Proceedings of the Linked Data in Linguistics Workshop @ LREC2022, pages 35–44
Marseille, 24 June 2022

Use Case: Romanian Language Resources in the LOD Paradigm

Verginica Barbu Mititelu, Elena Irimia, Vasile Păiș, Andrei-Marius Avram, Maria Mitrofan
Romanian Academy Research Institute for Artificial Intelligence
13 Calea 13 Septembrie, Bucharest, Romania
{vergi,elena,vasile,andrei.avram,maria}@racai.ro

Abstract

In this paper, we report on (i) the conversion of Romanian language resources to the Linked Open Data specifications and requirements, on (ii) their publication and (iii) interlinking with other language resources (for Romanian or for other languages). The pool of converted resources is made up of the Romanian Wordnet, the morphosyntactic and phonemic lexicon RoLEX, four treebanks, one for the general language (the Romanian Reference Treebank) and others for specialised domains (SiMoNERo for medicine, LegalNERo for the legal domain, PARSEME-Ro for verbal multiword expressions), frequency information on lemmas and tokens and word embeddings as extracted from the reference corpus for contemporary Romanian (CoRoLa) and a bi-modal (text and speech) corpus. We also present the limitations coming from the representation of the resources in Linked Data format. The metadata of LOD resources have been published in the LOD Cloud. The resources are available for download on our website and a SPARQL endpoint is also available for querying them.

Keywords: Romanian, OntoLex, LLOD Cloud

1. Introduction

According to the recently collected data[1] on the existence and availability of resources and technologies for different languages, Romanian is a language with fragmentary technological support[2]. This means a presence of between 3% and 10% in the catalogue of language resources available in the European Language Grid[3]. The vast majority of these resources are available in various formats adopted according to the requirements or needs of the projects in which they were created: e.g., the Romanian Wordnet was created in XML format (Tufiș et al., 2004a), the corpus annotated with verbal multiword expressions was released in CUPT format (Ramisch et al., 2018), etc.

In the last few years, within the Natural Language Processing group of the Romanian Academy Research Institute for Artificial Intelligence[4], steps have been taken to convert the resources developed herein throughout time (Tufiș, 2022) to the specifications of the Linked Open Data (LOD) paradigm, so as to ensure them the benefits derived from this: higher visibility, accessibility, contextualization (by linking them to other resources) and, eventually, further increase of the technological development of Romanian. The most important decision was to make them open. They have been made freely available and enriched with metadata, which have been added to the Linked Open Data Cloud[5] (LOD Cloud).

The representation of Romanian in the Linguistics LOD Cloud (LLOD Cloud) is not only due to our contribution. Four resources (all created in a multilingual context) had metadata already recorded in the LLOD Cloud when we started our endeavour. They were: EuroVoc[6], Universal Dependencies[7] Treebank Romanian, Multext-East[8] and Romanian WordNet (as part of Open Multilingual WordNet[9]). We present the resources we converted (providing a brief description of their content and of their representation in the LOD paradigm) in Section 2. Their publication methods are enumerated in Section 3. A presentation of the way in which they are interlinked among themselves or with other resources is available in Section 4. Some potential use cases are designed in Section 5, before concluding the paper.

2. Romanian Language Resources converted to LOD

During the last year we have added metadata of 8 more resources[10] for Romanian in the LLOD Cloud: the whole Romanian Wordnet (bigger than the one available in Open Multilingual WordNet), the morpho-phonemic lexicon RoLEX, four treebanks (the Romanian Reference Treebank RRT, the medical treebank SiMoNERo, the law treebank LegalNERo, the treebank annotated with verbal multiword expressions PARSEME-Ro), lemmas and tokens frequencies and word embeddings extracted from the corpus of contemporary Romanian (CoRoLa), and a bimodal (written and oral) corpus (RTASC). These have been chosen to be converted to the LOD specification because they are the main resources our group have created throughout time, they are still relevant in the international linguistic context and are, thus, worth being made more visible and accessible. We describe each resource below and present the decisions made about their conversion to the LOD specifications, as well as the limitations of this representation.

While there are more formats to represent Linked Data (N-Triples, RDF Turtle, JSON-LD and RDF/XML among the most common), we chose RDF Turtle[11] for its advantage of human readability, due to the possibility of defining prefixes in the beginning of the file. For syntactic

[1] Within the project European Language Equality (ELE) (https://european-language-equality.eu/) the LT support of the 24 official and 32 additional EU-languages as well as 33 endangered minority languages has been evaluated and a report on the state of the art in language technology and language-centric AI has been released for each language.

[2] https://european-language-equality.eu/wp-content/uploads/2022/03/ELE___Deliverable_D1_29__Language_Report_Romanian_.pdf

[3] https://live.european-language-grid.eu/

[4] www.racai.ro

[5] https://lod-cloud.net/

[6] https://op.europa.eu/en/web/eu-vocabularies/dataset/-/resource?uri=http://publications.europa.eu/resource/dataset/eurovoc

[7] https://universaldependencies.org/

[8] http://nl.ijs.si/ME/Vault/V4/

[9] http://compling.hss.ntu.edu.sg/omw/

[10] https://lod-cloud.net/datasets?search=racai

[11] https://www.w3.org/TR/turtle/

validation of the .ttl files, we used an open source tool[12]. Their semantic validation was done for the original format: e.g., for RoWN an in-house tool was developed (Tufiș et al., 2004a) for detecting semantic incorrectness of the resource. A semantic evaluation of the LD format of the resources is yet to be made, according to acknowledged criteria (Zaveri et al, 2015).

The URIs were created individually for each resource, using a resource-specific format. Of course, when different objects were reused between resources, potentially allowing for interlinking, the same URI was employed. The URIs do not correspond to real-world URLs, and thus individual objects cannot be accessed via the Internet. The resources are intended to be used either as a complete download of the entire file or through the provided SPARQL endpoints.

2.1 The Romanian Wordnet

2.1.1 Description of the Romanian Wordnet

The lexical ontology for Romanian, i.e. the Romanian Wordnet (RoWN, Tufiș and Barbu Mititelu, 2014), was developed by translating the Princeton WordNet (PWN, Miller, 1995; Fellbaum, 1998) synsets and transferring the relations between equivalent synsets. It contains 56,591 synsets in which 53,092 (noun, verb, adjective or adverb) literals occur. In PWN *semantic relations* are established between synsets, which are lexicalizations of concepts; thus this type of relations has cross-lingual validity (to a certain extent), which offers the grounds for transferring them between equivalent synsets in networks for other languages. RoWN was created and maintained in an XML format, with a DTD specific to the BalkaNet project principles[13].

