Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2773–2777
Marseille, 11–16 May 2020

Collection and Annotation of the Romanian Legal Corpus

Dan Tufiș, Maria Mitrofan, Vasile Păiș, Radu Ion, Andrei Coman
Research Institute for AI ”Mihai Drăgănescu”, Romanian Academy
”Calea 13 Septembrie”, Bucharest 050711, Romania
{tufis, maria, vasile, radu}@racai.ro

Abstract

We present the Romanian legislative corpus which is a valuable linguistic asset for the development of machine translation systems, especially for under-resourced languages. The knowledge that can be extracted from this resource is necessary for a deeper understanding of how law terminology is used and how it can be made more consistent. At this moment, the corpus contains more than 144k documents representing the legislative body of Romania. This corpus is processed and annotated at different levels: linguistically (tokenized, lemmatized and POS-tagged), dependency parsed, chunked, named entities identified and labeled with IATE terms and EUROVOC descriptors. Each annotated document has a CONLL-U Plus format consisting of 14 columns; in addition to the standard 10-column format, four other types of annotations were added. Moreover the repository will be periodically updated as new legislative texts are published. These will be automatically collected and transmitted to the processing and annotation pipeline. The access to the corpus is provided through ELRC infrastructure.

Keywords: corpus, annotation, law, Romanian

1. Introduction

In this paper we review the first results of the new project, ”Multilingual Resources for CEF.AT in the legal domain” (MARCELL) action[1] whose final goal is to enable enhancement of the automatic translation in CEF.AT [2] on the body of national legislation in seven EU official languages. For this task, all the seven teams cooperated in order to produce a comparable corpus aligned at the top-level domains identified by EUROVOC descriptors[3].

EUROVOC is a multilingual thesaurus maintained by the Publications Office of the European Union. It exists in the 24 official languages of the European Union. It is used by many major institutions in Europe which include: several governmental departments, regional and national parliaments in the continent, the Council of the European Union and the European Parliament. It contains keywords organized in 21 fields and 127 micro-thesauri and it serves as the basis for the domain names used in the European Union's terminology database: Interactive Terminology for Europe (IATE) [4]. IATE is the EU's terminology database and it is used for dissemination and management of EU-specific terminology. One of its main aims is to facilitate the task of translators working for the EU. Currently, it has over 8 million terms and uses the EUROVOC thesaurus as a domain classification system.

In the following we describe the activities and the results of the Romanian team. A general view of the project activities is given in another article (Váradi et al., 2020). For the Romanian language, the current legal database created includes more than 144k legislative processed documents, issued starting from 1881. Since the last round of document crawling, there were published, in one year, more than 40k new legal texts, not yet included into the database.

The first part of the paper presents the main goals of this project together with the process of collecting the corpus. The second part details the statistics of the corpus. The third part presents the annotation process of the corpus: part-of-speech tagging, dependency parsing, nominal phrases identification, named entity recognition and IATE and EUROVOC terms identification.

2. Collecting The Romanian Legal Corpus

2.1. Goals of the MARCELL Project

Since the techniques for processing and exploiting corpora have been developed and are not dependent on features of specific languages, text corpora have become the main source of research data in computer linguistics. Lately, it has become a common practice to use the web for corpus acquisition, but in general most of the texts gathered in a corpus have restrictions regarding the laws of intellectual property and licensing. However, there is a type of text data that is exempted from copyright protection (unlike web-based texts) – the law body, which is often specifically excluded from copyright laws. In law texts including constitution, acts, public notices and court judgements the used sub-language is partially restricted and also different from general language. Therefore, within this project, seven monolingual corpora of national legislation documents will be created in order to provide automatic translation on the body of national legislation (laws, decrees, regulations, etc) in seven countries: Bulgaria, Croatia, Hungary, Poland, Romania, Slovakia and Slovenia.

A related project was, years ago, JRC-Acquis, concerned with compiling a parallel corpus of the total body of European Union (EU) law applicable in the the EU Member States (Steinberger et al., 2006). Unlike JRC-Acquis corpus, the MARCELL comparable corpus addresses national laws in only 7 EU countries and is supposed to be a very good companion thematic data to JRC-Acquis with little (if

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

[2] https://ec.europa.eu/inea/sites/inea/files/building_block_dsi_introdocument_etranslation_v0.0.7.pdf

[3] https://eur-lex.europa.eu/browse/eurovoc.html

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

any) duplication.

