

Romanian WordNet 3.0

Table of Contents

1	BASIC INFORMATION	2
1.1	Resource composition	2
1.2	Representation of the resource (flat files, database, markup)	2
1.3	Character encoding	2
2	ADMINISTRATIVE INFORMATION	2
2.1	Contact person (name, address, affiliation, position, telephone, fax, e-mail)	2
2.2	Delivery medium (if relevant; description of the content of each piece of medium)	2
2.3	Copyright statement and information on IPR	3
3	TECHNICAL INFORMATION	3
3.1	Directories and files	3
3.2	Data structure of an entry	3
3.3	Resource size (no. of rules, MB occupied on disk).....	4
4	CONTENT INFORMATION	4
4.1	Type of the resource.....	4
4.2	The natural language(s) for the resource is applicable (if language dependent).....	4
4.3	Domain(s)/register(s) of the corpus	4
4.4	Annotations in the corpus (if an annotated corpus)	4
4.5	Intended application of the resource	5
4.6	Reliability of the annotations (automatically/manually assigned) – if any	5
5	REFERENCES	6

1 BASIC INFORMATION

1.1 Resource composition

Ro-WordNet (RoWN) (Tufiş et al., 2008; Tufiş et al., 2004b) is a lexical ontology following the Princeton WordNet (PWN) (Miller, 1990; Fellbaum, 1998) organizational principles. The synsets in RWN are aligned with PWN3.0 (Tufiş et al., 2012) and, additionally, they are associated with SUMO/MILO (Neal & Pease, 2001) concepts and labeled with DOMAINS3.0 (Bentivogli et al., 2004) categories.

1.2 Representation of the resource (flat files, database, markup)

RoWN is distributed as an XML file, observing the structure of EuroWordNet (Vossen, 1998) and BalkaNet wordnets (Tufiş, 2004; Tufiş et al., 2004a). The file can be loaded and browsed in VisDic (Horak & Smrz, 2004) (as well as in its descendant versions), the official editor and browser of the BalkaNet project.

1.3 Character encoding

The characters have been encoded in UTF8.

2 ADMINISTRATIVE INFORMATION

2.1 Contact person (name, address, affiliation, position, telephone, fax, e-mail)

NAME: Dan TUFİŞ

ADDRESS: Calea 13 Septembrie, No. 13, CASA ACADEMIEI, Bucharest 050711, ROMANIA

AFFILIATION: Research Institute for Artificial Intelligence, Romanian Academy

POSITION: Director

TELEPHONE: +40213 188 103

FAX: +40213 188 142

E-MAIL: tufis@racai.ro / danstef@racai.ro

2.2 Delivery medium (if relevant; description of the content of each piece of medium)

The resource will be uploaded on the MetaShare platform as an archive. It can be downloaded from: <http://nlptools.racai.ro/nlptools/index.php?page=rowan>

2.3 Copyright statement and information on IPR

The resource is free license-based for research purposes and fee license-based for commercial purposes.

3 TECHNICAL INFORMATION

3.1 Directories and files

RoWordNet_3.0 – the directory containing the following files:

- wnrom_3_v2.0.xml – the proper Romanian WordNet file
- .\VisDic_cfg_file\wnrom_3_v2.0.cfg – the VisDic configuration file for RoWN

The VisDic editor and browser (if needed) can be freely downloaded from <http://nlp.fi.muni.cz/projects/visdic/>.