2.1.2 Conversion of RoWN to LOD specifications

We used the following *OntoLex-lemon* classes and properties to represent RoWN, as *lemon*[14], the lexicon model for ontologies developed by the Ontology-Lexica (OntoLex) community group, is the recommended standard for wordnets[15]:

1. *ontolex: LexicalEntry* (mono- or multi-word), described by a lemma (*ontolex:CanonicalForm*), a part of speech (*wn:partOfSpeech*) and a reference to a *LexicalSense* object.

2. *ontolex: LexicalSense* (represents one of the meanings of the lexical entry and contains a reference to a synset in the network, encoded with the *ontolex: reference* property);

3. *ontolex:LexicalConcept*: encodes the synset referenced by the *LexicalSense*, described by a definition, an *ILI* (an id in the the collaborative interlingual index of concept for wordnets[16] (Bond et al., 2016)) and a part-of-speech (all defined by the wordnet specialised vocabulary wn).

The following listing shows an entry in the original XML format, corresponding to the lemma “pom” (En. “tree”). For simplicity, only one semantic relation (out of 26 in which this synset is involved) is presented here.

```
<SYNSET>
   <ID>rown-12651821-n</ID>
    <POS>n</POS>
     <SYNONYM>
           <LITERAL>pom<SENSE>1</SENSE></LITERAL>
   </SYNONYM>
   <DEF>nume generic pentru orice arbore sălbatic sau cultivat,
care produce fructe</DEF>
 <ILR>ENG30-13109733-n<TYPE>hypernym</TYPE></ILR>
   ...
</SYNSET>
```

Listing 1. A synset from RoWN in its original XML form

Listing 2 illustrates the LLOD encoding model for the same lemma “pom”. The LITERAL XML-attribute becomes the main class LexicalEntry in the *lemon* model, the SENSE attribute of the LITERAL is encoded as an ontolex:LexicalSense, while the synset, which was the basic entry in the XML file, is now just a reference property of the LexicalSense. The reference r*own:12651821-n* is described further in the file by the wn:partOfSpeech *n,* the wn:ili *103362* and a definition in Romanian. The LexicalEntry ID is generated by concatenating the literal, the pos and the synset number to differentiate it from the same literal in other possible synsets. The *lemon* variation and translation module *vartrans* is then used to express synset relations, by encoding the source, the target and the relation category (hypernym in our example).

```
rown:pom-n-12651821 a ontolex:LexicalEntry ;
          ontolex:canonicalForm [
                    ontolex:writtenRep "pom"@ro ] ;
          wn:partOfSpeech wn:n ;
          ontolex:Sense rown:pom-n-12651821-1 .
rown:pom-n-12651821-1 a ontolex:LexicalSense ;
          ontolex:reference rown:12651821-n .
...
rown:12651821-n a ontolex:LexicalConcept ;
          wn:partOfSpeech wn:n ;
          wn:ili ili:i103362 ;
          wn:definition [
                    rdf:value "nume generic pentru orice arbore
sălbatic sau cultivat, care produce fructe"@ro] .
…

vartrans:source    <https://www.racai.ro/p/llod/resources/rown-
3.0.ttl/12651821-n> ;
vartrans:category wn:hypernym ;
vartrans:target    <https://www.racai.ro/p/llod/resources/rown-
3.0.ttl/13109733-n> .
```

Listing 2. The same synset from RoWN in Turtle form

Some information from RoWN cannot be represented in LD format yet: (i) Balkan-specific concepts and (ii) a different treatment of PWN lexical relations. (i) During the BalkaNet project[17] (Tufiș et al., 2004b), some synsets lexicalizing Balkan-specific concepts have been implemented. They were added as hyponyms of synsets already translated from PWN. These new synsets have a specific identifier (BILI). Lacking a correspondent in the interlingual index (so they cannot be assigned an *ili*), these BILI synsets are not included in the LD format of RoWN. (ii) As far as the *lexical relations*[18] (antonymy and derivational relations) are concerned, their semantic

[12] https://github.com/IDLabResearch/TurtleValidator

[13] http://www.dblab.upatras.gr/balkanet/deliverables/d.2.pdf

[14] https://lemon-model.net

[15] http://bpmlod.github.io/report/WordNets/index.html

[16] https://github.com/globalwordnet/cili

[17] http://www.dblab.upatras.gr/balkanet/

[18] PWN distinguishes between semantic relations, holding between senses (so between synsets), and lexical relations, holding between word forms.

component has also been acknowledged in the development of RoWN: for example, the antonymy relation between two word forms can be safely exported to the synsets to which these two words belong as a conceptual opposition. This makes such relations transferable between equivalent synsets in other wordnets.

These two characteristics are relevant with respect to the LD representation of RoWN (as compared to that of other wordnets). Future work will seek to offer solutions for these cases.

2.2 RoLEX

2.2.1 Description of RoLEX

RoLEX is the most extensively validated phonemic lexicon available for the Romanian language. It contains 330,886 entries and it was initially developed in tabular format[19], with 6 columns: the lexical form, its lemma, a morphosyntactic description in the MSD[20] format, the syllabification with syllable boundaries marked by (.), the stress marked by placing (‘) in front of the stressed vowel, and the phonemic transcription using an extended version of Speech Assessment Methods Phonetic Alphabet (SAMPA)[21] for Romanian (see Listing 3).

It is an organically developed resource, based on the textual component of speech corpora collected in the ReTeRom project[22]. It is built on a list of lexical items extracted from a corpus of Romanian Wikipedia texts that was read by volunteers, news, interviews, talk shows, spontaneous speech, read fairy tales and novels. The morphosyntactic and lemma information associated with each item in the list was taken from a Romanian lexicon[23] that our team maintains, while syllabification, stress and phonemic transcription information was partially extracted from existing resources – RoSyllabiDict (Barbu, 2008) and MaRePhor (Toma et al., 2017) – and partially automatically generated using Stan et al.’s (2011) tool. On top of aggregating all this data, we applied a thorough curation, using techniques of automatic validation and correction, but also manual correction of automatically identified errors. The information about lemma and morphosyntactic descriptions was entirely manually validated and corrected in the original extensive Romanian lexicon TBL, while information about syllabification, stress and phonemic transcription has gone through a process of (1) automatic selection of entries with a high syllabification error probability (entries with abnormally long syllables, with syllables containing two vocals, with syllables containing one vowel followed by one or two semi-vowels, with syllables containing some specific letter groups “ce/ci/ge/gi/che/chi/ghe/ghi”, entries that contain proper nouns, etc.) followed by manual or automatic correction; (2) automatic selection of entries containing homographs that are not homophones: in these cases, stress can be correctly marked by taking into account information like lemma or the morphosyntactic description and its correct marking can consequently impact syllabification and phonemic transcription; (3) implementation of phonemic transcription rules based on correct syllabification and stress marking. All the corrections were made by two linguists; the work was distributed and inter-annotator agreement was not pursued, since correction tasks were rather trivial.