The main goals of the first phase of this action are: to produce a pre-processed (tokenized and morphologically tagged) monolingual corpus of national legislative texts for each of the seven languages; to identify the IATE and EUROVOC terms in texts and to classify the documents according to the 21 EUROVOC fields (top-level domains).

2.2. The corpus building process

The acquisition of the texts included in the Romanian legislative corpus was done via crawling. In order to collect the corpus, we used the Romanian legislative portal [5], which provides free access to all the legal documents issued since 1881. During this step, all the HTML-tags were eliminated together with style sheets, objects, tables, figures etc. From each file we collected only raw texts and the information needed to create metadata such as: document type, issuer, date, title and URL.

3. Corpus Statistics

The corpus contains more than 144k files ranging from 1881 to 2018. In Table 1 we present general statistics for the annotated corpus.

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 corpus

There are five main types of Romanian legal documents: governmental decisions (25%), ministerial orders (18%), decisions (16%), decrees (16%) and laws (6%).

After the statistics were calculated, we found that there are six main issuers of the documents: Government (28%), Ministers (19%), President (14%), Constitutional Court (12%), Parliament (6%) and National Authorities (4%).

Concerning the time-stamp, most of the published documents were issued after year 2000. Before 1990, almost 4,000 documents were issued and between 1990 and 2000 around 21,000 legal documents were published. After year 2000, the number of issued documents has increased and, on average, more than 6,000 documents were issued every year, reaching a total of 120,000 until 2018, in 19 years.

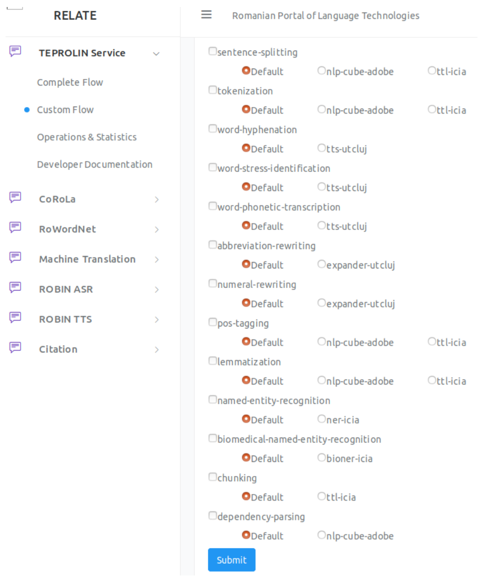
In terms of document length, there are around 6,000 short documents (less than 100 words per document, most of them being updates to other previously published legal documents), 70,000 documents contain between 100 and 500 words per document, more than 18,000 documents have around 1000 words per document and 52,000 contain more than 1000 words.

4. Corpus Annotation

4.1. Linguistic Annotation

The corpus is annotated in batches, as new documents are collected. The processing flow is part of the RELATE portal [6] (Păiș et al., 2019) and it includes text normalization, sentence splitting, tokenization, POS tagging, lemmatization, dependency parsing, named entity recognition and classification, chunking, IATE term annotation and top level EUROVOC labeling.

The preprocessing pipeline, excluding IATE and EUROVOC annotations, is done using the TEPROLIN text preprocessing platform (Ion, 2018), which was integrated into RELATE such that its output is as visually descriptive as possible. TEPROLIN can be easily configured to use different algorithms to do different parts of the text preprocessing pipeline and it only needs a list of desired text annotations to infer and construct the pipeline getting these annotations out (see Figure 1).



□sentence-splitting
● Default ○nlp-cube-adobe ○ttl-icia
□tokenization
● Default ○nlp-cube-adobe ○ttl-icia
□word-hyphenation
● Default ○tts-utcluj
□word-stress-identification
● Default ○tts-utcluj
□word-phonetic-transcription
● Default ○tts-utcluj
□abbreviation-rewriting
● Default ○expander-utcluj
□numeral-rewriting
● Default ○expander-utcluj
□pos-tagging
● Default ○nlp-cube-adobe ○ttl-icia
□lemmatization
● Default ○nlp-cube-adobe ○ttl-icia
□named-entity-recognition
● Default ○ner-icia
□biomedical-named-entity-recognition
● Default ○bioner-icia
□chunking
● Default ○ttl-icia
□dependency-parsing
● Default ○nlp-cube-adobe

Figure 1: Capture of TEPROLIN services

The sentence splitting, tokenization, POS tagging, lemmatization and chunking are done with TTL (Ion, 2007) which has a POS tagging accuracy of around 98% with the rich MSD tag set [7] and a lemmatization accuracy for unknown word of around 83%. TTL has been used to process the Romanian side of the Acquis-Communautaire and since then, it was systematically improved. The mentioned accuracy refers to JRC-Acquis text processing. As the MARCELL texts are of the same genre, we estimate the same accuracy as for JRC-Acquis. Dependency parsing is produced by NLP-Cube [8] (Boroș et al., 2018) which, according to the evaluations done in the CoNLL 2018 shared task "Multilingual Parsing from Raw Text to Universal Dependencies" [9], has a labelled attachment score of around 85% for Romanian.

