schweinfurth, Adelheid Glott, Sebastian Stüker, Thai-Son Nguyen, Felix Schneider, Thanh-Le Ha, Alex Waibel, et al. 2020. Removing european language barriers with innovative machine translation technology. In *Proceedings of the 1st International Workshop on Language Technology Platforms*, pages 44–49.

Awni Hannun, 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*.

Radu Ion. 2007. *Word sense disambiguation methods applied to English and Romanian*. Ph.D. thesis, Romanian Academy. In Romanian.

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

Radu Ion, 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.

Ye Jia, Ron J Weiss, Fadi Biadsy, Wolfgang Macherey, Melvin Johnson, Zhifeng Chen, and Yonghui Wu. 2019. Direct speech-to-speech translation with a sequence-to-sequence model. *Proc. Interspeech 2019*, pages 1123–1127.

Armand Joulin, 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*, pages 427–431, Valencia, Spain. Association for Computational Linguistics.

Tanmai Khanna, Jonathan N Washington, Francis M Tyers, Sevilay Bayatlı, Daniel G Swanson, Tommi A Pirinen, Irene Tang, and Hèctor Alòs i Font. 2021. Recent advances in apertium, a free/open-source rule-based machine translation platform for low-resource languages. *Machine Translation*, pages 1–28.

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

Maria Mitrofan and Vasile Păiș. 2022. Improving Romanian BioNER using a biologically inspired system. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 316–322, Dublin, Ireland. Association for Computational Linguistics.

Vasile Păiș, Maria Mitrofan, Carol Luca Gasan, Vlad Coneschi, and Alexandru Ianov. 2021. Named entity recognition in the Romanian legal domain. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 9–18.

Vasile Păiș. 2020. Multiple annotation pipelines inside the RELATE platform. In *The 15th International Conference on Linguistic Resources and Tools for Natural Language Processing*, pages 65–75.

Vasile Păiș, Radu Ion, Andrei-Marius Avram, Elena Irimia, Verginica Barbu Mititelu, and Maria Mitrofan. 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. IEEE.

Vasile Păiș, Dan Tufiș, and 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*, pages 181–192.

Vassil Panayotov, 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)*, pages 5206–5210. IEEE.

Vasile Păiș, Radu Ion, Andrei-Marius Avram, Maria Mitrofan, and Dan Tufiș. 2021a. In-depth evaluation of Romanian natural language processing pipelines. *Romanian Journal of Information Science and Technology (ROMJIST)*, 24(4):384–401.

Vasile Păiș, Elena Irimia, Radu Ion, Dan Tufiș, Maria Mitrofan, Verginica Barbu Mititelu, Andrei-Marius Avram, and Eric Curea. 2021b. Romanian text anonymization experiments from the CURLICAT project. In *The 17th Edition of the International Conference on Linguistic Resources and Tools for Natural Language Processing - CONSILR*.

Vasile Păiș and Dan Tufiș. 2018. Computing distributed representations of words using the CoRoLa corpus. *Proceedings of the Romanian Academy Series A - Mathematics Physics Technical Sciences Information Science*, 19(2):185–191.

Vasile Păiş, Maria Mitrofan, Carol Luca Gasan, Vlad Coneschi, and Alexandru Ianov. 2021. Named entity recognition in the Romanian legal domain. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pages 9–18, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Vasile Păiş and Dan Tufiş. 2022. Capitalization and punctuation restoration: a survey. *Artificial Intelligence Review*, 55(3):1681–1722.

Vasile Păiș. 2022. Punctuation recovery for romanian transcribed documents. In Vasile Păiș, editor, *Speech Recognition Technology and Applications*, pages 119–154. Nova Science Publishers.

Vasile Păiș and Verginica Barbu-Mititelu. 2022. Linguistic linked open data for speech processing. In Vasile Păiș, editor, *Speech Recognition Technology and Applications*, pages 155–188. Nova Science Publishers.

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

Georg Rehm, Dimitris Galanis, Penny Labropoulou, Stelios Piperidis, Martin Welß, Ricardo Usbeck, Joachim Köhler, Miltos Deligiannis, Katerina Gkirtzou, Johannes Fischer, Christian Chiarcos, Nils Feldhus, Julian Moreno-Schneider, Florian Kintzel, Elena Montiel, Víctor Rodríguez Doncel, John Philip McCrae, David Laqua, Irina Patricia Theile, Christian Dittmar, Kalina Bontcheva, Ian Roberts, Andrejs Vasiljevs, and Andis Lagzdiņš. 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*, pages 96–107, Marseille, France. European Language Resources Association.

Sarah Spiekermann. 2012. The challenges of privacy by design. *Communications of the ACM*, 55(7):38–40.

Adriana Stan, 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.

Milan Straka, 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.

Dan Tufiș and Verginica Barbu Mititelu. 2015. *The Lexical Ontology for Romanian*, pages 491–504. Springer International Publishing, Cham.

Dan Tufiș, Verginica Barbu Mititelu, Elena Irimia, Vasile Păiș, Radu Ion, Nils Diewald, Maria Mitrofan, and Mihaela Onofrei. 2019. Little strokes fell great oaks. creating CoRoLa, the reference corpus of contemporary Romanian. *Revue Roumaine de Linguistique*, 64(3):227 – 240.

Dan Tufiş, 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*, pages 2773–2777.

Tamás Váradi, 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*, pages 3761–3768, Marseille, France. European Language Resources Association.

Changhan Wang, Morgane Riviere, Ann Lee, Anne Wu, Chaitanya Talnikar, Daniel Haziza, Mary Williamson, Juan Pino, and Emmanuel Dupoux. 2021. VoxPopuli: A large-scale multilingual speech corpus for representation learning, semi-supervised learning and interpretation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 993–1003.