The Romanian Wordnet in Linked Open Data Format

Elena Irimia and **Verginica Barbu Mititelu**
Romanian Academy Research Institute for Artificial Intelligence
Bucharest, Romania
{elena,vergi}@racai.ro

Abstract

In this paper we present the standardization of the Romanian Wordnet by means of conversion to the Linked Open Data format. We describe the vocabularies used to encode data and metadata of this resource. The decisions made are in accordance with the characteristics of the Romanian Wordnet, which are the outcome of the development method, enrichment strategies and resources used for its creations. By interlinking with other resources, words in the Romanian Wordnet have now the pronunciation associated, as well as syntagmatic information, in the form of contexts of occurrences.

1 Introduction

The Romanian Wordnet (RoWN) as available today has been created starting with the BalkaNet project (Tufiș et al., 2004). The working methodology (Tufiş et al., 2004) followed mainly (see below) the expand approach (Vossen, 1996): synsets from the Princeton WordNet (Miller, 1995; Fellbaum, 1998) (PWN) were translated into Romanian and the relations between implemented synsets were transferred from corresponding PWN synsets. Using a bilingual electronic dictionary, the literals in the selected PWN synsets were first automatically translated and the Romanian equivalents were suggested to lexicographers as literals to be included in the Romanian synsets. For each selected word, its sense was chosen from the parsed electronic version of the Explanatory Dictionary of Romanian (DEX) (Coteanu and Mareș, 1996).

The selection[1] of the synsets to be implemented during BalkaNet was done so as to cover words with high frequency (according to corpora available at that moment), polysemy (according to the number of senses in DEX), as well as avoidance of dangling nodes in the RoWN structure (which meant that choosing a synset to implement in Romanian implied choosing all its unimplemented synsets in PWN up to the unique beginners of the hierarchies, in the case of nouns and verbs, which have a hierarchical structure). The synsets IDs were also transferred from PWN.

The BalkaNet team also aimed at reflecting some of the specificities of this geographic and cultural region in the wordnets under development. Consequently, a various number of such synsets were included in the wordnets: for Romanian, there were 541 synsets. They were included in the hierarchies mostly as hyponyms of existing synsets. Their IDs were generated so as to keep them distinct from those of the translated synsets. One such synset contains the literal *tobă* with the gloss "a type of cold cooked meat, containing pieces of chopped meat, fat, offal, all stuffed in a pig's stomach and suspended din aspic". It is a Romanian traditional cold dish, specific to Christmas season and looking like a wide sausage. For this reason it is a hyponym of the noun *cârnat*, which translates the English *sausage*.

Besides the automatic transfer of the semantic relations holding between equivalent English synsets, the Romanian team also transferred the lexical relations from PWN: these are relations marked at the literal (not synset) level in PWN. Examples include antonymy and derivation relations. In the case of the former, it was considered that this lexical relation has a conceptual counterpart: the semantic opposition between the concepts lexicalized by the words in antonymy relations (Miller, 1995). Consider the synsets {*sterile*, *unfertile*, *infertile*} (gloss: "incapable of reproducing") and the synset {*fertile*} (gloss: "capable of reproducing"). Antonymy is marked between *sterile* and *fertile* in PWN. However, speakers understand a semantic opposition between *fertile* and *infertile*, as well as between *fertile* and *unfertile*[2]. Given that there is no literal

[1] Further selections, in other projects in which the RoWN was enriched, were made so as to ensure the lexical coverage required by the respective projects.

[2] See this example: "By contrast, fertility is the ac-

```
<SYNSET>
  <ID>ENG30-09448090-n</ID>
  <POS>n</POS>
  <SYNONYM>
    <LITERAL>stratosferă
      <SENSE>1</SENSE>
    </LITERAL>
  </SYNONYM>
  <DEF>Stratul superior al atmosferei
  (situat deasupra troposferei),care
  începe la o înălțime de aproximativ 11 km
  de la suprafața Pământului. </DEF>
  <ILR>ENG30-08591680-n
    <TYPE>hypernym</TYPE>
  </ILR>
  <ILR>ENG30-09210604-n
    <TYPE>part_holonym</TYPE>
  </ILR>
</SYNSET>
```

Table 1: One synset associated to the literal "stratosferă" (en. "stratosphere")

correspondence between PWN and RoWN, which could have allowed for the transfer of antonymy at literal level, this assumption allowed for its transfer at the synset level. Thus, the RoWN equivalents of such synsets establish between them an antonymy relation. Table 1 shows an example of a RoWN synset in the original XML format.

