LegalNERo: A linked corpus for named entity recognition in the Romanian legal domain

Vasile Păis a,*, Maria Mitrofan a, Carol Luca Gasan a, Alexandru Ianov a, Corvin Ghită a, Vlad Silviu Coneschi a, and Andrei Onut a

Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy, Romania
E-mail: vasile@racai.ro

Abstract. LegalNERo is a manually annotated corpus for named entity recognition in the Romanian legal domain. It provides gold annotations for organizations, locations, persons, time and legal resources mentioned in legal documents. Furthermore, GeoNames identifiers are provided for location entities, when linking was possible. The resource is available in multiple formats, including span-based, token-based and RDF. The Linked Open Data version, in RDF-Turtle format, is available for both download and interrogation using a SPARQL endpoint.

Keywords: Named entity recognition, Linguistic linked data, Romanian language, Corpus

1. Introduction

Named entity recognition is the task of identifying named entities in text [40], like persons, locations, organizations, time, proteins, etc. Starting with 1995, within the MUC-6 conference [14], there have been periodic tasks on various aspects of named entity recognition, focusing on different entity types. For example, for the CoNLL-2003 shared task on language-independent named entity recognition [36], named entities were considered as "phrases that contain the names of persons, organizations and locations". However, this limited approach is not suitable for every domain. In this context, in the biomedical domain, a number of works addressed entities such as genes, proteins, diseases [17], cell type [33], chemicals [13], [19].

In the legal domain, the TREC conference had a dedicated track [7] administered by NIST 1 for evaluating the application of Information Retrieval (IR) methods to e-discovery in the context of the U.S. civil litigation from 2006 until 2011 [24]. The Competition on Legal Information Extraction and Entailment (COL-IEE) [20] run over multiple editions allowed further exploration of tools and algorithms for information extraction in the legal domain.

In the context of the international project "Multilingual Resources for CEF.AT in the legal domain" (MARCELL) 2 a large comparable corpus of legal documents for 7 languages was created [39]. This includes a monolingual sub-corpus for the Romanian language [37]. The Romanian corpus, as well as the other MARCELL corpora, was split at sentence and token level, lemmatized, and annotated at token level. Annotations comprise part-of-speech tags, dependency parsing, named entities and finally the corpus was enriched with IATE terms and EUROVOC descriptors. All these annotations were realized using automatic processes.

Named entities were identified using a general-purpose tool [26], available at that time for the Romanian language, that was not adapted to the legal domain, allowing only entities such as organization, persons, locations and time expressions. The tool was not trained on any legal texts.

Existing Romanian named entity corpora include: RONEC [10], Romanian TimeBank [11] and SiMoNERo [1]. The RONEC corpus contains 26,377 named entities, belonging to 16 different classes. The

*Corresponding author. E-mail: vasile@racai.ro.
1https://www.nist.gov/
2https://marcell-project.eu/
Romanian TimeBank corpus is an annotated parallel corpus for temporal information. This corpus contains 26,635 temporal named entities such as events, instances, signals, etc. SiMoNERo is a gold standard corpus for biomedical domain, manually annotated with four types of domain-specific named entities. SiMoNERo has 14,133 named entities distributed in 4,987 sentences. In this corpus, the NEs are in BIO format. All these corpora contain entities such as organizations, persons, locations, time expressions and biomedical entities. Nevertheless, none of these corpora contains legal texts or legal entities.

This paper presents a manually annotated corpus, comprised of documents from the MARCELL Romanian corpus, with named entities in the legal domain. We considered the classical entity types (organizations, persons, locations and time expressions) as they appear in legal documents and added a new entity type in the form of legal references to documents (such as laws, government decisions, orders, etc.).

The paper is structured as follows: in Section 3 we present the annotation process of the corpus, Section 4 describes different aspects of the corpus such as the annotation levels, the representation of the linked data and statistics of the corpus, Section 5 considers aspects regarding the usage of the RDF version of the corpus, Section 6 presents real use cases for the corpus and we finally conclude in Section 7.

