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

Vasile Păiș [a,*], Maria Mitrofan [a], Carol Luca Gasan [a], Alexandru Ianov [a], Corvin Ghiță [a], Vlad Silviu Coneschi [a], and Andrei Onuț [a]

[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 expressions and legal resources mentioned in legal documents. Furthermore, GeoNames identifiers are provided. The resource is available in multiple formats, including span-based, token-based and RDF. The Linked Open Data version is available for both download and querying using SPARQL.

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

1. Introduction

Named entity recognition (NER) is the task of identifying named entities (NE) in text [45], such as persons, locations, organizations, and proteins. Starting in 1995, within the MUC-6 conference [14], there have been periodic tasks on various aspects of NER. For example, for the CoNLL-2003 shared task [40], NEs were considered "phrases that contain the names of persons, organizations and locations". In this context, in the biomedical domain, a number of works have addressed entities such as genes, proteins, diseases [17], cell types [36], and chemicals [13], [18].

In the legal domain, the TREC conference had a dedicated track [7] 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 [26]. The Competition on Legal Information Extraction and Entailment (COLIEE) [19] ran over multiple editions allowed further exploration of tools and algorithms for information extraction in the legal domain.

This paper presents a manually annotated corpus comprising a subset of documents from the MARCELL Romanian corpus, with NEs 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).

The paper is structured as follows: in Section 2 we present related work; in Section 3 we introduce the annotation process; Section 4 describes different aspects of the corpus, such as the annotation levels, the representation of the linked data and statistics; Section 5 considers the usage of the RDF version; Section 6 presents use cases; Section 7 discusses quality and stability, and finally, we conclude in Section 8.

2. Related Work

In the context of the EU project "Multilingual Resources for CEF.AT in the legal domain" (MARCELL)[1] a large comparable corpus of legal documents for seven languages was created [43]. Comparable corpora can be exploited for improving machine translation [24] and parallel sentences can be identified automatically, potentially containing NEs. The Romanian sub-corpus [41], as well as the other MARCELL corpora, was split at sentence and token levels, lemma-

[*] Corresponding author. E-mail: vasile@racai.ro.

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

tized, annotated at token level (part-of-speech, dependency parsing, NEs), and finally the corpus was enriched with IATE terms and EUROVOC descriptors. All annotations in the Romanian MARCELL corpus were realized using automatic processes.

Existing Romanian NE corpora include: RONEC [10], Romanian TimeBank [11] and SiMoNERo [1]. The RONEC corpus contains 26,377 NEs, belonging to 16 different classes. The Romanian TimeBank is an annotated parallel corpus for temporal information. SiMoNERo is a gold standard corpus for the biomedical domain, manually annotated with four types of domain-specific NEs. SiMoNERo has 14,133 NEs distributed in 4,987 sentences. All these corpora contain entities such as organizations, persons, locations. None of these corpora contains legal texts or legal entities.

Dozier et al. [9] explore NER in legal documents such as US case law, depositions, pleadings and other trial documents. The types of entities include judges, attorneys, companies, jurisdictions, and courts.

Cardellino et al. [3] explored using the LKIF ontology [16] further mapped to the YAGO ontology [39] to train a NE recognizer, classifier and linker. The resulting system is applied to a corpus comprising judgements of the European Court of Human Rights. The authors recognize that in the legal domain NEs 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 NER systems in the case of legal contracts. The entity classes are person, organization, location, date, money value, reference, and other.

Leitner et al. [22] introduced a German legal NE corpus comprising seven coarse-grained and 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 expanded into law, ordinance or European legal norm.

In MARCELL, Romanian NEs were identified using a general-purpose tool [28], 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, but since no legal-domain tool was available it was used on this corpus. Thus, the need for a legal-domain NER corpus and system became apparent, which led us to the development of the LegalNERo corpus.

3. Annotation process

Annotation was performed by five human annotators under the supervision of two senior researchers at the Institute for Artificial Intelligence "Mihai Drăgănescu" of the Romanian Academy (RACAI). Annotators followed specific guidelines[2], inspired by the Linguistic Data Consortium (LDC) guidelines[3].

We considered five classes: person (PER), location (LOC), organization (ORG), time expressions (TIME) and legal document references (LEGAL). For person entities, we considered only person names. Titles and honorifics present in text near a person name were not included. Organizations must have some formally established association. Typical examples are businesses, government units and political parties. Locations are defined on a geographical basis. 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. Even though we only annotated the legal reference coarse-grained class, most of these entities can be mapped to fine-grained classes using automated processes, employing other linked data resources (Section 5).

Each annotator was given instructions on how to annotate the documents and then annotated a single document. We discussed issues or questions the annotators had. Subsequently, a collection of 100 documents was attributed to each annotator. Thirty documents (out of the 100) were shared with two other annotators, allowing us to compute inter-annotator agreement (IAA).

Corpus and account management for the annotators was realized through the RELATE platform [32]. Actual annotation was handled using BRAT [37], integrated into RELATE. This allowed the annotators to view one document, select the identified entity and then associate an entity type with the selected span.

After the annotation process, we computed IAA between each pair of annotators, using Cohen's Kappa. This was accomplished at token level and led to an average Kappa of 0.87. We further investigated the differences and were able to detect some recurring mistakes, such as the inclusion of indicative words in the entities (for example "orașul București"/"the city of

[2] https://relate.racai.ro/resources/legalnero/legalnero_annotation_guide.pdf

[3] https://www.ldc.upenn.edu/sites/www.ldc.upenn.edu/files/english-edt-v4.2.6.pdf

Bucharest" instead of just "București"/"Bucharest"). A script was created to correct these mistakes.