2 Conversion of RoWN to LOD format

Linked Data (LD) refers to a set of best practices in publishing structured data on the Web (Chiarcos et al., 2013). When an open type of license, namely Creative Commons (CC), is associated with a resource, then we talk about *linked open data* (LOD). The conversion of Romanian language resources to the LOD format is an internal project[3] of the Romanian Academy Research Institute for Artificial Intelligence, running in parallel with the NexusLinguarum COST Action[4]. We have already made several resources available in this format and, more important, this way some of them are made open to the community for the first time: this is also the case with RoWN, of which only a core has been freely available throughout time.

Original format	LOD format
domain_member_TOPIC	domain_topic
cause	causes
entailment	entails
domain_member_REGION	has_domain_region
domain_member_TOPIC	has_domain_topic
member_holonym	holo_member
part_holonym	holo_part
substance_holonym	holo_substance
member_meronym	mero_member
part_meronym	mero_part
substance_meronym	mero_substance
similar_to	similar
near_antonym	antonym

Table 2: Renaming of synset relations to comply to the LOD standards

The LOD format for RoWN was automatically generated using a conversion tool developed in C#. Preliminary actions that had to be taken were: (1) mapping RoWN to the CILI[5] IDs (through the PWN mapping) to enable its linking to the international network of wordnets[6] mapped to CILI, and (2) renaming some lexical and semantic relations to correspond to the LOD guidelines (see Table 2 for the renamed relations; the following relations kept their original name: attribute, hypernym, hyponym, instance_hypernym, instance_hyponym).

In accordance with the recommended standard for representing wordnets, our Turtle RDF LOD representation model is mainly based on the *OntoLex-Lemon* vocabulary (Cimiano et al., 2016) developed by the Ontology-Lexica community group (OntoLex), but is also supported by other useful vocabularies like the OWL Web Ontology Language[7], the wordnet specific ontology *wn*[8] and the variation and translation lemon module *vartrans*[9] to represent the various encoded properties. The serialisations in LMF-XML and JSON format are also available, but the linking with external resources was implemented only in the Turtle RDF format, which will, therefore, be the focus of this section.

tual production of live offspring and is the antonym of infertility", https://academic.oup.com/humrep/article/19/7/1497/2356621, accessed 19th Dec, 2022.

[3] https://www.racai.ro/p/llod/index_en.html

[4] https://nexuslinguarum.eu/

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

[6] Open Multilingual Wordnet (Bond and Foster, 2013)

[7] https://www.w3.org/TR/owl-guide/

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

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

As can be seen in Figure 1, the main level entry in the original XML format of RoWN was the synset, comprising an ID, the part-of-speech label (POS), a definition, the synonym set and a list of relations specified by their target synset (ILR1 and ILR2 objects) and their relation type (type1, type2). The synonym set was a list of different literals together with their associated senses, unique to the synset they belong to.

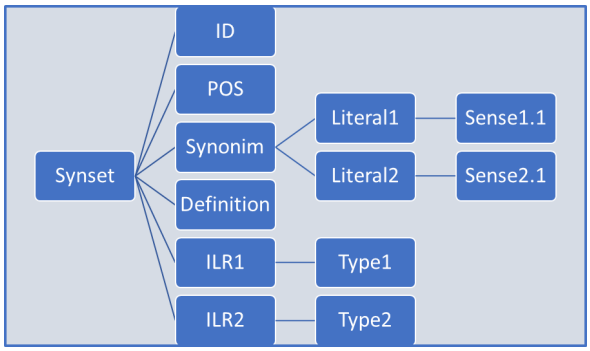

Figure 1: Diagram of the objects in the original XML format of RoWN

To comply with the OntoLex-Lemon model[10], the information in RoWN had to be restructured as shown in Figure 2. The color code for the nodes in the diagram is the following: *blue* stands for objects, *yellow* for properties and the correspondence between the new classes and properties and the ones in the original format (see Figure 1) are marked in *red*: e.g., each LexicalEntry in RoWN has an associated canonicalForm object and the ontolex:writtenRep property of this object has as value one of the literals in the synonym set of one of the original format synsets.

Basically, the information in the original file was organised around synonym sets (with specific meaning), accompanied by their associated lexical representations (literal1, literal2, etc.), while in the LOD format the data is organised around literals, accompanied by their possible meanings (represented as a list of senses: sense1, sense2, etc.).

The new format has four types of main entries:

- *ontolex:LexicalEntry*. Each LexicalEntry, representing a specific literal, has an associated ontolex:CanonicalForm object with an ontolex:writtenRep property and a list of declarations for ontolex:Sense objects that specify possible senses of the literal.

[10]see the guidelines at https://www.w3.org/2016/05/ontolex/

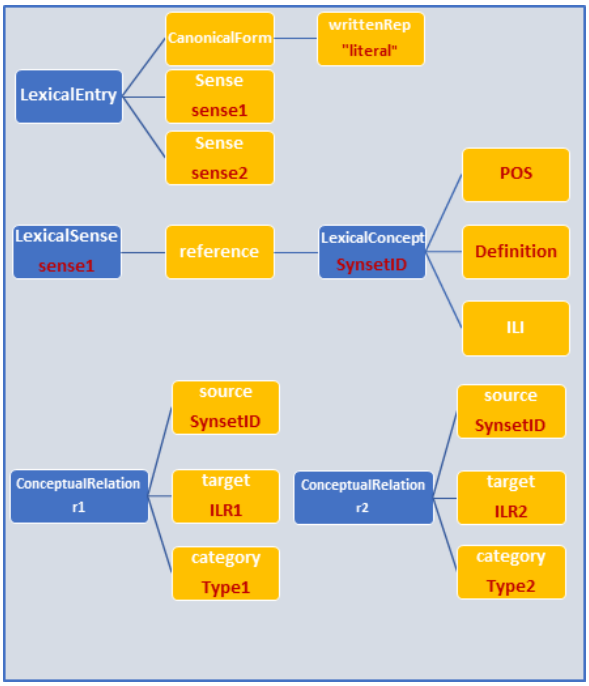

Figure 2: Diagram of the objects and properties used to represent information in the LOD format of RoWN, with correspondences with the original XML object labels (see Figure 1) marked in red.

- *ontolex:LexicalSense*. Each ontolex:Sense object is then described as a separate entry through an ontolex:reference to an ontolex:LexicalConcept whose value is a synset ID (previously copied in RoWN from PWN).

- *ontolex:LexicalConcept*. The LexicalConcept has, in turn, an associated part-of-speech (POS) description and a definition, encoded through wn:partOfSpeech and wn:definition, respectively. The recent ILI mapping is specified through the wn:ili property.

- a list of *vartrans:ConceptualRelation* objects associated to a specific LexicalConcept, encoding all the relations with other lexical concepts (synsets) in RoWN; the vartrans:target and vartrans:category properties are used to describe the relation's target synset and the relation type.

Table 3 shows the information in Table 1 (i.e., the XML representation of the concept *stratosferă* (EN. 'stratosphere')) converted to the LOD specifications.

```
<#stratosferă-n> a ontolex:LexicalEntry;
   ontolex: canonicalForm [
      ontolex:writtenRep "stratosferă"];
   wn: partOfSpeech wn:n;
   ontolex:Sense <#stratosferă-n-09448090-1>.

<#stratosferă-n-09448090-1> a
                  ontolex:LexicalSense;
   ontolex:reference <#09448090-n>.

<#09448090-n> a ontolex:LexicalConcept;
   wn:partOfSpeech wn:n ;
   owl:sameAs ili:i86260 ;
   wn:definition [
   rdf:value "Stratul superior al atmosferei (si-
      tuat deasupra troposferei), care începe la
      o înălțime de aproximativ 11 km de la su-
      prafața Pământului."@ro].

<#09448090-n-r1> a vartrans:ConceptualRe-
lation
vartrans:source <#09448090-n> ;
vartrans:category wn:hypernym ;
vartrans:target <#08591680-n> .

<#09448090-n-r2> a vartrans:ConceptualRe-
lation
vartrans:source <#09448090-n> ;
vartrans:category wn:holo_part ;
vartrans:target <#09210604-n> .
```

Table 3: The information associated to "stratosferă" in the LOD format

Object type	No. of objects
Lexical Entry	52,802
LexicalSense	85,277
LexicalConcept	59,348
Semantic Relation	138,592
CILI link	59,348
RoLEX sameAs link	16,196

Table 4: Statistics of objects and links in LOD RoWN

3 Interlinking

One of the important advantages LOD comes with is the possibility of putting language resources in a broader context, by means of interlinking them, which further ensures their FAIR characteristics (Wilkinson et al., 2016).