2. Related Work

One of the first papers to discuss named entity recognition in the legal domain is that of [9]. The authors explore named entity recognition and resolution in legal documents such as US case law, deposition, pleadings and other trial documents. The types of entities include judges, attorneys, companies, jurisdictions, and courts.

Cardellino et al. [3] explore using the LKIF3 ontology [16] further mapped to the YAGO4 ontology [35] in order to train a named entity recognizer, classifier and linker. The resulting system is then applied to a corpus comprising judgements of the European Court of Human Rights. The authors recognize that in the legal domain named entities are also names of laws, typified procedures and even concepts. Furthermore, when dealing with human annotators they observe that the classes and subclasses of Document, Organization and Person were the most consistent across annotators.

Glaser et al. [12] explored the suitability of named entity recognition systems in the case of legal contracts. The proposed entity classes are person, organization, location, date, money value, reference, and other. The "reference" entity is based on the work of [21], where references to legal norms are considered.

Leitner et al. [22] introduce a German legal named entity corpus comprising 7 coarse-grained classes which can be expanded into 19 fine-grained classes. In this case, a "person" entity can be classified into a regular person, a judge or a lawyer. Similarly, a "legal norm" entity can be further expanded into law, ordinance or European legal norm.

3. Annotation process

Annotation was performed by 5 human annotators, under the supervision of two senior researchers at the Institute for Artificial Intelligence "Mihai Drăgănescu" of the Romanian Academy (RACAI)5. Annotators followed specific guidelines, inspired in part by the Linguistic Data Consortium (LDC) guidelines for annotation of named entities6.

We considered 5 classes: person (PER), location (LOC), organization (ORG), time (TIME) and legal document references (LEGAL). For person entities, we considered only people names. Titles and honorifics present in text near a person name were not included in the entity. In the case of organizations, they must have some formally established association. Typical examples are businesses, government units and political parties. Locations are defined on a geographical basis and include countries, cities and other geographical areas. References are introduced similar to [21] and the coarse-grained class of [22], without additional sub-classes. Thus, they are references to legal documents such as laws, ordinances, government decisions, etc.

Each annotator was given instructions on how to annotate the documents and then annotated a single document (outside of the corpus). We then discussed any issues or questions the annotators had. Subsequently, a collection of 100 documents was attributed to each annotator. 30 documents (out of the 100) were also

3https://github.com/RinkeHoekstra/lkif-core
4https://yago-knowledge.org/
5https://www.rcai.ro/en/
shared with other two annotators. This aspect was hidden from the annotators during the process but allowed us to later compute inter-annotator agreement (IAA). Throughout the annotation process, we held periodic meetings to discuss any issues.

Corpus and account management for the annotators was realized through the RELATE platform [25]. Actual annotation was handled using the BRAT 7 annotation tool [34], integrated into the RELATE platform. This allowed the annotators to view one document at the time, select the identified entity with the mouse and then associate an entity type to the selected text span.

After the annotation process ended, we were able to compute inter-annotator agreement between each pair of annotators, using Coehn’s Kappa measure. This was accomplished at token level and to lead to an average Kappa of 0.87. Following this result, we further investigated the differences and we were able to detect some recurring mistakes with some of the annotators, such as inclusion of indicative words in the entities (for example "orasul București"/"the city of Bucharest" instead of just "București"/"Bucharest"). An automatic script was created to correct these types of mistakes.

Finally, we constructed an application to manually merge the common annotations into a single file. For each entity, the application shows all the other entities overlapping the same span (if they exist) and allows the user to select the entities that go in the final merged file. The application further makes it easy by highlighting entities found by multiple annotators.

Once all the common annotations were merged we re-computed Coehn’s Kappa measure between the merged corpus and each annotator. This produced an average Kappa of 0.89 and we consider this to be the final result.