Finally, we constructed an application to merge the annotations into a single file. For each entity, the application shows all other entities overlapping the same span and allows the user to select the entities that go in the merged file. To aid the user, the application highlights entities found by multiple annotators.

Once all common annotations were merged we recomputed Cohen'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. According to Landis and Koch [20], a Kappa value greater than 0.81 is indicative of an "almost perfect" agreement. The remaining disagreements account for mistakes made by individual annotators, such as missing an entity or a sub-entity. This is particularly reflected in potentially ambiguous situations such as the legal reference "Regulamentul CE nr. 765/2008" (en. "Council Regulation EC No 765/2008"). In this case certain annotators identified "CE" as an organization sub-entity, while others did not. In a few cases the end-of-sentence punctuation coincides with the dot indicating an abbreviation. Some annotators included the punctuation in the organization entity abbreviation, while others considered it to be sentence punctuation and did not include it (for example: "S.R.L." vs "S.R.L" or "AFER." vs "AFER").

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 by crawling from the public Romanian legislative portal[4]. As described in [43], the texts were extracted from HTML and converted into TXT files. For the LegalNERo corpus, we selected 370 documents of similar size, published between 2020-2021. We performed an initial check to ensure that the files contain correct Romanian characters (with diacritics) and do not contain tables or other structures that may impact the annotations.

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

We used UDPipe [38] on the text files for operations such as tokenization, lemmatization, part of speech tagging and dependency parsing. The resulting files were in CoNLL-U format[5]. We added a new column "RELATE:NE" (the 11th column) for NE annotations. We mapped the identified annotation text spans to tokens using a BIO notation format [35]. The associated entity annotation is 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 BIO scheme means that there is no support for overlapping entities. A token is associated with a single entity type. Therefore, we created two separate token-based annotations, stored in two folders: one for storing all entity types, without embedded entities, considering only the largest text spans, and another for storing only person, organization, and location entities and time expressions. Provision of the two folders means the corpus can be used either for legal domain annotations (considering the legal references) or for general annotations (the other entity types).

Initial annotations (BRAT and CoNLL-U Plus) were converted to RDF, specific to applications exploiting linked data. This increases the usability of the corpus and allows analysis using RDF queries and linking to external databases. Location entities can be resolved using geographical databases, such as GeoNames. The annotation is available in both CoNLL-U Plus files (column 12, "RELATE:GEONAMES") and in the RDF representation (Figures 1 and 2).

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. We considered multiple types of ontology-based representations:

- CoNLL-RDF representation [5],[6]. It directly translates from tab-separated CoNLL format to RDF by employing the prefix "conll" with the column name. It associates a token representation with the NLP Interchange Format (NIF) [15].

[4] http://legislatie.just.ro/

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

```
T1      LEGAL 2 36      LEGE nr. 185 din 17 octombrie 2019
T3      LEGAL 57 77     Legii nr. 227 / 2015
T5      LEGAL 86 98     Codul fiscal
```

Fig. 1. Span-based annotation in ann format.

```
# global.columns = ID FORM LEMMA UPOS XPOS FEATS HEAD DEPREL DEPS MISC RELATE:NE RELATE:
    GEONAMES
# sent_id = 1
# text = LEGE nr. 185 din 17 octombrie 2019 pentru modificarea Legii nr. 227 / 2015 privind
    Codul fiscal şi pentru modificarea art. 9 din Ordonanţa Guvernului nr. 105 / 199
1       LEGE    lege    VERB    Vmip3s  Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin
    0       root    _       SpacesBefore=\r\n       B-LEGAL _
2       nr.     nr.     NOUN    Yn      Abbr=Yes        1       obj     _       _       I-LEGAL
    _
3       185     185     NUM     Mc-p-d  Number=Plur|NumForm=Digit|NumType=Card  2       nummod
    _       _       I-LEGAL _
4       din     din     ADP     Spsa    AdpType=Prep|Case=Acc   6       case    _       _
    I-LEGAL _
5       17      17      NUM     Mc-p-d  Number=Plur|NumForm=Digit|NumType=Card  6       nummod
    _       _       I-LEGAL _
6       octombrie       octombrie       NOUN    Ncms-n  Definite=Ind|Gender=Masc|Number=Sing
    2       nmod    _       _       I-LEGAL _
7       2019    2019    NUM     Mc-p-d  Number=Plur|NumForm=Digit|NumType=Card  2       nummod
    _       SpacesAfter=\r\n        I-LEGAL _
```

Fig. 2. Token-based annotation in CoNLL-U Plus format.

- POWLA ontology [4], designed to support any kind of text-oriented annotation, allowing for "document layers" that contain the annotations.
- NERD ontology [34] provides classes such as "nerd:Location", "nerd:Person", "nerd:Organization" and "nerd:Time".
- European Legislation Identifier (ELI) ontology provides a framework for structuring metadata of legislative resources and publishing them as linked data. It provides the "eli:LegalResource" class which is defined as a work in a legislative corpus, applying to acts that have been legally enacted (whether or not they are still in force).
- GeoNames integrates geographical data such as the names of places in various languages, elevation, population and others. We linked location entities with GeoNames by using the feature identifiers. 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 3, and a description is given in Table 2. Some of the vocabularies from Table 1 were used only as part of metadata specification. Therefore, they do not appear in the diagram. A complete example is included in Appendix A.