3.1 Other wordnets

As already mentioned, a mapping of each synset in RoWN to CILI IDs was done by exploiting the mapping of RoWN to PWN 3.0 and the public availability of a PWN3.0-CILI mapping[11]. A total of 59,348 concepts from RoWN are, at the moment, linked to the corresponding concepts in any wordnet linked to CILI. The property *owl:sameAs* has also recently been used to directly link synsets in the LOD reprtesentation of RoWN and PWN 3.0.

3.2 RoLEX

RoLEX (Lőrincz et al., 2022) is a Romanian lexicon of 330,000 word forms having associated information about their lemma, morphosyntactic description (MSD[12]), syllabification, lexical stress and phonemic transcription with an extended version of Speech Assessment Methods Phonetic Alphabet[13] (SAMPA) for Romanian. An entry in the tabular format of RoLEX is presented in Table 5.

The original 6-column tabular format of RoLEX was also converted to LOD using the same OntoLex-Lemon model. Lemmas in the tabular format became ontolex:LexicalEntries that have a list of associated ontolex:lexicalForms. In turn, each lexicalForm has the MSD encoded using the POS property in the conll vocabulary and the remaining information described by the ontolex:writtenRep[14] and the ontolex:phoneticRep[15] properties.

In the Turtle RDF LOD version of RoLEX, a linking to ROWN was implemented by associating possible corresponding CILI IDs to each LexicalEntry. LexicalEntry labels in RoLEX were automatically matched with LexicalEntry labels in RoWN, and via all the associated LexicalSenses and respective LexicalConcepts, the corresponding CILI IDs were retrieved and encoded in RoLEX.

[11] https://github.com/globalwordnet/cili/blob/master/ili-map-pwn30.tab

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

[13] https://www.phon.ucl.ac.uk/home/sampa/

[14] "stratosfera"@ro, "stra.to.sfe.ra"@syl, "stratosf'era"@stress

[15] "s t r a t o s f e r a"@ro-RO-sampa

Column type	Value
word-form	stratosfera
lemma	stratosferă
MSD	Ncsfry
syllabification	stra.to.sfe.ra
stress marking	stratosf'era
phonetic transcription	s t r a t o s f e r a

Table 5: The tabular entry associated to the wordform "stratosfera", the singular nominative-accusative definite form of the lemma "stratosferă".

Recently, a direct linking of RoWN and RoLEX was also implemented, through LexicalEntry matching and using the owl:sameAs property. The matching was done by ignoring clitic pronouns (*o*, *i*, *se*, *-și*) existing in the labels associated to verbal entries in RoWN but being absent from RoLEX: 872 verbal reflexive and pronominal lemmas have been linked to their transitive forms. A number of 16,492 compound lexical entries in RoWN were not matched at all and therefore not linked to entries in RoLEX.

By linking these two resources, 16,196 literals in RoWN have a great deal of new linguistic information associated: their full inflected paradigms are now accessible, altogether with the respective morphosyntactic description, the pronunciation of each form, its syllabification, and the position of lexical stress in each form. Table 4 shows number of objects and links in the LOD RoWN format.

4 Use case scenarios

LD provides mechanisms for exploiting the resources' content, by means of their common elements; these are either identifiers (see ILI) or words co-occurring in several resources. The resources we have converted to LOD format are made available for querying as SPARQL endpoints[16]. This allows for federated queries[17] to be created and, thus, exploit the content of all these resources or only some of them. Such an example would be a conceptual search in a speech corpus, as described by Barbu Mititelu et al. (2022). The following steps are taken: (i) the input word (i.e., a possible lexicalization of a concept of interest) is looked up in RoWN and conceptually identical words (i.e., literals in the same synset, or synonyms) are retrieved; (ii) for each literal, its RoLEX entries are found by means of the ILI identifiers, and thus its inflectional paradigm is retrieved; (iii) these forms are then located in the files of a speech corpus.

The interlinked RoWN and RoLEX prove their usefulness in a Question Answering scenario related to COVID-19 (Ion et al., 2022). An important element for the system being able to find an answer in a set of documents was for it to be able to recognize all the various ways in which a question can be formulated. After the manual creation of several such possible formulations, two steps were taken for expanding them: (i) content words were associated to other semantically related words (synonyms, hypernyms) in RoWN by exploiting the semantic relations therein, and (ii) these newly found words were associated with their inflected forms from RoLEX, also taking advantage of the fact that the interlinking between these two resources was made with manual assignment in the case of homographs. The POS-tagging of the question and the morphosyntactic descriptions in RoLEX helped to find the inflected form necessary in each context.