2.2.2 Conversion of RoLEX to LOD specifications

OntoLex-lemon was the main frame of representation for the RoLEX conversion to LLOD specification. *Ontolex:LexicalEntry* was used to encode the unique lemmas in RoLEX. The inflectional paradigm of the lemma is encoded as a list of *ontolex:lexicalForm* and the set of senses traditionally associated with a LexicalEntry is absent, but compensated by interlinking RoLEX lexical entries with RoWN synsets via *ili*. We exemplify with the six inflected forms of the lemma “pom” (see Listing 3). They were converted in the *LexicalEntry* :lex_pom that has, among other properties, 6 ontolex:LexicalForm associated with it (see Listing 4).

pom	*=*	*Ncms-n*	*pom*	*pom*	*p o m*
pomul	*pom*	*Ncmsry*	*po.mul*	*p'omul*	*p o m u l*
pomului	*pom*	*Ncmsoy*	*po.mu.lui*	*p'omului*	*p o m u l u j*
pomi	*pom*	*Ncmp-n*	*pomi*	*pomi*	*p o m i_0*
pomii	*pom*	*Ncmpry*	*po.mii*	*p'omii*	*p o m i j*
pomilor	*pom*	*Ncmpoy*	*po.mi.lor*	*p'omilor*	*p o m i l o r*

Listing 3. Example of 6 entries associated to lemma “pom” from RoLEX in tabular format

```
:lex_pom_n a ontolex:LexicalEntry;
            rdfs:label "pom_n"@ro;
            ontolex:canonicalForm :form_pom_n;
            wn:partOfSpeech wn:n;
            wn:ili ili:i103362;
            wn:ili ili:i105570;
ontolex:lexicalForm :form_pom_n_noun_ind_masc_sing;
ontolex:lexicalForm :form_pom_n_noun_ind_masc_plur;
ontolex:lexicalForm:form_pom_n_noun_acc_nom_def_masc_pl
ur;
ontolex:lexicalForm:form_pom_n_noun_dat_gen_def_masc_plu
r;
ontolex:lexicalForm:form_pom_n_noun_acc_nom_def_masc_si
ng;
ontolex:lexicalForm:form_pom_n_noun_dat_gen_def_masc_sin
g .
```

Listing 4. Example of an ontolex:LexicalEntry from RoLEX

We used *ontolex:canonicalForm* property to specify the canonical form (i.e. lemma) associated with the specific lexical entry: :form_pom_n. The canonical form is later described through the property *ontolex:writtenRep* as "pom"@ro. Although encoding the same information as ontolex:canonicalForm, *rdfs:label* is also used, as OntoLex recommends it for compatibility with RDFS-based representation systems. The *partOfSpeech* property from the *wn*[24] vocabulary encodes the part-of-speech information “n” (noun).

To generate lexicalForm labels, we used the MSD tag uniquely associated with each form and expanded it in a Universal Dependencies (UD) feature list[25]. For example, the label form_pom_n_noun_acc_nom_def_masc_plur

[19] RoLEX in tabular format is freely available for download at https://www.racai.ro/p/reterom/results.html.

[20] https://github.com/clarinsi/mte-msd/blob/master/tables/msd-canon-ro.tbl

[21] https://www.phon.ucl.ac.uk/home/sampa/romanian.htm

[22] https://www.racai.ro/p/reterom/

[23] https://raw.githubusercontent.com/racai-ai/Rodna/master/data/resources/tbl.wordform.ro,

[24] http://globalwordnet.github.io/schemas/wn

[25] https://universaldependencies.org/u/feat/

was created starting from the "Ncmpry" MSD, encoded by the *conll:POS* property as can be seen in Listing 5:

```
:form_pom_n_noun_acc_nom_def_masc_plur a ontolex:Form;
          ontolex:writtenRep "pomii"@ro;
          conll:POS "Ncmpry";
          ontolex:writtenRep "po.mii"@syl;
          ontolex:writtenRep "p'omii"@stress;
          ontolex:phoneticRep "p o m i j"@ro-RO-sampa .
```

Listing 5. Description of an ontolex:Form in RoLEX

Finally, the OntoLex properties *writtenRep* and *phoneticRep* are used to represent the lexicalisation and the phonemic transcription associated with the form. In the absence of specific OntoLex properties, we also used writtenRep to encode syllabification and stress.

2.3 Treebanks

2.3.1 Description of the Romanian Treebanks

RoRefTrees (RRT) (Barbu Mititelu, 2018) is the reference treebank for standard Romanian. It has 9,523 sentences containing 218k tokens. It covers a variety of genres (legal, news, fiction, medical, science, academic writing, Wikipedia) and reflects the contemporary language. The RRT treebank is distributed in the UD releases and is freely available[26].

SiMoNERo (Barbu Mititelu and Mitrofan, 2020) is the medical treebank for the Romanian language and has three levels of annotation: gold standard morphological annotation, hand validated annotation with medical named entities and syntactic annotation in compliance with UD specifications. SiMoNERo contains texts from the BioRo corpus (Mitrofan and Tufiș, 2018) belonging mainly to three main medical domains: diabetes, cardiology and endocrinology. All the texts are extracted from three types of documents: medical scientific journal articles, scientific medical books and medical blog posts. Currently, the treebank contains 4,681 sentences distributed in 146k tokens. SiMoNERo also has 14,133 medical named entities distributed in the four types: anatomical parts (ANAT), chemicals (CHEM), disorders (DISO) and procedures (PROC). The TTL tool (Ion, 2007) was used for tokenization, lemmatization, part-of-speech tagging and dependency parsing, while the dependency parsing level was added using NLP-Cube (Boroș et. al, 2018). The treebank is also distributed in the UD releases and is freely available for download[27].

LegalNERo (Păiș et. al, 2021a; Păiș and Mitrofan, 2021) is the first legal treebank for the Romanian language and it also has three annotation levels: morphological, syntactic and named entities annotation. The LegalNERo corpus contains a total of 370 documents selected from MARCELL-RO corpus (Tufiș et al., 2020) and 265k tokens. It provides gold annotations for five entity classes: organisations (ORG), locations (LOC), persons (PER), time expressions (TIME) and legal resources mentioned in legal documents (LEGAL). The UDPipe tool (Straka et al., 2016) was used for tokenization, lemmatization, part-of-speech tagging and dependency parsing. The treebank is available[28] in multiple formats, including span-based, token-based and RDF.