The corpus comprises multiple documents, represented as "powla:Document" elements. Each document is organized into three layers, corresponding to sentences, tokens and NE text spans. Each object's layer is indicated by the "powla:hasLayer" attribute and the layer is linked to a document using the "powla:hasDocument" attribute. Tokens are linked to sentences, using the "nif:sentence" and "powla:hasParent". The sentences document layer contains objects of type "nif:Sentence". To maintain the order of sentences, the "nif:nextSentence" and "nif:previousSentence" relations are used. Furthermore, the "nif:firstWord" and "nif:lastWord" attributes are employed to allow direct access to the first and last tokens of a sentence.

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 ("nif:begin Index") and end positions ("nif:endIndex") for associated strings. Furthermore, the GeoNames feature identifier ("gn:Feature") is specified when available for corresponding "nerd:Location" entities.

Table 1
Used vocabularies

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

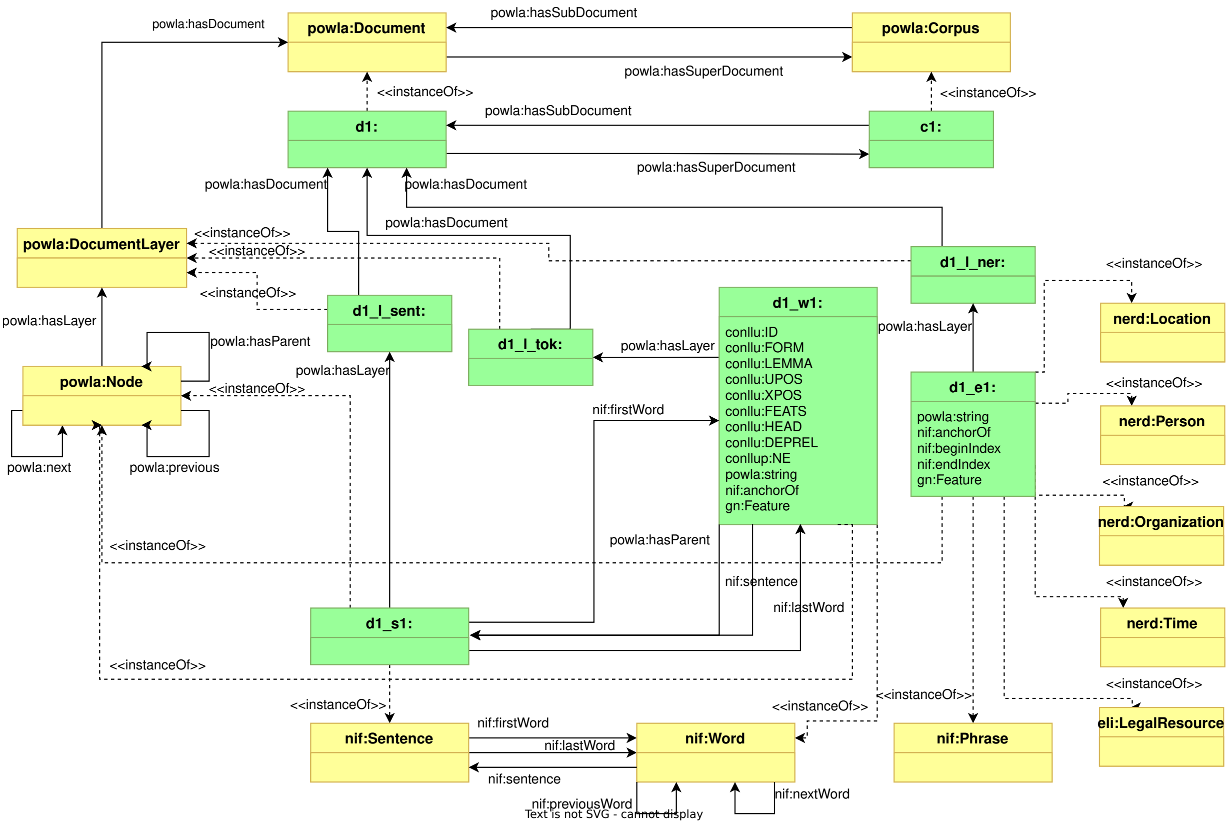

Fig. 3. Key concepts and relationships.

An example encoding in the LegalNERo corpus is given in Appendix A. where a token d1_s1_1 of type Word for the "nif" ontology and of type Node for the "powla" ontology is described. The example given in Appendix A corresponds to the sample ann and CoNLL-U Plus annotations given in Figures 1 and 2.

4.3. *Statistics*

Since the corpus is available in multiple representations, we follow each facet and present the corresponding statistics (see Table 3). Table 4 presents the distribution of the annotated tokens. The legal references class (LEGAL) contains 2,851 organizations (ORG) and 3,301 time (TIME) expressions. This format of the corpus also contains 1,411 GeoNames identifiers linked with the locations (LOC), where there is a complete overlap between the NE and GeoNames identifier. Table 5 gives the span-based statistics.

5. Using the RDF version of LegalNERo

The LegalNERo corpus [33] is available for download from Zeonodo[6] as a single archive containing all representations described in this paper. In the "rdf" folder there is a file containing all triples in RDF-Turtle. A SPARQL endpoint[7] is available from the RELATE platform, offered via Apache Jena Fuseki, with a graphical query interface[8]. Figure 4 presents a SPARQL query to list legal references. This type of queries is useful in creating gazetteer resources.

Additional examples are provided in Figures 5 and 6. 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 6 makes use of the token layer and displays organization entities, tokenized, with associated UPOS tags concatenated. In this example, only entities comprised of up to five tokens are considered. This type of query is useful in finding patterns associated with the NEs present in the corpus. Patterns can then be used with simpler pattern-based NER systems, such as Stanford RegexNER, available from the Stanford CoreNLP [23] package.

```
PREFIX : <http://racai.ro/legalnero>
PREFIX powla: <http://purl.org/powla/powla.
    owl#>
PREFIX eli: <http://data.europa.eu/eli/
    ontology#>
SELECT ?id ?ent
WHERE {
    ?id a eli:LegalResource .
    ?id powla:string ?ent .
} LIMIT 5
```

```
:d338_e16   "Normelor metodologice de
    aplicare a Legii nr. 232/2016"
:d107_e7    "Referatul de aprobare al Direcţ
    iei relaţii cu presa, afaceri europene şi
     relaţii internaţionale nr. S8
    4.536/4.04.2019"
:d85_e20    "Legea nr. 13/2008"
:d291_e1    "ORDIN nr. 625 din 25 aprilie
    2019"
:d319_e1    "ORDIN nr. 1.155 din 9 august
    2019"
```

Fig. 4. SPARQL query to list legal references and result.

The advantage of having a linked data resource comes from the ability to interlink it with other resources. Recently, a number of other Romanian resources were converted to linked data format [2] and are available on the same SPARQL server as LegalNERo. This enables complex federated queries to be performed across multiple resources. An example is the refinement of the legal reference class into multiple fine-grained classes. Considering the classes "law", "decision", "government decision", and "government ordinances", they correspond to the presence in the entity of different forms of the Romanian words "lege", "decizie", "hotărâre", and "ordonanță". Thus, we can exploit the RoLEX lexicon to obtain the word forms associated with the words and then use these to classify the existing entities (see Figure 7).

6. Corpus usage

In accordance with the LegalNERo corpus, we developed two NER models [27]: one for all the entities and expressions, and another dealing only with persons, locations, organizations and time expressions. These models are based on a recurrent neural network with a final CRF layer, trained using NeuroNER [8]. To improve performance, we used pre-trained word embeddings [31] representations trained on the Represen-

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

[7] https://relate.racai.ro/datasets/legalnero/query

[8] https://relate.racai.ro/datasets/dataset.html?tab=query&ds=/legalnero

Table 2

Brief description of the key concepts and relations

Ontology	**Concept/Relation**	**Brief Description**
POWLA	Corpus	Represents general corpus information
POWLA	Document	Individual documents
POWLA	DocumentLayer	Represents the different views associated with a document: sentences, tokens, named entities
POWLA	Node	Individual sentences, tokens, entities
POWLA	hasSuperDocument	Links a document to the corpus
POWLA	hasSubDocument	Links the corpus to the document
POWLA	hasDocument	Links a layer to the corresponding document
POWLA	hasLayer	Links a node (sentence, token, entity) to the corresponding document layer
POWLA	hasParent	Links a token to a sentence
NIF	Sentence	Represents a sentence
NIF	Word	Represents a token
NIF	Phrase	A named entity
NIF	firstWord	The first word in a sentence
NIF	lastWord	The last word in a sentence
NIF	sentence	Links a word to the corresponding sentence
NIF	nextWord	The next word in a sentence
NIF	previousWord	The previous word in a sentence
NERD	Location	Entity of type Location
NERD	Person	Entity of type Person
NERD	Organization	Entity of type Organization
NERD	Time	Expression of type Time
ELI	LegalResource	Expression of type Legal reference

Table 3

Key statistics

Category	**Value**
Text Files	370
Tokens	265,335
Sentences	8,284
Unique lemma	12,887
Triples	5,761,781

tative Corpus of Contemporary Romanian Language (CoRoLa) [42]. The models were integrated in RELATE [29, 32] and are available for online usage and download[9], together with the embeddings[10].

```
PREFIX : <http://racai.ro/legalnero>
PREFIX powla: <http://purl.org/powla/powla.
    owl#>
PREFIX gn: <http://www.geonames.org/ontology#
    >
PREFIX nerd: <http://nerd.eurecom.fr/ontology
    #>
SELECT ?id ?ent ?geo
WHERE {
    ?id a nerd:Location .
    ?id powla:string ?ent .
    ?id gn:Feature ?geo .
} LIMIT 5
```

Fig. 5. SPARQL query to list location entities with associated GeoNames identifiers.

In the context of the "Curated Multilingual Language Resources for CEF.AT" (CURLICAT) project [44], we aim to develop an anonymization solution for Romanian. Part of this solution, we need the identification of NEs. Of course the purpose is not to anonymize legislation, but we consider that the NER models developed based on LegalNERo, 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 in RELATE[11].