4. Corpus description

4.1. Annotation levels

Raw text files were extracted from the Romanian part of the MARCELL corpus. They contain national legislation gathered via crawling from the public Romanian legislative portal8. As described in [39], the texts were extracted from the original HTML format and converted into TXT files. For the purposes of constructing the LegalNERo corpus, we selected a number of 370 documents of similar size, issued in the last two years (2020-2021). We also performed an initial check to make sure the files contain correct Romanian characters (with diacritics) and do not contain tables or other structures that may impact the annotations.

As described in Section 3, annotation was performed using the BRAT tool integrated into the RELATE platform. Thus, the primary annotation output is represented by BRAT-specific files. Each line contains an entity ID, followed by the entity type, the text span (start and end characters) and the actual text. This annotation format allows for multiple annotations in overlapping spans.

We used UDPipe9 on the text files for automatic operations such as tokenization, lemmatization, part of speech tagging and dependency parsing. The resulting files were in CoNLL-U format10. This format can be extended with additional columns by following the CoNLL-U Plus guidelines and adding in a special metadata line the description of the new columns. Using this approach we added a new column "RELATIVE:NE" (the 11th column) for named entity annotations. We mapped the identified annotation text spans to tokens using a BIO annotation format [32]. This implies that each token has an additional annotation with the associated entity, prefixed with one of "B-" (for entity beginning) or "I-" (for a token inside the entity). Tokens that are not part of any entity are annotated with "O" ("outside").

The use of the BIO annotation scheme means there is no direct support for overlapping entities. A token is associated with a single entity type. Therefore, we created two separate token-based annotations, stored in two corresponding folders: one for storing all the entity types, without embedded entities, considering only the largest text spans, and another for storing only person, organization, location and time entities.

Provision of the two CoNLL-U Plus folders means the corpus can be easily used either for legal domain annotations (considering the legal references) or for general annotations (considering the other entity types).

Initial annotations (BRAT and CoNLL-U Plus) were converted to RDF format, specific to applications exploiting linked data. This increases the usability of the corpus as well as allows analysis of the corpus using RDF queries and linking to external databases.

---

7https://brat.nlplab.org/index.html
8http://legislatie.just.ro/
9https://ufal.mff.cuni.cz/udpipe
10https://universaldependencies.org/format.html
4.2. Linked data representation

Already having the text span annotations (in BRAT format) and the token-based annotations (in CoNLL-U Plus format) we were faced with the problem of designing a schema useful for linked data applications. First, we considered the CoNLL-RDF representation [5],[6]. It directly translates from tab-separated CoNLL format to RDF by employing the prefix "conll" together with the column name. It further associates a token representation with the NLP Interchange Format (NIF) ontology [15], by declaring it as a "nif:Word" element, linked to a "nif:Sentence".

We further investigated the POWLA [4] ontology. This was also used by [6] complementary to the NIF ontology. Unlike other approaches, POWLA is not tied to a specific selection of annotation layers, but it is designed to support any kind of text-oriented annotation. For this purpose, POWLA allows specifying "document layers" which then contain the actual annotations. This is very similar to our situation, where we have an annotation layer comprising the text spans associated with entities (corresponding to the BRAT format) and the token-based annotations (corresponding to the CoNLL format).

For the named entity annotations, we employed the NERD ontology [30]. It was previously mentioned in [31] that NERD can be used together with the NIF ontology. It provides classes such as "nerd:Location", "nerd:Person", "nerd:Organization" and "nerd:Time" that can be used for the corresponding entities. Nevertheless, there is no direct specification for legal references.

The European Legislation Identifier (ELI) ontology provides a descriptive framework for structuring metadata of legislative resources and publishing them as linked data. Its primary purpose is to describe relationships between national and European legislative resources. It provides the "eli:LegalResource" class which is defined as a work in a legislative corpus, which applies to acts that have been legally enacted (whether or not they are still in force).