3.2 Data structure of an entry

The structure of an entry in RWN is exemplified below:

```
<SYNSET>
  <ID>ENG30-xxxxxxxx-C </ID>
  <POS>cat</POS>
  <SYNONYM>
    [<LITERAL>literal
  <SENSE>k</SENSE>
  </LITERAL>]+
  </SYNONYM>
  <DEF> a definition </DEF>
  [<BCS>n</BCS>]
  [<ILR>synset-ID<TYPE>name-of-relation</TYPE></ILR>]+
  [<DOMAIN>a domain</DOMAIN>]+
  [<SUMO>a sumo-concept<TYPE> a type of mapping</TYPE></SUMO>]
</SYNSET>
```

The structure of an entry for a non-lexicalized synset is the following:

```
<SYNSET>
  <ID>ENG30-xxxxxxxx-C </ID>
  <POS>cat</POS>
  <NL>yes</NL>
  <SYNONYM></SYNONYM>
  <DEF> a definition </DEF>
  [<BCS>n</BCS>]
  [<ILR>synset-ID<TYPE>name-of-relation</TYPE></ILR>]+
  [<DOMAIN>a domain</DOMAIN>]+
  [<SUMO>a sumo-concept<TYPE> a type of mapping</TYPE></SUMO>]
</SYNSET>
```

3.3 Resource size (no. of rules, MB occupied on disk)

The current (validated) version contains 59,015 synsets, with the following distribution:

Noun synsets	Verb synsets	Adj. synsets	Adv. synsets	Total
41,063	10,064	4,822	3,066	59,015

The needed disk space is about 29.3 Mb.

4 CONTENT INFORMATION

4.1 Type of the resource

This is a language-dependent resource (Romanian).

4.2 The natural language(s) for the resource is applicable (if language dependent)

The language of the lexical ontology is Romanian. Via alignment with PWN, it is virtually a bilingual English-Romanian dictionary.

4.3 Domain(s)/register(s) of the corpus

This resource falls into General Domain.

4.4 Annotations in the corpus (if an annotated corpus)

There are four types of entries, all of them having the same structure: entries for nouns, for verbs, for adjectives and for adverbs.

See section 3.2:

The value of the <ID> tag is a unique identifier for the aligned synset in PWN3.0 (the numerical value is the offset of the respective synset in the PWN database). The trailing character C in the ID value is one of N, V, A, R.

The value of the <POS> is one of the N, V, A, R (identical to the character C) identifying the part of speech of the literals in the current synset. One should notice that in the Romanian wordnet the adjectival satellites (marked with the category S in PWN) are included into the A category.

Under the tag <SYNONYM> there are one or more <LITERAL> immediately followed by a sense number. Unlike in PWN, here the numbering is not related to the frequency of the respective sense of the literal, but it follows the numbering conventions from the Romanian Explanatory Dictionary (DEX), the reference dictionary by the Romanian Academy. In the case of non-lexicalized concepts, the tag <SYNONYM> is empty.

The tag <DEF> marks up the definition from DEX. In some cases (namely when the respective sense was not documented in DEX, the definition is a professional translation of the corresponding PWD definition).

The <BCS> tag is optional and it marks up the so called base concept synsets. The value of the tag is 1, 2 or 3, according to what was called in BalkaNet BCS1, BCS2 and BCS3 synsets (see Tufiş et al, 2004).

The current synset entry contains one or more relations towards other synsets. This information is encoded as: [<ILR>synset-ID<TYPE>name-of-relation</TYPE></ILR>]+ where the <ILR> tag (Internal Language Relation) uniquely identifies the target synset of the relation specified by the tag <TYPE>. The relations are transferred from PWN3.0.

The tag <DOMAIN> is one of the labels specified by the DOMAINS-3 taxonomy and it was imported from the PWN3.0 via synset alignment, as well.

The tag <SUMO> marks up the SUMO/MILO concept transferred from the SUM/MILO - PWN3.0 alignment via PWN-RWN synset alignment. The tag <TYPE> embedded into content of the <SUMO> tag describes the type of mapping: “=” defining exact mapping and “+” defining an approximate mapping (the SUMO concept is more general than the meaning of the current entry).

The <NL> tag signals the non-lexicalized concepts in Romanian. For them there is no literal in the <SYNONYM> tag, but there is a gloss.

The design procedure of the RWN followed the *conceptual density principle* (Tufiş et al., 2004) in a top-down strategy and the literals chosen for implementation were selected on the basis of frequency and definitional productivity (the number of entries in DEX definitions containing the specific literals). The lexical stock covers the basic general language vocabulary of Romanian.