5 Access to LOD RoWN

The LOD format of RoWN is available for download on the website of the internal LOD project (see Section 4). This is the first time the whole Romanian Wordnet is made freely available for download. Previously (Pianta et al., 2002), only a core of it was accessible. Only by means of the dedicated API (Dumitrescu et al., 2018) could any kind of information therein be exploited. A SPARQL endpoint is also made available for it on the SPARQL Apache Jena Fuseki server installed on one of our servers. The resource's metadata has already been registered in the LOD Cloud[18], as well as in the European Language Grid catalogue[19].

6 Conclusions and future work

Although not currently under development, RoWN is still considered a valuable resource for the Romanian language, as shown by its recent use in a query expansion task (Ion et al., 2022) and its evaluation in a word similarity task (Barbu and Barbu Mititelu, 2022).

We have presented here its conversion to LOD specifications, a new format that can help RoWN

[16] https://relate.racai.ro/datasets/

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

[18] https://lod-cloud.net/#

[19] https://live.european-language-grid.eu/catalogue/search/Romanian%20wordnet

become a more FAIR resource. In the future, we are going to add the Balkan-specific concepts and derivational relations to the LOD RoWN and then reuse the resource in its interlinked format in various scenarios.

References

Eduard Barbu and Verginica Barbu Mititelu. 2022. Evaluating computational models of similarity against a human rated dataset. *Baltic Journal of Modern Computing*, 10(3):295–306.

Verginica Barbu Mititelu, Elena Irimia, Vasile Pais, Andrei-Marius Avram, and Maria Mitrofan. 2022. Use case: Romanian language resources in the LOD paradigm. In *Proceedings of the 8th Workshop on Linked Data in Linguistics within the 13th Language Resources and Evaluation Conference*, pages 35–44, Marseille, France. European Language Resources Association.

Francis Bond and Ryan Foster. 2013. Linking and extending an open multilingual wordnet. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1352–1362.

Christian Chiarcos, John McCrae, Philipp Cimiano, and Christiane Fellbaum. 2013. *Towards Open Data for Linguistics: Linguistic Linked Data*, pages 7–25. Springer Berlin Heidelberg, Berlin, Heidelberg.

Philipp Cimiano, John P McCrae, and Paul Buitelaar. 2016. Lexicon model for ontologies: Community report. *W3C Ontology-Lexicon Community Group*.

Ion Coteanu and Lucreția Mareș, editors. 1996. *Dicționarul explicativ al limbii române, ediția a II-a*. Editura Univers Enciclopedic, Bucharest.

Stefan Daniel Dumitrescu, Andrei Marius Avram, Luciana Morogan, and Stefan-Adrian Toma. 2018. Rowordnet–a python api for the romanian wordnet. In *2018 10th International Conference on Electronics, Computers and Artificial Intelligence (ECAI)*, pages 1–6. IEEE.

Christiane Fellbaum. 1998. *WordNet: An Electronic Lexical Database*. MIT Press, Cambridge, MA.

Radu Ion, Andrei-Marius Avram, Vasile Păiș, Maria Mitrofan, Verginica Barbu Mititelu, Elena Irimia, and Valentin Badea. 2022. An open-domain QA system for e-governance. In *Proceedings of the Fifth International Conference Computational Linguistics in Bulgaria (CLiB)*, pages 105–112.

Beáta Lőrincz, Elena Irimia, Adriana Stan, and Verginica Barbu Mititelu. 2022. RoLEX: The development of an extended Romanian lexical dataset and its evaluation at predicting concurrent lexical information. *Natural Language Engineering*, page 1–26.

George A. Miller. 1995. Wordnet: A lexical database for english. *Communications of the ACM*, 38(11):39–41.

Emanuele Pianta, Luisa Bentivogli, and Christian Girardi. 2002. MultiWordNet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*.

Dan Tufiş, Eduard Barbu, Verginica Barbu Mititelu, Radu Ion, and Luigi Bozianu. 2004. The romanian wordnet. *Romanian Journal of Information Science and Technology Special Issue*, 7(1–2):107–124.

Dan Tufiș, Dan Cristea, and Sofia Stamou. 2004. Balkanet: Aims, methods, results and perspectives. a general overview. *Romanian Journal of Information Science and Technology Special Issue*, 7(1–2):9–43.

Piek Vossen. 1996. Right or wrong: Combining lexical resources in the eurowordnet project. In *Proceedings of the 7th EURALEX International Congress*, pages 715–728, Göteborg, Sweden. Novum Grafiska AB.

Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific data*, 3:1–9.