The GeoNames database is integrating geographical data such as names of places in various languages, elevation, population and others from various sources. According to the information available on the website, it contains over 25 million geographical names and consists of over 11 million unique features whereof 4.8 million populated places and 13 million alternate names. We linked location entities with the GeoNames database by using the feature identifiers associated with each GeoNames feature. The annotation was performed automatically and then manually validated.

Table 1 presents the vocabularies used in the corpus. The key concepts and relationships expressed in the dataset are visualized in Figure 1. Some of the vocabularies from Table 1 were used only as part of metadata specification, therefore they do not appear in the diagram. We used the graphical ontology editor OWL-GrEd [2] for constructing the diagram shown in Figure 1.

The corpus is comprised of multiple documents, represented as "powla:Document" elements. Each document is organized into three layers ("powla:DocumentLayer") corresponding to sentences, tokens and named entity text spans. Tokens are linked to the corresponding sentences and offer all the CoNLL-U Plus information, including word form, lemma, universal part-of-speech, language-specific part-of-speech, morphological features, dependency information, named entity type and GeoNames identifier.

The named entities document layer contains elements from the NERD and European Legislation Identifier ontologies. The elements also inherit from "nif:Phrase", thus specifying the beginning and end positions for the associated strings. Furthermore, the GeoNames feature identifier is specified when available for corresponding "nerd:Location" entities.

4.3. Statistics

Since the corpus is available in multiple representations (raw text, span-based annotations, token-based annotations and linked data RDF), we follow each facet and present the corresponding statistics. In Table 2 are presented general corpus statistics. There are

https://www.geonames.org/about.html
Table 1

<table>
<thead>
<tr>
<th>Prefix</th>
<th>Name</th>
<th>URI</th>
</tr>
</thead>
<tbody>
<tr>
<td>nif</td>
<td>NLP Interchange Format (NIF)</td>
<td><a href="http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#">http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#</a></td>
</tr>
<tr>
<td>powla</td>
<td>POWLA Ontology</td>
<td><a href="http://purl.org/powla/powla.owl#">http://purl.org/powla/powla.owl#</a></td>
</tr>
<tr>
<td>nerd</td>
<td>NERD Ontology</td>
<td><a href="http://nerd.eurecom.fr/ontology#">http://nerd.eurecom.fr/ontology#</a></td>
</tr>
<tr>
<td>conllu</td>
<td>CoNLL-U tabular format</td>
<td><a href="https://universaldependencies.org/format.html#">https://universaldependencies.org/format.html#</a></td>
</tr>
<tr>
<td>conllup</td>
<td>CoNLL-U Plus format</td>
<td><a href="https://universaldependencies.org/ext-format.html#">https://universaldependencies.org/ext-format.html#</a></td>
</tr>
<tr>
<td>eli</td>
<td>European Legislation Identifier (ELI)</td>
<td><a href="http://data.europa.eu/eli/ontology#">http://data.europa.eu/eli/ontology#</a></td>
</tr>
<tr>
<td>gn</td>
<td>GeoNames</td>
<td><a href="http://www.geonames.org/ontology#">http://www.geonames.org/ontology#</a></td>
</tr>
<tr>
<td>rdf</td>
<td>RDF</td>
<td><a href="http://www.w3.org/1999/02/22-rdf-syntax-ns#">http://www.w3.org/1999/02/22-rdf-syntax-ns#</a></td>
</tr>
<tr>
<td>rdfs</td>
<td>RDF Schema</td>
<td><a href="http://www.w3.org/2000/01/rdf-schema#">http://www.w3.org/2000/01/rdf-schema#</a></td>
</tr>
<tr>
<td>owl</td>
<td>OWL</td>
<td><a href="http://www.w3.org/2002/07/owl#">http://www.w3.org/2002/07/owl#</a></td>
</tr>
<tr>
<td>dcat</td>
<td>DCAT 2 Vocabulary</td>
<td><a href="http://www.w3.org/ns/dcat#">http://www.w3.org/ns/dcat#</a></td>
</tr>
<tr>
<td>dct</td>
<td>DCMI Metadata Terms</td>
<td><a href="http://purl.org/dc/terms/">http://purl.org/dc/terms/</a></td>
</tr>
<tr>
<td>skos</td>
<td>SKOS Simple Knowledge Organization System</td>
<td><a href="http://www.w3.org/2004/02/skos/core#">http://www.w3.org/2004/02/skos/core#</a></td>
</tr>
<tr>
<td>xsd</td>
<td>XSD</td>
<td><a href="http://www.w3.org/2001/XMLSchema#">http://www.w3.org/2001/XMLSchema#</a></td>
</tr>
<tr>
<td>prov</td>
<td>PROV</td>
<td><a href="http://www.w3.org/ns/prov#">http://www.w3.org/ns/prov#</a></td>
</tr>
<tr>
<td>foaf</td>
<td>FOAF</td>
<td><a href="http://xmlns.com/foaf/0.1/">http://xmlns.com/foaf/0.1/</a></td>
</tr>
<tr>
<td>pav</td>
<td>PAV - Provenance, Authoring and Versioning</td>
<td><a href="http://pav-ontology.github.io/pav/">http://pav-ontology.github.io/pav/</a></td>
</tr>
</tbody>
</table>

Table 3 presents the distribution of the annotated tokens conllup files of the corpus. It can be seen legal documents references class (LEGAL) contains 265,335 tokens distributed in 370 documents. The average length of a sentence is 32.02 tokens, which is above the 16.06 tokens/sentence, the average sentence length in ROMBAC [18], a balanced Romanian corpus, containing legal, news, fiction, biographical and medical texts.
5. Using the RDF version of LegalNERo

The LegalNERo corpus [29] is available for download from the Zenodo platform\(^{13}\) as a single archive containing all the different representations described in this paper, stored into dedicated folders. In the "rdf" folder there is a single file containing all the triples in RDF-Turtle format. In addition to the download option, a SPARQL endpoint\(^{14}\) is available from the RELATE platform, hosted by the Institute for Artificial Intelligence "Mihai Drăganescu" of the Romanian Academy.

The SPARQL endpoint is offered via an Apache Jena Fuseki server\(^{15}\). A simple graphical query interface, provided by the same server implementation, is available\(^{16}\). This allows a user to interact with the LegalNERo corpus by means of SPARQL queries and visualize results in table format. Figure 2 presents a SPARQL query to list legal references found in the corpus. It also shows the user interface displaying data in table form. This type of queries is useful in creating gazetteer resources specific to named entity recognition systems. The query can easily be adapted to produce lists of different types of entities.

Additional query examples are provided in Figures 3 and 4. In the first case, the SPARQL query allows listing of location entities with associated GeoNames identifiers. The result will contain only those entities that have a GeoNames identifier. Figure 4 makes use of the token layer available in the corpus and displays organization entities, tokenized, with the associated UPOS tags concatenated. In this example, only entities comprised of up to 5 tokens are considered. This type of query is useful in finding patterns associated with the named entities present in the corpus. Patterns can then be used with simpler pattern-based NER systems, such as Stanford RegexNER\(^{17}\), available from the Stanford CoreNLP [23] package.

6. Corpus usage

In accordance with the multiple facets of the LegalNERo corpus, we developed two NER models: one for all the entities and one dealing only with persons, locations, organizations and time entities. These models are based on a recurrent neural network with a final CRF layer, trained using the NeuroNER\(^{18}\) toolkit [8]. To improve the model’s performance, we used pre-trained word embeddings [28] representations trained on the Representative Corpus of Contemporary Romanian Language (CoRoLa) [38]. The models were integrated in the RELATE [25] platform and are available for online interrogation and download\(^{19}\), together with the used word embeddings\(^{20}\).