4.5 Intended application of the resource

The lexical ontology has been used in practically all NLP-enhanced applications developed at RACAI: tagging, lemmatization, word-sense disambiguation (Tufiş et al., 2004c), word alignment, collocation extraction, document classification, question-answering, machine translation and others (Tufiş & Barbu, 2004).

4.6 Reliability of the annotations (automatically/manually assigned) – if any

The lexical ontology has been based on several reference published dictionaries: Explanatory Dictionary of Romanian, Dictionary of Synonyms, Dictionary of Antonyms. The mapping to the translation equivalent synsets from Princeton WordNet has been manually done by experienced lexicographers and NLP researchers. Based on the manual synset alignment, the semantic

relations have been automatically transferred from PWN onto RWN, while the lexical relations were transferred (when it was possible) under the validation of a lexicographer.

5 REFERENCES

References on the Romanian WordNet:

Dan Tufiș, Verginica Barbu Mititelu, Dan Ștefănescu, Radu Ion. (2012). *The Lexical Ontology for Romanian*. In Nicoletta Calzolari and Nancy Ide (eds.): *Language Resources and Evaluation, Special Issue on Wordnets*, Springer 2012, ISSN 1574-020X.

Dan Tufiș, Radu Ion, Luigi Bozianu, Alexandru Ceaușu, Dan Ștefănescu. (2008). *RO-Wordnet*. In *Proceedings of the 4th Global WordNet Association Conference*, Szeged, Hungary.

Dan Tufiș (ed.). (2004). *Special Issue on BalkaNet*, Romanian Academy, vol7, no. 2-3, ISSN 1453-8245 .

Dan Tufiș, D. Cristea, S. Stamou. (2004a). In *BalkaNet: Aims, Methods, Results and Perspectives. A General Overview*. In *Romanian Journal on Information Science and Technology*, Dan Tufiș (ed.) Special Issue on BalkaNet, Romanian Academy, vol7, no. 2-3, 2004, pp. 9-34, ISSN 1453-8245

Dan Tufiș, Eduard Barbu, Verginica Barbu Mititelu, Radu Ion, Luigi Bozianu. (2004b). *The Romanian Wordnet*. *Romanian Journal on Information Science and Technology*, vol. 7, no. 2-3 (pp. 107-124).

Dan Tufiș, Radu Ion, Nancy Ide. (2004c). *Word sense disambiguation as a wordnets validation method in Balkanet*. In *Proceedings of the 4th LREC Conference*, Lisbon, 741-744; 1071-1074

Dan Tufiș, Eduard Barbu. (2004). *A Methodology and Associated Tools for Building Interlingual Wordnets*. In *Proceedings of LREC2004*, Lisbon, Portugal (pp. 1067-1070).

References to Princeton WordNet, DOMAINS taxonomy and Sumo/MILO ontology:

Horak, A., Smrz, P. (2004). *VisDic - Wordnet Browsing and Editing Tool*. In *Proceedings of the Second International WordNet Conference - GWC 2004*. Brno, Czech Republic : Masaryk University, 2003. pp. 136-141. ISBN 80-210-3302-9.

Bentivogli, L., Forner, P., Magnini, B., Pianta, E. (2004). *Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing*. *Proceedings of COLING 2004 Workshop on "Multilingual Linguistic Resources"* (pp. 101-108).

Niles, I., Pease, A. (2001). *Towards a Standard Upper Ontology*. *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems* (pp. 2-9).

Miller, G.A., Beckwith, R., Fellbaum, C., Gross, D., Miller, K. (1990). *Introduction to WordNet: An On-Line Lexical Database*. *International Journal of Lexicography*, Vol. 3, No. 4 (pp. 235-244).

Fellbaum, Ch. (Ed.) (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.

Vossen, P. (Ed.) (1998). *EuroWordNet: A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers.