ROMANIAN WORDNET 2.0

1. BASIC INFORMATION

   1.1 *Lexicon type (wordform, explanatory, terminological lexicon, wordnet, etc),*
   Ro-WordNet (RWN) is a lexical ontology following the Princeton WordNet (PWN) organizational principles. The synsets in RWN are aligned with PWN3.0 and, additionally, they are associated with SUMO/MILO concepts and labeled with DOMAINS3.0 categories.

   1.2 *Representation of the lexicon (flat files, database, markup)*
   RWN is distributed as an XML file, observing the structure of BalkaNet wordnets. The file can be loaded and browsed in VisDic (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: Dan Tufis
   Address: Calea 13 Septembrie, no. 13, 050711
   Affiliation: Research Institute for Artificial Intelligence, Romanian Academy
   Position: Director
   Telephone: +4021 3188103
   Fax: +40 21 3188142
   e-mail: tufis@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.

   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
   WNROM – the directory containing the following files:
   wnrom.xml – the proper Romanian WordNet file
   wnrom.cfg – the VisDic configuration file for RWN
   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 Lexicon size (nmb. of lexical items, KB occupied on disk)

The current (validated) version contains 30,006 synsets, with the following distribution:

Noun synsets	**Verb synsets**	**Adj. synsets**	**Adv. synsets**	**Total**
21158	7163	851	834	30006

The needed disk space is about 14 000 Kb.

4. CONTENT INFORMATION

4.1 The natural language(s) of the lexicon

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

4. 2 Entry Type

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

4.3 Attributes and their values

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, R, A.

The value of the <POS> is one of the N, V, R, A (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 Tufiș 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.

4.4 Coverage of the lexicon

The design procedure of the RWN followed the *conceptual density principle* (Tufiș 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 lexicon

The lexical ontology has been used in practically all NLP-enhanced applications developed at RACAI: tagging, lemmatization, word-sense disambiguation, word alignment, collocation extraction, document classification, question-answering, machine translation.

4.6 POS assignment

The part of speech assignment is the one in the Explanatory Dictionary of Romanian.

4.7 Reliability (automatically/manually constructed)

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 was possible) under the validation of a lexicographer.

5. RELEVANT REFERENCES AND OTHER INFORMATION

References on the Romanian WordNet

1. Dan Tufiş. Radu Ion, Luigi Bozianu, Alexandru Ceauşu, Dan Ştefănescu: RO-Wordnet. In *Proceedings of the 4th Global WordNet Association Conference*, January 22-25, 2008, Szeged, Hungary
2. Dan Tufiş (ed.) *Special Issue on BalkaNet, Romanian Academy*, vol7, no. 2-3, 2004, ISSN 1453-8245
3. Dan Tufiş, D. Cristea, S. Stamou. BalkaNet: Aims, Methods, Results and Perspectives. A General Overview. In *Romanian Journal on Information Science and Technology*, Dan Tufiş (ed.) Special Issue on BalkaNet, Romanian Academy, vol7, no. 2-3, 2004, pp. 9-34, ISSN 1453-8245
4. Tufiș, D., Barbu, E., Barbu Mititelu, V., Ion, R., Bozianu, L. (2004b). The Romanian Wordnet. *Romanian Journal on Information Science and Technology*, vol. 7, no. 2-3 (pp. 107-124).
5. Dan Tufiş, Radu Ion, Nancy Ide. Word sense disambiguation as a wordnets validation method in Balkanet. In *Proceedings of the 4th LREC Conference*, Lisbon, 2004, 741-744; 1071-1074
6. Tufiș, D. & Barbu, E. (2004). A Methodology and Associated Tools for Building Interlingual Wordnets. In *Proceedings of LREC2004*, Lisbon, Portugal (pp. 1067-1070).

A larger version (not entirely validated) of Romanian WordNet can be browsed at the web address www.racai.ro/wnbrowser.

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

1. 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).
2. 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).
3. 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).
4. Fellbaum, Ch. (Ed.) (1998). *WordNet: An electronic lexical database*. Cambridge, MA: MIT Press.
5. Vossen, P. (Ed.) (1998). *A Multilingual Database with Lexical Semantic Networks*. Dordrecht: Kluwer Academic Publishers.

The Princeton WordNet 3.0 can be freely downloaded from the address: http://wordnet.princeton.edu/wordnet/download/ and can be used on-line at the address: http://wordnetweb.princeton.edu/perl/webwn