Table 2 presents the distribution of content words in the

[5] `http://legislatie.just.ro/`

[6] `https://relate.racai.ro`

[7] `http://nl.ijs.si/ME/V4/msd/html/msd-ro.html`

[8] `https://github.com/adobe/NLP-Cube`

[9] `https://universaldependencies.org/conll18/results-las.html`, see the "Per treebank LAS-F1" section and scroll down to the "ro_rrt" table.

legal corpus. As it can be seen, nouns are the most frequent ones, followed by adjectives, highlighting the fact that law terms typically consist of the nouns, noun groups and other nominal constructions.

Tag	Number of tags
Nouns	21,792,691
Verbs	4,361,160
Adjectives	4,960,693
Adverbs	860,172

Table 2: Statistics of the content words

4.2. Annotation with NER

Part of the overall annotations required within the MARCELL project is named entity recognition. In this context, we used the module integrated in the RELATE platform (Păiș et al., 2019). This is a general named entity recognizer for Romanian language implemented using Conditional Random Fields (CRF), based on the Stanford NER (Finkel et al., 2005) software package. It is enhanced with the Romanian word embeddings (Păiș and Tufiș, 2018) learned from the CoRoLa corpus (Tufiș et al., 2019). Furthermore, it uses the embeddings web service from the RELATE platform in order to obtain at runtime representations for previously unknown Romanian words. In total, a number of four entity classes can be identified: person (PER), organization (ORG), location (LOC) and time (TIME). In accordance with the MARCELL project specifications, the resulting annotation is added in IOB format (Sang and Veenstra, 1999), in a dedicated column in the resulting annotated file. Table 3 presents the distribution of named entity annotation over each of four entity classes. The NER accuracy, which was estimated on 500 randomly selected documents) is the lowest among our processing modules (64.1%) with the TIME annotation significantly better (94.42%).

NE type	Number of entities
LOC	574,400
ORG	2,096,680
PER	1,153,053
TIME	1,357,692

Table 3: Named entities distribution

4.3. Annotation with IATE terms and EUROVOC descriptors

In order to tackle the identification of IATE terms (for Romanian, IATE terminology consists of about 55,000 terms) and EUROVOC descriptors, we developed a linear time, approximate-string-matching algorithm that combines several string-matching techniques and language specific properties (Coman et al., 2019). Because of the computational limitations imposed by approximate-string-matching algorithms, we attempted to convert this issue into a perfect-string-matching one, which we could then tackle with well-known linear-time techniques.

After concluding that classical distance measures between strings (i.e. Levenshtein Distance, Hamming Distance) were not suitable for massive data linguistic processing, we aimed to create a function which could give identical results for forms of the same term, and different results for forms of different terms, thus enabling us to check the matching between any two forms. Therefore, we introduced the concept of a Compression Function, which aimed to provide a pseudo-lemmatization based solely on word structure. Its definition was established through several steps and optimizations, which are progressively discussed throughout the paper (Coman et al., 2019). The final form of the Compression Function took a structure containing only Romanian alphabet letters, punctuation and spacing, and only kept the first letter of each word, the consonants, the spacing and, in some cases, capitalization, in order to construct the image (e.g. ”Navigație RNAV” → ”nvgț RNV”). The function could be written in pseudocode as follows (Algorithm 1).

```
Function Normalize (string S) is
    S ← RemovePunctuation(S)
    S ← SelectiveLowercase(S)
end
Function Compress (string S) is
    S ← Normalize(S)
    S ← Keep the spacing/consonants/first letter of
      each word
end
```

Algorithm 1: Compression Function

The final algorithm was based on the Aho-Corasick (Aho and Corasick, 1975) data structure and the previously defined Compression Function. It also introduced a processing separation between short terms (having at most 4 consonants) and long terms (all the other terms) in order to increase identification accuracy. Short terms were directly inserted into an Aho-Corasick structure, through which the corpus was also directly passed in order to identify the matches. On the other hand, long terms were first passed through the Compression Function (pseudo-lemmatized), then inserted into a different Aho-Corasick structure, through which we passed the image of the corpus through the Compression Function (Algorithm 2).