7. Quality and stability

The underlying structures used to construct the linked data representation of LegalNERo are stable in

[9] https://relate.racai.ro/index.php?path=ner/demo

[10] http://relate.racai.ro/index.php?path=corola/we

[11] https://relate.racai.ro/index.php?path=roanon/anonymize

Table 4
NEs statistics on conllup files (token-based)

Dataset	LEGAL	PER	LOC	ORG	TIME	GEO	TOTAL tokens
conllup_PER_LOC_ORG_TIME	-	2,099	3,144	22,328	8,422	1,411	35,993
conllup_LEGAL_PER_LOC_ORG_TIME	24,687	2,099	3,144	19,477	5,121	1,411	54,528

Table 5
NEs statistics on .ann files (span-based)

Dataset	LEGAL	PER	LOC	ORG	TIME	GEO	TOTAL NEs
ann_PER_LOC_ORG_TIME	-	914	2,276	6,209	4,643	-	14,042
ann_LEGALL_PER_LOC_ORG_TIME	3,387	914	2,276	4,824	2,213	-	13,614
ann_LEGAL_PER_LOC_ORG_TIME_overlap	3,387	914	2,276	6,209	4,643	-	17,429

```
PREFIX : <http://racai.ro/legalnero>
PREFIX nif: <http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core#>
PREFIX conllu: <https://universaldependencies.org/format.html#>
PREFIX conllup: <https://universaldependencies.org/ext-format.html#>
SELECT ?id ?wp1 ?wp2 ?wp3 ?wp4 ?wp5
WHERE {
  ?id conllup:NE "B-ORG". ?id conllu:FORM ?w1. ?id conllu:UPOS ?p1.
  OPTIONAL{ ?id nif:nextWord ?i2. ?i2 conllup:NE "I-ORG". ?i2 conllu:FORM ?w2. ?i2 conllu:UPOS
    ?p2.
    OPTIONAL{ ?i2 nif:nextWord ?i3. ?i3 conllup:NE "I-ORG". ?i3 conllu:FORM ?w3. ?i3 conllu:
    UPOS ?p3.
      OPTIONAL{ ?i3 nif:nextWord ?i4. ?i4 conllup:NE "I-ORG". ?i4 conllu:FORM ?w4. ?i4 conllu:
    UPOS ?p4.
        OPTIONAL{ ?i4 nif:nextWord ?i5. ?i5 conllup:NE "I-ORG". ?i5 conllu:FORM ?w5. ?i5 conllu
    :UPOS ?p5.
  } } } }
  BIND(CONCAT(STR(?w1),"/",STR(?p1)) as ?wp1) .
  BIND(CONCAT(STR(?w2),"/",STR(?p2)) as ?wp2) .
  BIND(CONCAT(STR(?w3),"/",STR(?p3)) as ?wp3) .
  BIND(CONCAT(STR(?w4),"/",STR(?p4)) as ?wp4) .
  BIND(CONCAT(STR(?w5),"/",STR(?p5)) as ?wp5) .
} LIMIT 5
```

Fig. 6. SPARQL query to list organization entities at token level (comprising up to five tokens) with associated UPOS tags.

the sense that we do not plan to change any of the classes or available attributes. Additions will be in the form of new data, following the same structure. Furthermore, any future extensions with regard to available NEs will follow the same general structure, with new classes being added, without removing existing information. Finally, the versioning system provided by Zenodo allows retrieval of the corpus at any point in time and ensures its continuous availability.

Metrics for assessing linked data quality have been proposed [25, 46], considering both the data content and the metadata. The embedded metadata provides an indication of the dataset provenance, improving also the trustworthiness of the dataset. We offer the corpus under a Creative Commons licence (CC BY-ND 4.0). This information is also part of the metadata (machine-readable) and indicated in human-readable format on the corpus download page.

Before release, the dataset was checked for syntactic validity and no errors have been found. Furthermore, there are no consistency issues with regard to the data structure (no misplaced classes or properties, no inconsistent values). With regard to the actual annotations, the only inconsistencies may arise from the IAA (Section 3). The SPARQL endpoint is offered from a research server, shared with other projects, and thus performance metrics (such as low latency, high throughput or scalability) were not considered.

```
PREFIX : <http://racai.ro/legalnero>
PREFIX powla: <http://purl.org/powla/powla.
    owl#>
PREFIX eli: <http://data.europa.eu/eli/
    ontology#>
PREFIX ontolex: <http://www.w3.org/ns/lemon/
    ontolex>
SELECT DISTINCT ?canonicalForm ?entString
WHERE {
  SERVICE <https://relate.racai.ro/datasets/
    rolex/sparql> {
    ?idCanonical ontolex:writtenRep ?
    writtenRep .
    FILTER (?writtenRep in
      ( "lege"@ro, "decizie"@ro, "hotărâre"
    @ro, "ordonanţă"@ro)).
    ?idEntry ontolex:canonicalForm ?
    idCanonical .
    ?idCanonical ontolex:writtenRep ?
    canonicalForm.
    ?idEntry ontolex:lexicalForm ?idForm.
    ?idForm ontolex:writtenRep ?formWritten.
    FILTER (lang(?formWritten)="ro").
    BIND (str(?formWritten) as ?formString).
  }
  ?idEnt a eli:LegalResource .
  ?idEnt powla:string ?entString.
  FILTER (REGEX(str(?entString),CONCAT("\\b"
    ,?formString,"\\b"),"i")).
} LIMIT 5
```

Fig. 7. Classify legal reference entities into fine-grained classes.

8. Conclusion and future work

This paper introduced the LegalNERo corpus, a manually annotated corpus for NER considering legal references in the Romanian language, providing span-based annotations, token-based annotations and RDF-Turtle format. The corpus represents a subset of the MARCELL [43] legislative corpus; for certain applications these corpora could be used together. LegalNERo also provides 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 expressions. We further offer a SPARQL endpoint. Finally, the corpus was integrated into the Linked Open Data Cloud[12].

Our aim is to further use this corpus to construct an improved NER system for the Romanian legal domain. Currently available models 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 [28] previously used to automatically annotate the Romanian Legal Corpus [41]. We re-evaluated the old NER system [28] on the LegalNERo corpus and it achieved only a 49.38% average F1 score (with individual F1 scores 84.06% for time expressions, 56.7% for organizations, 26.85% for locations, and 19.3% for persons). This difference comes from the lack of legal-domain text used in training the old system. Even though the size of the corpus is small, it has proven useful in improving the performance of the NER system and can be considered an important first resource for NER for the Romanian language in the legal domain. Considering additional techniques, such as word embedding combinations [30] could prove beneficial in improving the overall performance.

[12] https://lod-cloud.net/dataset/racai-legalnero

Acknowledgements

Part of this work 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.

References

[1] Verginica Barbu Mititelu and Maria Mitrofan. The Romanian medical treebank-SiMoNERo. In *Proceedings of the The 15th Edition of the International Conference on Linguistic Resources and Tools for Natural Language Processing – ConsILR-2020*, pages 7–16, online, 2020.

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

[3] Cristian Cardellino, Milagro Teruel, Laura Alonso Alemany, and Serena Villata. A low-cost, high-coverage legal named entity recognizer, classifier and linker. In *Proceedings of the 16th edition of the International Conference on Articial Intelligence and Law*, pages 9–18, 2017. DOI: 10.1145/3086512.3086514.

[4] Christian Chiarcos. POWLA: Modeling Linguistic Corpora in OWL/DL. In *Proceedings of the 9th International Conference on The Semantic Web: Research and Applications*, ESWC'12, pages 225–239, Berlin, Heidelberg, 2012. Springer-Verlag. ISBN 9783642302831.

[5] Christian Chiarcos and Christian Fäth. CoNLL-RDF: Linked corpora done in an NLP-friendly way. In Jorge Gracia, Francis Bond, John P. McCrae, Paul Buitelaar, Christian Chiarcos, and

Sebastian Hellmann, editors, *Language, Data, and Knowledge*, pages 74–88, Cham, 2017. Springer International Publishing. ISBN 978-3-319-59888-8.

[6] Christian Chiarcos and Luis Glaser. A tree extension for CoNLL-RDF. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7161–7169, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.885.

[7] Gordon V. Cormack, Maura R. Grossman, Bruce Hedin, and Douglas W. Oard. Overview of the TREC 2010 legal track. In Ellen M. Voorhees and Lori P. Buckland, editors, *Proceedings of The Nineteenth Text REtrieval Conference, TREC 2010, Gaithersburg, Maryland, USA, November 16-19, 2010*, volume 500-294 of *NIST Special Publication*. National Institute of Standards and Technology (NIST), 2010. URL https://trec.nist.gov/pubs/trec19/papers/LEGAL10.OVERVIEW.pdf.

[8] Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 97–102, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL https://aclanthology.org/D17-2017. DOI: 10.18653/v1/D17-2017.

[9] Christopher Dozier, Ravikumar Kondadadi, Marc Light, Arun Vachher, Sriharsha Veeramachaneni, and Ramdev Wudali. *Named Entity Recognition and Resolution in Legal Text*, pages 27–43. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010. ISBN 978-3-642-12837-0.

[10] Ștefan Daniel Dumitrescu and Andrei-Marius Avram. Introducing RONEC - the Romanian named entity corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4436–4443, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.546.

[11] Corina Forăscu and Dan Tufiş. Romanian TimeBank: An annotated parallel corpus for temporal information. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3762–3766, 2012.

[12] Ingo Glaser, Bernhard Waltl, and Florian Matthes. Named entity recognition, extraction, and linking in German legal contracts. In *IRIS: Internationales Rechtsinformatik Symposium*, pages 325–334, 2018.

[13] Aitor Gonzalez-Agirre, Montserrat Marimon, Ander Intxaurrondo, Obdulia Rabal, Marta Villegas, and Martin Krallinger. PharmaCoNER: Pharmacological substances, compounds and proteins named entity recognition track. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 1–10, Hong Kong, China, November 2019. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/D19-5701. DOI: 10.18653/v1/D19-5701.

[14] Ralph Grishman and Beth Sundheim. Message Understanding Conference- 6: A brief history. In *COLING 1996 Volume 1: The 16th International Conference on Computational Linguistics*, 1996. URL https://www.aclweb.org/anthology/C96-1079.

[15] Sebastian Hellmann, Jens Lehmann, Sören Auer, and Martin Brümmer. Integrating nlp using linked data. In Harith Alani, Lalana Kagal, Achille Fokoue, Paul Groth, Chris Biemann, Josiane Xavier Parreira, Lora Aroyo, Natasha Noy, Chris Welty, and Krzysztof Janowicz, editors, *The Semantic Web – ISWC 2013*, pages 98–113, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg. ISBN 978-3-642-41338-4.

[16] Rinke Hoekstra, Joost Breuker, Marcello Di Bello, and Alexander Boer. Lkif core: Principled ontology development for the legal domain. In *Proceedings of the 2009 Conference on Law, Ontologies and the Semantic Web: Channelling the Legal Information Flood*, pages 21–52, NLD, 2009. IOS Press. ISBN 9781586039424.

[17] Yuting Hu and Suzan Verberne. Named entity recognition for Chinese biomedical patents. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 627–637, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. . URL https://www.aclweb.org/anthology/2020.coling-main.54.

[18] Radu Ion, Vasile Păiș, and Maria Mitrofan. RACAI's system at PharmaCoNER 2019. In *Proceedings of The 5th Workshop on BioNLP Open Shared Tasks*, pages 90–99, Hong Kong, China, November 2019. Association for Computational Linguistics. . URL https://www.aclweb.org/anthology/D19-5714.

[19] Yoshinobu Kano, Mi-Young Kim, Masaharu Yoshioka, Yao Lu, Juliano Rabelo, Naoki Kiyota, Randy Goebel, and Ken Satoh. Coliee-2018: Evaluation of the competition on legal information extraction and entailment. In Kazuhiro Kojima, Maki Sakamoto, Koji Mineshima, and Ken Satoh, editors, *New Frontiers in Artificial Intelligence*, pages 177–192, Cham, 2019. Springer International Publishing. ISBN 978-3-030-31605-1.

[20] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174, 1977.

[21] Jörg Landthaler, Bernhard Waltl, and Florian Matthes. Unveiling references in legal texts-implicit versus explicit network structures. In *IRIS: Internationales Rechtsinformatik Symposium*, volume 8, pages 71–8, 2016.

[22] Elena Leitner, Georg Rehm, and Julian Moreno-Schneider. Fine-grained named entity recognition in legal documents. In Maribel Acosta, Philippe Cudré-Mauroux, Maria Maleshkova, Tassilo Pellegrini, Harald Sack, and York Sure-Vetter, editors, *Semantic Systems. The Power of AI and Knowledge Graphs*, pages 272–287, Cham, 2019. Springer International Publishing. ISBN 978-3-030-33220-4.

[23] Christopher D. Manning, Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard, and David McClosky. The Stanford CoreNLP natural language processing toolkit. In *Association for Computational Linguistics (ACL) System Demonstrations*, pages 55–60, 2014. URL http://www.aclweb.org/anthology/P/P14/P14-5010.

[24] Dragos Stefan Munteanu and Daniel Marcu. Improving machine translation performance by exploiting non-parallel corpora. *Computational Linguistics*, 31(4):477–504, 12 2005. ISSN 0891-2017. . URL https://doi.org/10.1162/089120105775299168.

[25] Aparna Nayak, Bojan Božić, and Luca Longo. Linked data quality assessment: A survey. In Chengzhong Xu, Yunni Xia, Yuchao Zhang, and Liang-Jie Zhang, editors, *Web Services – ICWS 2021*, pages 63–76, Cham, 2022. Springer International Publishing. ISBN 978-3-030-96140-4.

[26] Douglas W. Oard, Jason R. Baron, Bruce Hedin, David D. Lewis, and Stephen Tomlinson. Evaluation of information retrieval for e-discovery. *Artif. Intell. Law*, 18(4):347–386, De-

cember 2010. ISSN 0924-8463. . URL https://doi.org/10.1007/s10506-010-9093-9.

[27] Vasile Păiş, Maria Mitrofan, Carol Luca Gasan, Vlad Coneschi, and Alexandru Ianov. 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, November 2021. Association for Computational Linguistics. URL https://aclanthology.org/2021.nllp-1.2. DOI: 10.18653/v1/2021.nllp-1.2.

[28] Vasile Păiș. *Contributions to semantic processing of texts; Identification of entities and relations between textual units; Case study on Romanian language*. PhD thesis, Romanian Academy, 2019.

[29] Vasile Păiș. Multiple annotation pipelines inside the RELATE platform. In *The 15th International Conference on Linguistic Resources and Tools for Natural Language Processing*, pages 65–75, online, 2020. URL https://profs.info.uaic.ro/~consilr/wp-content/uploads/2021/03/volum-ConsILR-v-4-final-revizuit.pdf#page=73.

[30] Vasile Păiș and Maria Mitrofan. Assessing multiple word embeddings for named entity recognition of professions and occupations in health-related social media. In *Proceedings of the Sixth Social Media Mining for Health (#SMM4H) Workshop and Shared Task*, pages 128–130, Mexico City, Mexico, June 2021. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2021.smm4h-1.27.

[31] Vasile Păiș and Dan Tufiș. 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, 2018.

[32] Vasile Păiș, Radu Ion, and Dan Tufiș. A processing platform relating data and tools for Romanian language. In *Proceedings of the 1st International Workshop on Language Technology Platforms*, pages 81–88, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-64-1. URL https://www.aclweb.org/anthology/2020.iwltp-1.13.

[33] Vasile Păiș, Maria Mitrofan, Carol Luca Gasan, Alexandru Ianov, Corvin Ghiță, Vlad Silviu Coneschi, and Andrei Onuț. Romanian named entity recognition in the legal domain (LegalNERo), May 2021. URL https://doi.org/10.5281/zenodo.4772094.

[34] Giuseppe Rizzo and Raphaël Troncy. NERD: A framework for unifying named entity recognition and disambiguation extraction tools. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 73–76, Avignon, France, April 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E12-2015.

[35] Erik F Sang and Jorn Veenstra. 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, 1999.

[36] Burr Settles. Biomedical named entity recognition using conditional random fields and rich feature sets. In *Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLP-BA/BioNLP)*, pages 107–110, Geneva, Switzerland, August 28th and 29th 2004. COLING. URL https://www.aclweb.org/anthology/W04-1221.

[37] Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. brat: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France, April 2012. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/E12-2021.

[38] Milan Straka and Jana Straková. Tokenizing, pos tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada, August 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/K/K17/K17-3009.pdf. DOI: 10.18653/v1/K17-3009.

[39] Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. Yago: A core of semantic knowledge. In *Proceedings of the 16th International Conference on World Wide Web*, WWW '07, pages 697–706, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-654-7. URL http://doi.acm.org/10.1145/1242572.1242667. DOI: 10.1145/1242572.1242667.