In the context of the "Curated Multilingual Language Resources for CEF.AT" (CURLICAT) project\(^{21}\), we aim to develop an anonymization solution for Romanian language. Part of this solution, we need the identification of named entities present in a given document. Of course the purpose is not to anonymize legislation (which does not require anonymization), but we consider that the NER models developed based on the LegalNERo corpus, have the ability to complement other models developed on more general corpora and rule-based approaches. A current prototype of the anonymization solution is available through the RELATE platform and can be used online\(^{22}\).

---

\(^{13}\)https://doi.org/10.5281/zenodo.4772094
\(^{14}\)https://relate.racai.ro/datasets/legalnero/sparql
\(^{15}\)https://jena.apache.org/documentation/fuseki2/
\(^{16}\)https://relate.racai.ro/datasets/dataset.html?tab=query&ds=legalnero
\(^{17}\)https://nlp.stanford.edu/software/regexner.html
\(^{18}\)http://neuroner.com/
\(^{19}\)https://relate.racai.ro/index.php?path=mer/demo
\(^{20}\)http://relate.racai.ro/index.php?path=corola/we
\(^{21}\)https://curlicat-project.eu/
\(^{22}\)https://relate.racai.ro/index.php?path=roanon/anonymize
Table 3
NEs statistics on conllup files (token-based)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>LEGAL</th>
<th>PER</th>
<th>LOC</th>
<th>ORG</th>
<th>TIME</th>
<th>GEO</th>
<th>TOTAL tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td>conllup_PER_LOC_ORG_TIME</td>
<td>-</td>
<td>2,099</td>
<td>3,144</td>
<td>22,328</td>
<td>8,422</td>
<td>1,411</td>
<td>35,993</td>
</tr>
<tr>
<td>conllup_LEGAL_PER_LOC_ORG_TIME</td>
<td>24,687</td>
<td>2,099</td>
<td>3,144</td>
<td>19,477</td>
<td>5,121</td>
<td>1,411</td>
<td>54,528</td>
</tr>
</tbody>
</table>

Table 4
NEs statistics on .ann files (span-based)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>LEGAL</th>
<th>PER</th>
<th>LOC</th>
<th>ORG</th>
<th>TIME</th>
<th>GEO</th>
<th>TOTAL NEs</th>
</tr>
</thead>
<tbody>
<tr>
<td>ann_PER_LOC_ORG_TIME</td>
<td>-</td>
<td>914</td>
<td>2,276</td>
<td>6,209</td>
<td>4,643</td>
<td>-</td>
<td>14,042</td>
</tr>
<tr>
<td>ann_LEGALL_PER_LOC_ORG_TIME</td>
<td>3,387</td>
<td>914</td>
<td>2,276</td>
<td>4,824</td>
<td>2,213</td>
<td>-</td>
<td>13,614</td>
</tr>
<tr>
<td>ann_LEGAL_PER_LOC_ORG_TIME_overlap</td>
<td>3,387</td>
<td>914</td>
<td>2,276</td>
<td>6,209</td>
<td>4,643</td>
<td>-</td>
<td>17,429</td>
</tr>
</tbody>
</table>

Fig. 2. SPARQL query to list legal references and corresponding result.

7. Conclusions and future work

This paper introduced the LegalNERo corpus. It is a manually annotated corpus for named entity recognition considering legal references in the Romanian language and also enhanced with GeoNames identifiers. The corpus represents a subset of the larger MARCELL [39] parallel legislative corpus, therefore for certain applications these corpora could be used together. LegalNERo provides also annotations for sub-entities present inside the legal references. This can be exploited to allow usage of the corpus for training more classic NER systems considering only persons, locations, organizations and time entities.

We offer the corpus under a Creative Commons license (CC BY-ND 4.0). The downloadable version comes with different perspectives on the data, including span-based annotations, token-based annotations and RDF-Turtle format. We further offer a SPARQL endpoint allowing online interaction with the corpus.
Finally, the corpus was integrated in the Linked Open Data Cloud \(^{23}\).