It is worth mentioning that, unlike other string-matching algorithms like Levenshtein Automata, Aho-Corasick does not impose the need to process multi-word terms separately. Each term is passed through the Compression Function regardless of its structure, then the corpus is “fed” character by character to the structure. The identified terms from those two Aho-Corasick structures (represented by ShortTerms/LongTerms in the previous pseudocode) were then merged and inserted in the legal corpus. Overall, there were 51,517,877 matches, with an average of 347 IATE terms matched per document.

In order to compute the accuracy of the algorithm, we took into consideration both the fraction of positive-matches which we identified and the fraction of matches which are false-positives. Thus, an evaluation over a testing sample yielded an accuracy of 98%-99%. Due to the size of the

```
Function ProcessShortTerms is
    AhoCorasick ShortTerms
    for each IATE short term T do
        insert Normalize(T) in ShortTerms
    end
    for each document D do
        pass Normalize(D) through ShortTerms
    end
end
Function ProcessLongTerms is
    AhoCorasick LongTerms
    for each IATE long term T do
        insert Compress(T) in LongTerms
    end
    for each document D do
        pass Compress(D) through LongTerms
    end
end
```

Algorithm 2: Aho-Corasick Processing

sample, the actual value is expected to be slightly lower.

In each document, the legal terms identified from IATE and EUROVOC can be found on columns 13 and 14 respectively. The IATE and EUROVOC labels are prefixed with a number counting the terms in the current document. For multi-word terms, this counter allows correct term identification. In figure 2, MONITORUL (English “the instructor”) is the first term in the current document, identified by the IATE code 1394636 and EUROVOC descriptor 3206 (Education and Communication). However, MONITORUL OFICIAL (English “the official monitor”) is a different term (the second) with IATE code 3522817 for which three EUROVOC descriptors applies: 3221, 7206, 7231.

4.4. File structure and metadata

In order to enable further analysis for all seven action languages, the format of the processed documents is the same, irrespective of the language. Each document has a CoNLL-U Plus[10] format and begins with a line describing the columns followed by a newdoc marker holding the file id (# newdoc id = ro.legal). Each sentence in a document is labelled by a unique ID (# sent_id = ro_legal.4), followed by the text of the respective sentence (# text = . . .) and then the vertical analysis, CoNLL-U Plus with 14 columns, of the tokens occurring in the sentence. Each file also contains the corresponding in-line metadata: the title of the document, date of issue, document type and URL. For the purpose of local corpus management, we also created standoff metadata to be used in the KORAP platform we use for the exploitation of the CoRoLa national corpus (Tufiș et al., 2019). The structure of a line is the following, the line fields being tab-separated (see Figure 2): ID, FORM, LEMMA, UDPOS, XPOS, FEATS, HEAD, DEPREL, _, _, NER, CHUNK, IATE, EUROVOC.

After the language specific processing the documents are archived and sent to the next processing hub: the multilingual clustering and comparable documents semantic alignment phase begins.

5. Availability

The corpus is stored in a uniform representation format and is already made available to ELRC[11]. Using the ELRC infrastructure and protocols, the corpus can be accessed in two forms: raw legislative documents or linguistically annotated legislative documents. Moreover, periodically new documents will be added to the corpus, pre-processed, annotated and classified. This process will assure that language-specific features and changes will be captured.

6. Conclusions

In this paper, we described the process of creating a large-scale monolingual corpus of national legislation documents enhanced with different types of annotations. Identifying the terms from both IATE and EUROVOC makes this resource very useful in the development of machine translation systems. Moreover, the work presented in this paper emphasizes the fact that the construction of domain-specific corpora also involves putting work and effort into developing domain-specific annotation tools.

We are planning to classify all the documents according to the 21 top-level EUROVOC categories. Several approaches will be used in order to determine the optimal one. Currently, the classification of the documents is done based on the most frequent EUROVOC category in each document, but we are also working on a classification based on word embeddings and on another one using the JRC Eurovoc Indexer JEX[12], which is pre-trained for all EU official languages.