PARSEME-Ro (Barbu Mititelu et al., 2019a) is a journalistic corpus automatically morpho-syntactically annotated using UDPipe. It was further manually enriched with semantic information of the type verbal multiword expressions (VMWEs), within the PARSEME project (Savary et al., 2018). Three main types of VMWEs were annotated: light verb constructions, verbal idioms and reflexive verbs. The corpus is also freely available[29] together with the corpora annotated for other languages (Ramisch et al., 2020).

2.3.2 Conversion of the Romanian treebanks to LOD specifications

The Romanian treebanks were converted to the LOD specifications by using the CoNLL-U (the original format of the corpus) to an RDF graph (the target format for LOD) tool that was developed by the Applied Computational Linguistics (ACoLi) laboratory (Chiarcos et al., 2017). This tool converted the CoNLL-U files into the Turtle format.

We used the NIF[30] format to achieve interoperability between the sentences and the words found in the Romanian treebanks. Thus, we employ objects of type *nif:Sentence* to denote the id of the sentence and its text and objects of type *nif:Word* together with *conll* properties to describe the words:

- *conll:ID* - the word index
- *conll:WORD* - the original word, as it is found in the sentence
- *conll:LEMMA* - the lemma of the word form
- *conll:UPOS* - the universal part-of-speech
- *conll:POS* - the extended part-of-speech
- *conll:FEAT* - list of morphological features
- *conll:HEAD* - the head of the word in the dependency tree of the sentence
- *conll:EDGE* - the syntactic relation established with the head
- *conll:MISC* - the semantic information added to some of the treebanks, namely the medical entities in SiMoNERo, the legal ones on LegalNERo, the verbal multiword expressions in PARSEME-Ro.

We also specify the next sentence using the *nif:nextSentence* object and the next word with the *nif:nextWord* object. An example of a sentence and its first word from RRT in the LOD format is given in Listing 6.

```
# sent_id = dev-6
# text = Prin însăși natura lucrurilor era imposibil.
:rrt_dev_s5_0 nif:nextSentence :rrt_dev_s6_0 .
:rrt_dev_s6_0 a nif:Sentence;
   rdfs:comment
      "sent_id = dev-6
      text = Prin însăși natura lucrurilor era imposibil." .
:rrt_dev_s6_1 a nif:Word;
    conll:WORD "Prin";
    conll:EDGE "case";
    conll:FEAT "AdpType=Prep|Case=Acc";
    conll:HEAD :rrt_dev_s6_3;
```

[26] https://github.com/UniversalDependencies/UD_Romanian-RRT

[27] https://github.com/UniversalDependencies/UD_Romanian-SiMoNERo

[28] https://zenodo.org/record/4772095#.YkHJT3pByM8

[29] https://lindat.mff.cuni.cz/repository/xmlui/handle/11234/1-3367

[30] https://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core/nif-core.html

```
conll:ID "1";
conll:LEMMA "prin";
conll:POS "Spsa";
conll:UPOS "ADP";
nif:nextWord :rrt_dev_s6_2 .
```

Listing 6. Example from the RRT treebank

2.4 Corpus-based Frequencies and Embeddings

2.4.1 Description of CoRoLa

The representative corpus of contemporary Romanian (CoRoLa) (Tufiş et al., 2019) was created as a priority project of the Romanian Academy and made available to the community in 2017: its 1.2 billion words cover imaginative, journalistic, scientific, administrative, law, memoirs, and blogpost texts, assignable to 4 domains (Society, Science, Nature, Arts & Culture) and tens of subdomains thereof. Each file has associated metadata, with information about the text author, title, year of publication, text type, domain, subdomain, source, etc. The vast majority of the texts included in the corpus are cleared with respect to the Intellectual Property Right, but for some we only have the right to store and process them. Thus, the corpus is not downloadable, though it is indexed and queryable in the KorAP platform[31] (Diewald et al., 2016; Diewald et al., 2019). Besides the written component, CoRoLa also contains an oral component, made up of over 100 hours of speech, which is indexed and can be queried on a different platform[32].

2.4.2 Conversion of CoRoLa-based frequencies to LOD specifications

One of the ways in which the corpus can be exploited is by creating corpus-driven data, such as word and lemma frequencies. This information, initially available as simple lists, were converted to linked data specifications.

For conversion purposes, we used the OntoLex-FRAC[33] specification, designed by OntoLex for frequency, attestation and corpus information. An automatic conversion tool, written in Java, was developed, taking as input a CSV file with frequencies and producing a file in RDF Turtle format. For both word and lemma frequencies we used *frac:frequency* objects of type *frac:CorpusFrequency* to hold the actual values as integers. An example is given below. All the vocabularies used for the representation are available in Table 1 and an example is given in Listing 7.

```
:sat a ontolex:LexicalEntry;
      ontolex:canonicalForm "sat"@ro;
      frac:frequency [
        a frac:CorpusFrequency;
        rdf:value "33059"^^xsd:int;
        dct:source <http://corola.racai.ro/> ].
```

Listing 7. Word frequency representation

2.4.3 Conversion of CoRoLa-based word embeddings to LOD specifications

Multiple word embeddings representations were trained using the CoRoLa corpus (Păiş and Tufiş, 2018b). For conversion to OntoLex-FRAC specifications we considered only the best performing model as described by Păiş and Tufiş (2018a). The model contains vector representations of dimension 300 for 250,942 words. Data is encoded using the *frac:Embedding* class. An example is given in Listing 8.

```
:de a ontolex:LexicalEntry;
  ontolex:canonicalForm "de"@ro;
  frac:embedding [
    a :CoRoLaEmbeddings_300;
    rdf:value "0.058826 0.050749 0.094646 0.059437 0.014913 -
0.14699 0.31223 0.25699 0.020498 0.1497 -0.045657 -0.16574 -
0.14085 0.053746 -0.016113 -0.11879 0.11086 -0.086826 -
0.11564 -0.1137 -0.21041 0.12873 0.074748 -0.1439 -0.11781 -
0.14723 0.080661 0.18918 0.079647 -0.043609 -0.024831
0.058612 0.0028617 0.074098 0.048036 .......... ] .
```

Listing 8. Word embedding representation

2.5 ROBIN Technical Acquisition Speech Corpus (RTASC)

2.5.1 Description of RTASC

RTASC (Păiş et al., 2021b; 2021c) is a bimodal (speech and text) corpus, resulting from the ROBIN[34] project. The dataset was initially created to facilitate the construction of a dialogue component (Ion et al., 2020) for human-robot interaction in the context of a micro-world scenario of purchasing computers. It allowed us to improve the performance of our general speech recognition system (Avram et al., 2020) by approximately 16 WER on this domain. RTASC contains 3,786 audio files, with a total duration of 6h25m, read by multiple Romanian native speakers, associated with 711 text files. Being a read speech corpus, the audio files are aligned with the text variants. The text component was processed automatically in the RELATE platform (Păiş et al., 2020; Păiş, 2020), being tokenized, lemmatized and enhanced with morphosyntactic annotations and dependency parsing.