Our aim is to further use this corpus to construct an improved NER system for the legal domain, in the Romanian language. Currently available models, presented in Section 6, achieved an average F1 score of 84\% (considering all entities) and 84.70\% (without the legal reference entity type). This already presents an improved performance compared to the one \([26]\) previously used to automatically annotate the Romanian Legal Corpus \([37]\) (part of the larger MARCELL corpus). Nevertheless, considering additional techniques, such as word embeddings combinations \([27]\) could prove beneficial in improving the overall performance.

Acknowledgements

Part of this work was conducted in the context of the "Curated Multilingual Language Resources for CEFAT" (CURLICAT) project, CEF-TC-2019-1 – Automated Translation grant agreement number INEA/CEF/ICT/A2019/1926831.

References

\(^{23}\)https://lod-cloud.net/dataset/racai-legalnero

\[\text{SELECT ?id ?ent ?geo where }\]
\[
<table>
<thead>
<tr>
<th>?id</th>
<th>?ent</th>
<th>?geo</th>
</tr>
</thead>
</table>
| id | n:rd:location | .
| id | p:owl:location | .
| id | g:ontology:feature | .
| \} LIMIT 25

Fig. 3. SPARQL query to list location entities with associated GeoNames identifiers.

\[\text{prefix : } <\text{http://racai.ro/legalnero}>\]
\[\text{prefix p:owl: } <\text{http://purl.org/owl/owl2/owl#}>\]
\[\text{prefix n:rd: } <\text{http://nord.eurecom.fr/ontology#}>\]
\[\text{prefix g:ontology: } <\text{http://www.geonames.org/ontology#}>\]

\[\text{prefix : } <\text{http://racai.ro/legalnero}>\]
\[\text{prefix p:owl: } <\text{http://purl.org/owl/owl2/owl#}>\]
\[\text{prefix n:rd: } <\text{http://nord.eurecom.fr/ontology#}>\]
\[\text{prefix g:ontology: } <\text{http://www.geonames.org/ontology#}>\]

\[\text{prefix : } <\text{http://racai.ro/legalnero}>\]
\[\text{prefix p:owl: } <\text{http://purl.org/owl/owl2/owl#}>\]
\[\text{prefix n:rd: } <\text{http://nord.eurecom.fr/ontology#}>\]
\[\text{prefix g:ontology: } <\text{http://www.geonames.org/ontology#}>\]

\[\text{prefix : } <\text{http://racai.ro/legalnero}>\]
\[\text{prefix p:owl: } <\text{http://purl.org/owl/owl2/owl#}>\]
\[\text{prefix n:rd: } <\text{http://nord.eurecom.fr/ontology#}>\]
\[\text{prefix g:ontology: } <\text{http://www.geonames.org/ontology#}>\]


WHERE {
    ?id conllu:NE "B-ORG" .
    OPTIONAL { ?id conllu:HEAD1 ?id2 .
                ?id2 conllu:NE "I-ORG" .
                ?id2 conllu:FORM ?word2 .
    }
    OPTIONAL { ?id conllu:HEAD2 ?id3 .
                ?id3 conllu:NE "I-ORG" .
                ?id3 conllu:FORM ?word3 .
    }
                ?id4 conllu:NE "I-ORG" .
    }
    OPTIONAL { ?id conllu:HEAD5 ?id5 .
                ?id5 conllu:NE "I-ORG" .
                ?id5 conllu:FORM ?word5 .
    }
}

 BIND(CONCAT(STR(?word1),"/",STR(?pos1)) AS ?wp1).
 BIND(CONCAT(STR(?word2),"/",STR(?pos2)) AS ?wp2).
 BIND(CONCAT(STR(?word3),"/",STR(?pos3)) AS ?wp3).
 BIND(CONCAT(STR(?word4),"/",STR(?pos4)) AS ?wp4).
 BIND(CONCAT(STR(?word5),"/",STR(?pos5)) AS ?wp5).

Fig. 4. SPARQL query to list organization entities at token level (comprising up to 5 tokens) with associated UPOS tags.


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