One of the main goals of the MARCELL project is to ensure sustainability by continuous feeding of the repository with new incoming data and ensuring time-persistence and low maintenance times of the processing pipelines against the OS updates and other changes between hosts and environments. In this context, the size of the Romanian legal corpus is expected to increase in both raw and annotated data. Furthermore, the Romanian language-specific processing flow, as all language-specific flows will be containerized, using Docker or similar technologies.

7. Acknowledgements

This research was supported by the EC grant no. INEA/CEF/ICT/A2017/1565710 for the Action no. 2017-EU-IA-0136 entitled “Multilingual Resources for CEF.AT in the legal domain” (MARCELL).

8. Bibliographical References

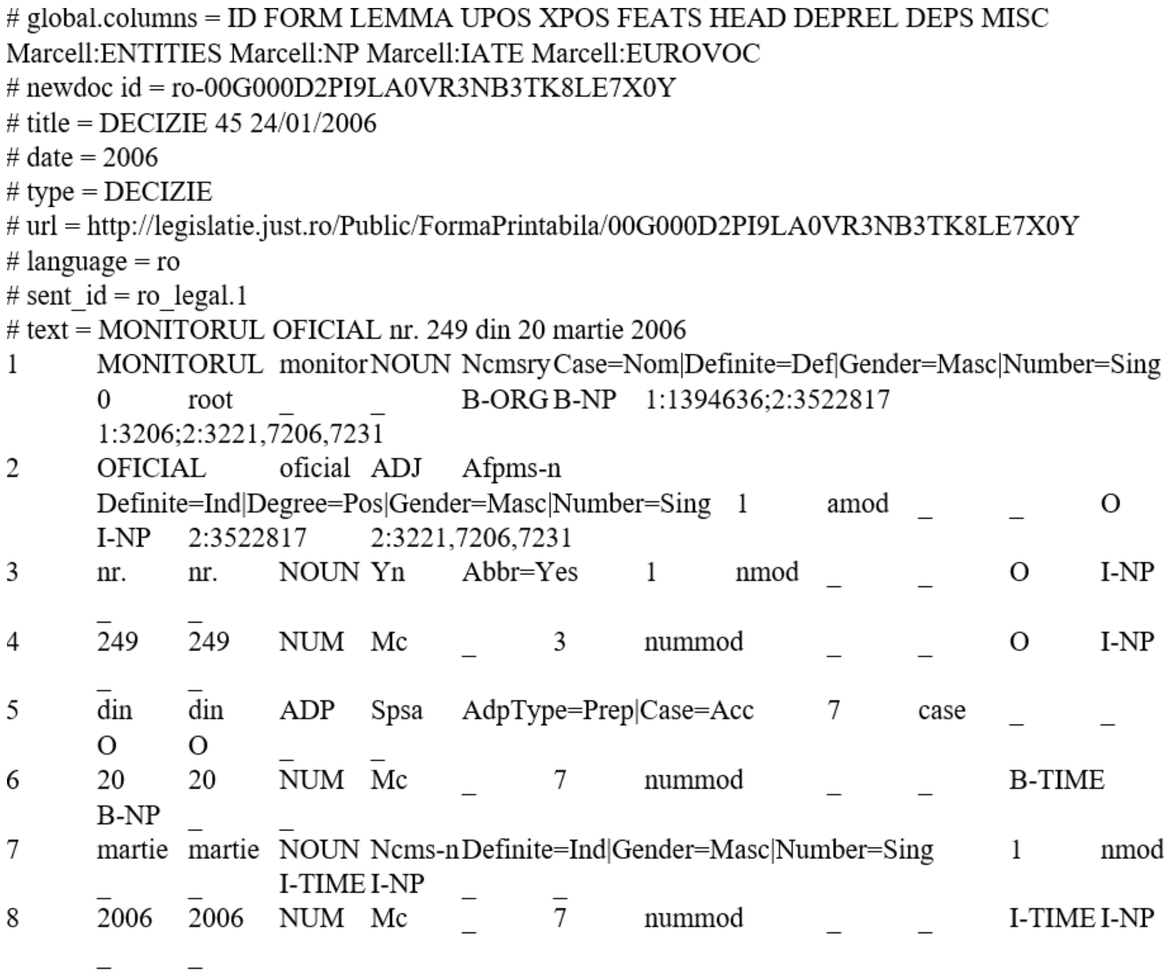
Aho, A. V. and Corasick, M. J. (1975). Efficient string matching: an aid to bibliographic search. *Communications of the ACM*, 18(6):333–340.