[40] Erik F. Tjong Kim Sang and Fien De Meulder. Introduction to the CoNLL-2003 shared task: Language-independent named entity recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147, 2003. URL https://www.aclweb.org/anthology/W03-0419.

[41] Dan Tufiș, Maria Mitrofan, Vasile Păiș, Radu Ion, and Andrei Coman. Collection and annotation of the Romanian legal corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2773–2777, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.337.

[42] Dan Tufiș, Verginica Barbu Mititelu, Elena Irimia, Vasile Păiș, 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.

[43] 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. The MARCELL legislative corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 3761–3768, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://www.aclweb.org/anthology/2020.lrec-1.464.

[44] Tamás Váradi, Bence Nyéki, Svetla Koeva, Marko Tadić, Vanja Štefanec, Maciej Ogrodniczuk, Bartlomiej Nitoń, Piotr Pęzik, Verginica Barbu Mititelu, Elena Irimia, Maria Mitrofan, Vasile Păiș, Dan Tufiș, Radovan Garabík, Simon Krek, and Andraž Repar. Introducing the CURLICAT corpora: Seven-language domain specific annotated corpora from curated sources. In *Proceedings of the Language Resources and Evaluation Conference*, pages 100–108, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.11.

[45] Vikas Yadav and Steven Bethard. A survey on recent advances in named entity recognition from deep learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1182.

[46] Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Sören Auer. Quality assessment for linked data: a survey. *Semantic Web*, 7:63–93, 2016. ISSN 2210-4968. URL https://doi.org/10.3233/SW-150175. DOI: 10.3233/SW-150175.

Appendix A. Example encoding in the LegalNERo corpus

```
:c1 a powla:Corpus,dcat:Dataset,prov:Entity;
    dct:title "LegalNERo" ;
    dcat:theme :ner ;
    dcat:distribution :legalnero.zip ;
    dct:publisher :racai ;
    powla:documentID "LegalNERo" .
:d1 a powla:Document ;
    powla:documentID "
    mj_00000G0NTTBOIDQWK933SGOQ93YB6CS6" ;
    powla:hasSuperDocument :c1 .
:d1_l_sent a powla:DocumentLayer ;
    powla:hasDocument :d1 .
:d1_l_tok a powla:DocumentLayer ;
    powla:hasDocument :d1 .
:d1_l_ner a powla:DocumentLayer ;
    powla:hasDocument :d1 .
:d1_s1 a nif:Sentence, powla:Node ;
    nif:firstWord :d1_s1_1 ;
    nif:lastWord :d1_s1_66 ;
    nif:word :d1_s1_1 ;
    nif:word :d1_s1_2 ;
    powla:hasLayer :d1_l_sent .
:d1_s1_1 a nif:Word, powla:Node ;
    nif:nextWord :d1_s1_2 ;
    powla:string "LEGE" ;
    nif:anchorOf "LEGE" ;
    conllu:ID "1" ;
    conllu:FORM "LEGE" ;
    conllu:LEMMA "lege" ;
    conllu:UPOS "VERB" ;
    conllu:XPOS "Vmip3s" ;
    conllu:FEATS "Mood=Ind|Number=Sing|Person
    =3|Tense=Pres|VerbForm=Fin" ;
    conllu:HEAD "0" ;
    conllu:DEPREL "root" ;
    conllu:MISC "SpacesBefore=\\r\\n" ;
    conllup:NE "B-LEGAL" ;
    nif:sentence :d1_s1 ;
    powla:hasParent :d1_s1 ;
    powla:next :d1_s1_2 ;
    powla:hasLayer :d1_l_tok .
:d1_s1 nif:nextSentence :d1_s2 .
:d1_s2 nif:previousSentence :d1_s1 .
```

```
:d1_s1 powla:next :d1_s2 .
:d1_s2 powla:previous :d1_s1 .
:d1_e1 a nerd:Thing, eli:LegalResource, powla
    :Node, nif:Phrase ;
    powla:hasLayer :d1_l_ner ;
    powla:string "LEGE nr. 185 din 17
    octombrie 2019" ;
    nif:anchorOf "LEGE nr. 185 din 17
    octombrie 2019" ;
    nif:beginIndex "2" ;
    nif:endIndex "36" .
:d1_e2 a nerd:Time, powla:Node, nif:Phrase ;
    powla:hasLayer :d1_l_ner ;
    powla:string "17 octombrie 2019" ;
    nif:anchorOf "17 octombrie 2019" ;
    nif:beginIndex "19" ;
    nif:endIndex "36" .
```

The corpus object is encoded as "c1" (classes "powla:Corpus", "dcat:Dataset"). Documents ("powla:Document") contain "powla:DocumentLayer" for sentences ("d1_l_sent"), tokens ("d1_l_tok"), and entities ("d1_l_ner"), and are linked to the corpus ("powla:hasSuperDocument") (see "d1"). Sentences ("nif:Sentence") are linked to the appropriate document layer. Individual tokens ("nif:Word") contain conllu annotations and are linked to the sentence ("nif:sentence") and to the tokens layer, using the relation "powla:hasLayer" (see "d1_s1_1"). Span-based entities (see "d1_e1") belong to NERD classes or "eli:LegalResource". These are linked to the NER layer and contain the start ("nif:beginIndex") and end ("nif:endIndex") indexes. Object names are encoded using an initial letter ("c" - corpus, "d" - document, "s" - sentence, "e" - entity) followed by a number ("d1_s1" means the first sentence of the first document).