2.5.2 Conversion of RTASC to LOD specifications

The processed text component of the corpus became a treebank. Therefore, the conversion followed a process similar to the one described in Section 2.3.2 for the other treebanks. However, RTASC also contains the speech component which was linked to the texts. Furthermore, additional metadata, such as the recording devices used was included using the Studio Ontology Framework[35] (Fazekas and Sandler, 2011). An example speaker representation is in Listing 9.

```
:speaker_1
    a ma:Person, foaf:Person, studio:Device;
    foaf:gender "m";
    studio:microphone "Realtek HD Audio/Speedlink SL-8703-
BK" .
```

Listing 9. Speaker description in the RTASC corpus

[31] https://korap.racai.ro/

[32] http://89.38.230.23/corola_sound_search/index.php

[33] https://github.com/ontolex/frequency-attestation-corpus-information

[34] http://aimas.cs.pub.ro/robin/en/

[35] http://isophonics.org/content/studio-ontology

Audio files were added using the Ontology for Media Resources 1.0[36]. Thus a document in the corpus reflects its bimodal characteristics by making use of both *powla:Document* and *ma:DataTrack* classes. Then the actual audio file is represented using the class *ma:AudioTrack* referencing the wav file using the property *ma:hasFragment*. Additionally, the language, format and sampling rate are specified as well as the speaker who recorded the file. The link with the text representation is given as a *ma:hasSubtitling* property. An example is provided in Listing 10.

```
:d1 a powla:Document, ma:DataTrack ;
  powla:documentID "S0" ;
  powla:hasSuperDocument :c1 ;
  ma:hasLanguage [ rdfs:label "ro" ] .

:S0_4_wav a ma:MediaResource ;
  ma:hasTrack :S0_4_wav_audio ;
  ma:hasSubtitling :d1 .

:S0_4_wav_audio a ma:AudioTrack ;
  ma:hasLanguage [ rdfs:label "ro" ] ;
  ma:hasFormat "audio/wav" ;
  ma:samplingRate "44.1" ;
  ma:hasContributor :speaker_4 ;
  ma:hasFragment <S0_4.wav> .
```

Listing 10. Linking audio files in the RTASC corpus

2.6. LOD Vocabularies Usage

In our endeavour, we purposefully focused on reusing vocabularies already widely used in the LOD community and we did not create new classes or properties to represent our specificities, but adapted already existing ones. Tables 1 and 2 summarise the vocabularies used in each of the converted resources to represent the encoded data and metadata, respectively.

	RoWN	RoLEX	RRT	LegalNERo	SiMoNERo	PARSEME-Ro	CoRoLa-based freq	RTASC
ontolex-lemon[37]	x	x					x	
wn	x	x						
ili[38]	x	x						
frac[39]							x	
rdfs[40]		x	x	x	x	x		
rdf[41]							x	
vartrans[42]	x							
nif[43]			x	x	x	x		x
powla[44]								x
conll[45]		x	x	x	x	x	x	
conllu[46]							x	
ma[47]								x
xsd[48]							x	x

Table 1. Vocabularies used to encode data in Romanian resources converted to LOD.

	RoWN	RoLEX	RRT	LegalNERo	SiMoNERo	PARSEME-Ro	CoRoLa-based freq	RTASC
dct[49]							x	
foaf[50]								x
studio[51]								x
owl[52]								x
dcat[53]								x
prov[54]								x
pav[55]								x
dc[56]	x	x	x	x	x	x		
cc[57]	x	x	x	x	x	x		
schema[58]	x	x	x	x	x	x		

Table 2. Vocabularies used to encode metadata in Romanian resources converted to LOD.

3. Publication of Romanian LOD Resources

Two ways of publishing (Verborgh, 2021) resources in the LD format have been adopted: data dump and availability of SPARQL endpoint. A dedicated page[59] was created for these resources, where they can be downloaded from.

A SPARQL Apache Jena Fuseki server has been installed on one of our servers. It can upload RDF-Turtle files, similar to those produced in this project, and then allow them to be queried online by providing a SPARQL endpoint. The server can be accessed at https://relate.racai.ro/datasets/. It first presents a list of

[36] https://www.w3.org/TR/mediaont-10/

[37] Ontology-lexicon interface: http://www.w3.org/ns/lemon/ontolex#

[38] https://github.com/globalwordnet/cili/blob/master/ili.ttl

[39] Frequency, attestation and corpus information: http://www.w3.org/ns/lemon/frac#

[40] RDF Schema: http://www.w3.org/2000/01/rdf-schema#

[41] RDF: http://www.w3.org/1999/02/22-rdf-syntax-ns#

[42] http://www.w3.org/ns/lemon/vartrans

[43] http://persistence.uni-leipzig.org/nlp2rdf/ontologies/nif-core\#

[44] http://purl.org/powla/powla.owl\#

[45] http://ufal.mff.cuni.cz/conll2009-st/task-description.html#

[46] https://universaldependencies.org/format.html\#

[47] http://www.w3.org/ns/ma-ont\#

[48] XSD: http://www.w3.org/2001/XMLSchema\#

[49] DCMI Metadata Terms: http://purl.org/dc/terms/

[50] http://xmlns.com/foaf/0.1/

[51] http://isophonics.net/content/studio-ontology

[52] http://www.w3.org/2002/07/owl\#

[53] http://www.w3.org/ns/dcat\#

[54] http://www.w3.org/ns/prov\#

[55] http://pav-ontology.github.io/pav/

[56] http://purl.org/dc/elements/1.1/

[57] http://creativecommons.org/ns

[58] http://schema.org/

[59] https://www.racai.ro/p/llod

available resources and then allows the user to select the desired resource and run a query.

Another way of advertising our resources is registering their metadata in the LOD Cloud. They are now retrieved while browsing the cloud.

4. Interlinking of Romanian LOD Resources

The resources we converted to LOD specifications are either linked to other resources or among themselves.

The use of *wn:ili* property in the representation of RoWN ensures its linking to the other wordnets linked to it. The mapping to ILI was done automatically via the intrinsic mapping to PWN3.0. The mapping of ILI to different versions of PWN[60] is publicly offered by the Global Wordnet initiative. Through *wn:ili* property, 59,348 links were created to concepts in any wordnet linked to ILI.

RoLEX is linked to RoWN via ILI, thus ensuring the phonemic description for the words therein. Lemma forms in RoLEX are linked to all occurrences of the respective word in RoWN. However, homographs (i.e., words spelt identically but pronounced differently) are manually semantically disambiguated and then linked to the RoWN correspondent. The linking was automatic, by matching LexicalEntry labels/URIs in RoLEX with corresponding LexicalEntry labels in RoWN and recovering all the synsets (ontolex:LexicalSense) associated with them. As we mentioned, all the LexicalSense descriptions have a *wn:ili* property, and all these properties were extracted and transported to their corresponding LexicalEntry label/URI in RoLEX. As a LexicalEntry object in RoLEX is matched with more Lexical Entry objects in RoWN (e.g. *RoLEX:lex_pom* is matched with both *RoWN:pom-n-12651821* and *RoWN:pom-n-13104059*), each entry in RoLEX has a list of corresponding ILI.

Location entities in the LegalNERO corpus were mapped to the GeoNames resource, when linking was possible. This was automatically performed by using a lookup script created in the Python language that matched GeoNames entries with Location entities at lemma level. For each of the matched Location entities, the corpus was subsequently enriched with the GeoNames code id as an unique identifier for the GeoNames resource. In this process there were 1,411 Location entities that have been matched with a GeoNames code.

5. Use cases

The power of LOD is seen when combining multiple resources to produce powerful usage scenarios. This interlinking process exploits the internal structure of these resources and the points they have in common. These can be either identifiers (like those in ILI) or specific word forms found in multiple resources (for example the lemma of a word in the treebanks, expressed by the conll:lemma property can be linked to the *ontolex:canonicalForm* property of the ontolex:LexicalEntry class in RoLEX or RoWN).

Having all the resources available as SPARQL endpoints (see Section 3) allows for formulating complex SPARQL queries exploiting multiple datasets, using the SERVICE keyword (otherwise known as a federated query[61]). A powerful example of a complex federated query exploiting multiple resources is a conceptual search in a speech corpus. In this case, the concept is first looked up in RoWN and related concepts are retrieved. Then, the RoLEX lexicon can be employed to extract the word forms associated with the identified concepts, making use of the ILI identifiers. Finally, the resulting words are looked up in the speech corpus (in our case RTASC, see Section 2.5). For the result, the user obtains a list of audio files containing words related to the concept used in the initial query. This is depicted in Figure 1. The associated SPARQL query implementing the process is given in Listing 11. Finally, example results obtained from running the query through the Fuseki-provided SPARQL endpoints is given in Figure 2.

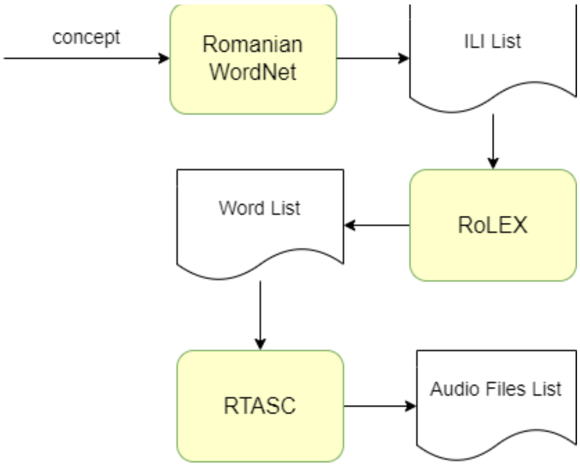


Figure 1. Example concept-based searching in a speech corpus by means of a federated query exploiting multiple Romanian resources

```
SELECT ?ili ?formString ?audio
WHERE {
 { SELECT DISTINCT ?ili1 WHERE {
   ?idConcept ontolex:writtenRep "calculator"@ro .
   ?idLex a ontolex:LexicalEntry .
   ?idLex ontolex:canonicalForm ?idConcept .
   ?idLex ontolex:Sense ?sense .
   ?sense a ontolex:LexicalSense .
   ?sense ontolex:reference ?ref .
   ?ref a ontolex:LexicalConcept .
   ?ref wn:ili ?ili .
   ?x vartrans:source ?ref .
   ?x vartrans:category wn:hypernym .
   ?x vartrans:target ?ref1 .
   ?senser1 ontolex:reference ?ref1 .
   ?ref1 wn:ili ?ili1 .
   ?idr1 ontolex:Sense ?senser1 .
   ?idr1w ontolex:writtenRep ?w .
   ?idr1 ontolex:canonicalForm ?idr1w .
   }}

 SERVICE <https://relate.racai.ro/datasets/rolex/sparql> {
   ?idRolex wn:ili ?ili1 .
   ?idRolex ontolex:lexicalForm ?idRolexForm .
   ?idRolexForm ontolex:writtenRep ?formWritten .
   FILTER (lang(?formWritten)="ro") .
   BIND (str(?formWritten) as ?formString) .
 }
  SERVICE <https://relate.racai.ro/datasets/rtasc/sparql> {
   ?id_tok conllu:FORM ?formString .
   ?id_tok powla:hasLayer ?id_layer .
```

[60] https://github.com/globalwordnet/cili

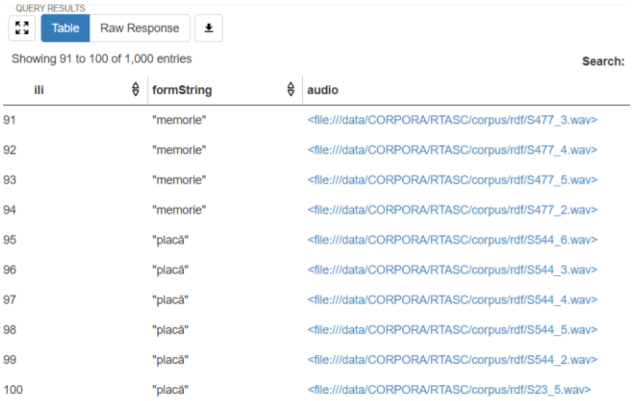
[61] https://www.w3.org/TR/sparql11-federated-query/

```
  ?id_layer powla:hasDocument ?id_doc .

  ?idr a ma:MediaResource .
  ?idr ma:hasSubtitling ?id_doc .
  ?idr ma:hasTrack ?ida .

  ?ida a ma:AudioTrack .
  ?ida ma:hasFragment ?audio .
 }
}
```

Listing 11. Federated query providing concept-based retrieval of speech files, combining the Romanian Wordnet, the RoLEX lexicon and the RTASC speech corpus



Showing 91 to 100 of 1,000 entries

	ili	formString	audio
91		"memorie"	<file:///data/CORPORA/RTASC/corpus/rdf/S477_3.wav>
92		"memorie"	<file:///data/CORPORA/RTASC/corpus/rdf/S477_4.wav>
93		"memorie"	<file:///data/CORPORA/RTASC/corpus/rdf/S477_5.wav>
94		"memorie"	<file:///data/CORPORA/RTASC/corpus/rdf/S477_2.wav>
95		"placă"	<file:///data/CORPORA/RTASC/corpus/rdf/S544_6.wav>
96		"placă"	<file:///data/CORPORA/RTASC/corpus/rdf/S544_3.wav>
97		"placă"	<file:///data/CORPORA/RTASC/corpus/rdf/S544_4.wav>
98		"placă"	<file:///data/CORPORA/RTASC/corpus/rdf/S544_5.wav>
99		"placă"	<file:///data/CORPORA/RTASC/corpus/rdf/S544_2.wav>
100		"placă"	<file:///data/CORPORA/RTASC/corpus/rdf/S23_5.wav>

Figure 2. Example results when searching for concepts related to the word "calculator" ("computer")

Apart from this example, other usage scenarios of different degrees of complexity can be envisaged, exploiting other features of the resources. Thus, one can start with a phonetic search in the RoLEX lexicon (possibly as a result of obtaining phonemes from an ASR system) and then go to RoWN to retrieve concepts. In case of multiple words having similar phonetic representations (as indicated by RoLEX results) one can make use of the frequencies list extracted from the CoRoLa corpus to identify the most likely word form.

It is possible to make use of the interlinking process to query multilingual resources. Thus, one can start with a phonetic query in RoLEX, identify relevant RoWN concepts and then employ the ILI identifier to find similar concepts in another language (or multiple languages). Finally, the result can be used to perform concept-based retrieval in the foreign language(s).

The presence of references to the GeoNames ontology in the LegalNERo corpus opens up the possibility of performing queries filtered by geospatial criteria. These can be performed directly or combined with other resources (RoLEX, RoWN) to obtain more complex results.

6. Conclusions

We have taken important steps towards making a pool of Romanian language resources available to the community in LD format. They have been released in an open manner and are accessible in standard formats, reusing existing vocabularies, and can be queried using a SPARQL endpoint. They are now ready to be exploited to the potential offered by Linked Data. Our conversion of resources to LD specifications has led to what seems to be a duplication of resources in the LLOD Cloud. RoWN was already available from MultiWordNet and RRT was already available as the Romanian treebank in UD. However, the version of RoWN that we have converted is the whole resource that has been created, as opposed to the sample that is available in MultiWordNet. The RRT we converted is a newer version (including some recent corrections) of that already available in the LLOD Cloud.

Some resources require further enhancement by adding extra information: e.g., derivational relations existing between Romanian word senses to be added to the RoWN (Barbu Mititelu, 2012), links between the verbs in RRT and their corresponding senses in the valence lexicon of Romanian verbs (Barbu et al., 2022), links between the verbal multiword expressions in PARSEME-Ro corpus and their corresponding senses in RoWN (Barbu Mititelu et al., 2019b), etc.

The treebanks that are released via UD pose the problem of keeping track of their biannual versions. A solution to this will be seeked.

With regard to the sustainability of the resources, developed tools allow execution on primary data, such as the RoWN development format. Therefore, when new versions of the resources become available, they will be exported to the LOD format. Nevertheless, this is a manual process, requiring a developer to also execute the corresponding LOD export tool.

10. Bibliographical References

Avram, A. M., Păiş, V., and Tufis, D. (2020). Towards a Romanian end-to-end automatic speech recognition based on Deepspeech2. In *Proc. Rom. Acad*. Ser. A, Vol. 21, pp. 395-402.

Barbu, A.-M. (2008). Romanian lexical data bases: Inflected and syllabic forms dictionaries. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, pp. 1937-1941.

Barbu, A.-M., Barbu Mititelu, V., Mititelu, C. (2022). Aligning the Romanian Reference Treebank and the Valence Lexicon of Romanian Verbs. In *Proceedings of LREC 2022* (in press).

Barbu Mititelu, V. (2012). Adding Morpho-Semantic Relations To The Romanian Wordnet. In *Proceedings of LREC2012*, Istanbul, Turkey, pp. 2596-2601.

Barbu Mititelu, V. (2018). Modern Syntactic Analysis of Romanian. In Ofelia Ichim, Luminiţa Botoşineanu, Daniela Butnaru, Marius-Radu Clim, Ofelia Ichim, Veronica Olariu (eds.), *Clasic şi modern în cercetarea filologică românească actuală*, Iaşi, Publishing House of "Alexandru Ioan Cuza" University, pp. 67-78.

Barbu Mititelu, V., Cristescu, M. and Onofrei, M. (2019a). The Romanian Corpus Annotated with Verbal Multiword Expressions. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*, Florence, Italy, pp 13–21.

Barbu Mititelu, V., Stoyanova, I., Leseva, S., Mitrofan, M., Dimitrova, T. and Todorova, M. (2019b). Hear about Verbal Multiword Expressions in the Bulgarian and the Romanian Wordnets Straight from the Horse's Mouth. In *Proceedings of the Joint Workshop on Multiword Expressions and WordNet (MWE-WN 2019)*. Association for Computational Linguistics, Florence, Italy, pp. 2-12.

Barbu Mititelu, V. and Mitrofan, M. (2020). The Romanian Medical Treebank – SiMoNERo. In

Proceedings of the 15th International Conference "Linguistic Resources and Tools for Natural Language Processing", Editura Universității A. I. Cuza, Iași, pp. 7–16.

Boroş, T., Dumitrescu, S.D. and Burtica, 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 Dependencies*, pp. 171-179.

Chiarcos C. and Fäth C. (2017). CoNLL-RDF: Linked Corpora Done in an NLP-Friendly Way. In Gracia J., Bond F., McCrae J., Buitelaar P., Chiarcos C., Hellmann S. (eds) *Language, Data, and Knowledge. LDK*. pp 74-88.

Diewald, N., Hanl, M., Margaretha, E., Bingel, J., Kupietz, M., Banski, P. and Witt, A. (2016). KorAP architecture – Diving in the Deep Sea of Corpus Data. In N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S.Piperidis (Eds.), *Proceedings of the 10th International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia, pp. 3586–3591.

Diewald, N., Barbu Mititelu, V., Kupietz, M. (2019). The KorAP User Interface. Accessing CoRoLa via KorAP. *Revue Roumaine de Linguistique*, 64(3): 265–277.

Fazekas, G. and Sandler, M. B. (2011). The studio ontology framework. In 12th International Society for Music Information Retrieval Conference (ISMIR), pp. 471–476.

Fellbaum, Ch. (1998, ed.). *WordNet: An Electronic Lexical Database*. Cambridge, MA: MIT Press.

Bond, F., Vossen, P., McCrae, J. P., & Fellbaum, C. (2016). Cili: the collaborative interlingual index. In *Proceedings of the 8th Global WordNet Conference (GWC)*, pp. 50-57.

Ion, R. (2007). *Word sense disambiguation methods applied to English and Romanian*. PhD diss., PhD thesis (in Romanian). Romanian Academy, Bucharest.

Ion, R., Badea, V.G., Cioroiu, G., Barbu Mititelu, V., Irimia, E., Mitrofan, M., and Tufiș, D. (2020). A Dialog Manager for Micro-Worlds. In *Studies in Informatics and Control*. 29(4):411-420

Miller, G.A. (1995). WordNet: A Lexical Database for English. *Communications of the ACM*, 38(11): 39-41.

Mitrofan, M. and Tufiş, D. (2018). BioRo: The biomedical corpus for the Romanian language. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Păiş, V. and Tufiş, D. (2018a). Computing distributed representations of words using the CoRoLa corpus. In Proceedings of the Romanian Academy, series A, 403-410.

Păiș, Vasile and Tufiș, Dan. (2018b). More Romanian word embeddings from the RETEROM project. In Proceedings of the International Conference on Linguistic Resources and Tools for Processing Romanian Language - CONSILR, 91-100.

Păiş, V. (2020). Multiple annotation pipelines inside the RELATE platform. In *Proceedings of the 15th International Conference on Linguistic Resources and Tools for Natural Language Processing*. pp. 65--75.

Păiș, V., Tufiș, D. and Ion, R. (2020). A Processing Platform Relating Data and Tools for Romanian Language. In *Proceedings of The 12th Language Resources and Evaluation Conference*. European Language Resources Association, Marseille, France, pp. 81-88.

Păiș, V. and Mitrofan, M. (2021). Towards a named entity recognition system in the Romanian legal domain using a linked open data corpus. In *Workshop on Deep Learning and Neural Approaches for Linguistic Data*. Skopje, North Macedonia, pp. 16--17.

Păiș, V., Mitrofan, M., Gasan, C.L., Coneschi, V. and Ianov, A. (2021a). Named Entity Recognition in the Romanian Legal Domain. In *Proceedings of the Natural Legal Language Processing Workshop 2021*, pp. 9-18.

Păiș, V., Ion, R., Avram, A.M., Irimia, E., Barbu-Mititelu, V., and Mitrofan, M. (2021b). Human-Machine Interaction Speech Corpus from the ROBIN project. In *Proceedings of the 2021 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pp. 91-96.

Păiș, V., Ion, R., Barbu Mititelu, V., Irimia, E., Mitrofan, M., and Avram, A. (2021c). ROBIN Technical Acquisition Speech Corpus. Zenodo, https://doi.org/10.5281/zenodo.4626539.

Ramisch, C., Cordeiro, S.R., Savary, A., Vincze, V., Barbu Mititelu, V., Bhatia, A., Buljan, M., Candito, M., Gantar, P., Giouli, V., Güngör, T., Hawwari, A., Iñurrieta, U., Kovalevskaitė, J., Krek, S., Lichte, T., Liebeskind, C., Monti, J., Parra Escartín, C., QasemiZadeh, B., Ramisch, R., Schneider, N., Stoyanova, I., Vaidya, A. and Walsh, A. (2018). Edition 1.1 of the PARSEME Shared Task on Automatic Identification of Verbal Multiword Expressions. In the *Proceedings of Joint Workshop on Linguistic Annotation, Multiword Expressions and Constructions (LAW-MWE-CxG-2018)*, workshop at COLING 2018, Santa Fe, USA, pp. 222-240.

Ramisch, C. et al., 2020, *Annotated corpora and tools of the PARSEME Shared Task on Semi-Supervised Identification of Verbal Multiword Expressions (edition 1.2)*, LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University, http://hdl.handle.net/11234/1-3367.

Toma, Ş.A, Stan, A., Pura, M.L and Bârsan, T. (2017). MaRePhoR—An open access machine-readable phonetic dictionary for Romanian. In *2017 International Conference on Speech Technology and Human-Computer Dialogue (SpeD)*, pp. 1-6.

Savary, A., Candito, M., Barbu Mititelu, V., Bejček, E., Cap, F., Čéplö, S., Cordeiro, S., Eryiğit, G., Giouli, V., Van Gompel, M., HaCohen-Kerner, Y., Kovalevskaitė, J., Krek, S., Liebeskind, C., Monti, J., Parra Escartín, C., van der Plas, L., Qasemi Zadeh, B., Ramisch, C. and Vincze, V. (2018). PARSEME multilingual corpus of verbal multiword expressions. In S. Markantonatou, C. Ramisch, A. Savary and V. Vincze (Eds.), *Multiword expressions at length and in depth: Extended papers from the MWE 2017 workshop Phraseology and Multiword Expressions*. Berlin: Language Science Press, pp. 87–147.

Stan, A., Yamagishi, J., King, S. and Aylett, M. (2011). The Romanian speech synthesis (RSS) corpus: Building a high quality HMM-based speech synthesis system using a high sampling rate. *Speech Communication* 53, no. 3 (2011): 442-450.

Straka, M., Hajič, J. and Straková, J. (2016). UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, Portorož, Slovenia.

Tufiș, D., Barbu, E., Barbu Mititelu, V., Ion, R., Bozianu, L. (2004a). The Romanian Wordnet. *Romanian Journal* *of Information Science and Technology*, vol. 7, nr. 1-2, pp. 107-124

Tufiş, D., Cristea, D., Stamou, S. (2004b). BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. In *Romanian Journal on Information Science and Technology*, 7(2-3): 9-34.

Tufiş, D. and Barbu Mititelu, V. (2014). The Lexical Ontology for Romanian. In N. Gala, R. Rapp, N. Bel-Enguix (Eds.), *Language Production, Cognition, and the Lexicon*, series Text, Speech and Language Technology, vol. 48. Springer, pp. 491-504.

Tufiș, D., Barbu Mititelu, V., Irimia, E., Păiș, V., Ion, R., Diewald, N., Mitrofan, M., Onofrei, M. (2019). Little Strokes Fell Great Oaks. Creating CoRoLa, the Reference Corpus of Contemporary Romanian. *Revue Roumaine de Linguistique*, 64(3): 227–240.

Tufiș, D., Mitrofan, M., Păiș, V., Ion, R. and Coman, A. (2020). Collection and Annotation of the Romanian Legal Corpus. In *Proceedings of The 12th Language Resources and Evaluation Conference. European Language Resources Association*, Marseille, France, pp. 2766-2770.

Tufiș, D. (2022). Romanian Language Technology – a view from an academic perspective. *International Journal of Computers Communication and Control*, vol. 17, no. 1, 2022.

Zaveri, A., Rula, A., Maurino, A., Pietrobon, R., Lehmann, J., Auer, S. (2015). Quality assessment for Linked Data: A Survey. *Semantic Web*, vol. 7, pp. 63-93.