Boroș, T., Dumitrescu, S. D., and Burtică, R. (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*

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

[11] http://www.lr-coordination.eu/

[12] https://ec.europa.eu/jrc/en/language-technologies/jrc-eurovoc-indexer



```
# global.columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC
Marcell:ENTITIES Marcell:NP Marcell:IATE Marcell:EUROVOC
# newdoc id = ro-00G000D2PI9LA0VR3NB3TK8LE7X0Y
# title = DECIZIE 45 24/01/2006
# date = 2006
# type = DECIZIE
# url = http://legislatie.just.ro/Public/FormaPrintabila/00G000D2PI9LA0VR3NB3TK8LE7X0Y
# language = ro
# sent_id = ro_legal.1
# text = MONITORUL OFICIAL nr. 249 din 20 martie 2006
1       MONITORUL  monitor NOUN  Ncmsry Case=Nom|Definite=Def|Gender=Masc|Number=Sing
        0       root    _      _      B-ORG B-NP   1:1394636;2:3522817
        1:3206;2:3221,7206,7231
2       OFICIAL    oficial  ADJ   Afpms-n
        Definite=Ind|Degree=Pos|Gender=Masc|Number=Sing  1      amod    _       _       O
        I-NP    2:3522817       2:3221,7206,7231
3       nr.     nr.     NOUN  Yn      Abbr=Yes      1       nmod    _       _       O       I-NP
        _       _
4       249     249     NUM   Mc      _       3       nummod          _       _       O       I-NP
        _       _
5       din     din     ADP   Spsa    AdpType=Prep|Case=Acc        7       case    _       _
        O       O       _     _
6       20      20      NUM   Mc      _       7       nummod          _       _       B-TIME
        B-NP    _       _
7       martie  martie  NOUN  Ncms-n Definite=Ind|Gender=Masc|Number=Sing         1       nmod
        _       _       I-TIME I-NP   _       _
8       2006    2006    NUM   Mc      _       7       nummod          _       _       I-TIME I-NP
        _       _
```

Figure 2: Example of an annotated sentence together with the document metadata

Dependencies, pages 171–179. Association for Computational Linguistics (ACL), October.

Coman, A., Mitrofan, M., and Tufiș, D. (2019). Automatic identification and classification of legal terms in romanian law texts identification and classification. In *Proceedings of the International Conference on Linguistic Resources and Tools for Processing Romanian Language – CONSILR*, pages 39–49.

Finkel, J. R., Grenager, T., and Manning, C. (2005). Incorporating non-local information into information extraction systems by gibbs sampling. In *Proceedings of the 43rd annual meeting on association for computational linguistics*, pages 363–370. Association for Computational Linguistics.

Ion, R. (2007). *Word Sense Disambiguation Methods Applied to English and Romanian*. Ph.D. thesis, Romanian Academy.

Ion, R., (2018). *TEPROLIN: An Extensible, Online Text Preprocessing Platform for Romanian*, pages 69–76. Editura Universității "Alexandru Ioan Cuza", Iași.

Păiș, V. and Tufiș, D. (2018). Computing distributed representations of words using the corola corpus. In *Proceedings of the Romanian Academy, series A*, pages 403–410.

Păiș, V., Tufiș, D., and Ion, R. (2019). Integration of Romanian NLP tools into the RELATE platform. In *Proceedings of the International Conference on Linguistic Resources and Tools for Processing Romanian Language – CONSILR*, pages 181–192.

Sang, E. F. and Veenstra, J. (1999). Representing text chunks. In *Proceedings of the ninth conference on European chapter of the Association for Computational Linguistics*, pages 173–179. Association for Computational Linguistics.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiș, D., and Varga, D. (2006). The jrc-acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the 5th LREC Conference*, pages 2142–2147, Genoa, Italy, May.

Tufiș, D., Mititelu, V. B., Irimia, E., Paiș, V., Ion, R., Diewald, N., Mitrofan, M., and Onofrei, M. (2019). Little strokes fell great oaks. creating corola, the reference corpus of contemporary romanian. In *Revue roumaine de linguistique, no.3*, pages 227–240.

Váradi, T., Koeva, S., Yalamov, M., Tadić, M., Sass, B., Nitoń, B., Ogrodniczuk, M., Pezik, P., Barbu Mititelu, V., Ion, R., Irimia, E., Mitrofan, M., Păiș, V., Tufiș, D., Garabik, R., Krek, S., Repar, A., and Rihtar, M. (2020). The MARCELL Legislative Corpus. In *this volume*, Marseille, France.