Verginica Barbu Mititelu
(Bucharest)

THE LATEST DEVELOPMENTS OF THE ROMANIAN WORDNET

Abstract: The wordnet-style lexicography has imposed itself as one of the most representative and useful method for the organization of lexical knowledge, both from the human user's and the computational perspective. In this article we present the Romanian wordnet: methodology, content, up-to-date statistics. We lay special emphasis on the latest activity carried on in its almost twelve-year ceaseless development, namely the addition of morphologic and semantic links between pairs of words of the type root word – derived word. In our undertaking we followed two different approaches: a language-internal one and a transfer one. Adding such relations in our wordnet leads to the increase in the density of semantic and of lexical relations, thus making the resource more appropriate for various applications.

Key words: wordnet, Romanian language, derivation, semantic relations, lexical relations

Introduction

Our everyday life has been invaded by various electronic applications that facilitate our search for information, that help us find answers to questions and even understand texts written in languages from which we do not know any word. The performance of these applications depends greatly on the language resources they use.

The linguists are confronted now with the challenging task of developing resources in electronic format. From lexicography to morphology, syntax, dialectology, language teaching, all linguistics domains need to keep pace with this evolution and create resources whose utility can be proved both for linguistic purposes and for computational applications.

The Natural Language Processing group of the "Mihai Drăgănescu" Research Institute for Artificial Intelligence of the Romanian Academy has been involved, since its foundation, both into the creation of resources and into the development of applications. The latter make use of the former and the results give feedback, among other things, about the quality of the resources. It is a process that underlies the necessity for well qualified linguists able to get involved in the creation and maintenance of these resources. Probably the best that we have, the Romanian wordnet (RoWN) will be presented below, with focus on its latest developments.

A different type of lexicogaphic resource

Lexicography is the domain of creating dictionaries. Nowadays, we can identify three profiles of the dictionary users:

- the traditional type – any user who prefers looking up printed dictionaries;
- the modern type – usually, a user of computers, of Internet, more or less dependent on these means of communication; they prefer looking up dictionaries on smart phones, on various gadgets for reading electronic books, etc.;
- the technical type – represented by specialists in natural language processing, who develop applications that require lexical knowledge.

Each of them needs a different lexicographic product. The first one will use a "traditional" dictionary, i.e. a printed one. The second – a dictionary in electronic format, containing the same type of information as the printed one. One advantage is the fact that there is no restriction on the number of entries. An immediate consequence would be the lack of rigorous criteria for selecting the material to be included. For technicians we cannot talk about dictionaries, but about language resources with a content that can be processed by computers with the aim of improving the results of various applications they develop.

One type of user that seems neglected above is the linguist specialist. Depending on their work method, we consider that linguists can be distributed in any of the three categories.

The lexical knowledge repository that we present below is dedicated primarily to language engineers, although others will also find it extremely useful and easy to look up.

The last decades experience has proved that using semantic networks in the form of wordnets for representing lexical knowledge is appropriate for computer use. For more than twenty years we can talk about a wordnet-style lexicography, modeled after Princeton WordNet (PWN) (Miller 1995; Fellbaum 1998). This has been realized manually by a team of lexicographers. Its last version contains 155 287 words (http://wordnet.princeton.edu/wordnet/man/wnstats.7WN.html). The figure is impressive, indeed, but one needs to mention that PWN contains simple literals, but also numerous relatively stable combinations of words. What is more, the lack of space restrictions and of clear criteria for selecting the words to be included in the network lead to the representation of many proper nouns and of plenty of terms from various domains. Nevertheless, their presence is justifiable through the needs of various applications: named entity recognition, terms recognition, term frequency calculus for establishing the degree of relevance of a document for a certain domain with the aim of information extraction.

In a wordnet, there are nodes and arcs. The former contain sets of words with associated sense numbers; the latter are semantic relations holding between nodes. Two or more words belong to the same node (or synset) if they have one common meaning and if they occur in more or less the same contexts. To make the meaning clearer, each synset has a gloss. Only nouns, verbs, adjectives and adverbs are in-

cluded in the wordnet. The semantic relations in the wordnet vary according to the part of speech of the words in the linked synsets. Thus, hyponymy and meronymy create the nouns hierarchies, troponymy and lexical entailment organize the verbs, while descriptive adjectives are grouped in clusters, around two antonymic adjectives as heads of opposing groups; the rest of the adjectives and the adverbs have no organization.

The Romanian Wordnet

The development of the Romanian Wordnet (RoWN) started in 2001 within the BalkaNet project (Tufiş 2004), within which wordnets were developed for Bulgarian, Czech, Greek, Romanian, Slovene, and Turkish.

A team of researchers from the Research Institute for Artificial Intelligence of the Romanian Academy and from the Faculty of Informatics of the "Al. I. Cuza" University of Iaşi and also students from two master programmes (one from the Faculty of Informatics of "Al. I. Cuza" University of Iaşi and one from the Faculty of Letters of the University of Bucharest) developed almost 18 000 synsets, conceptually aligned to Princeton WordNet 2.0 and through it to the synsets of all the BalkaNet wordnets. After the end of this project, the Research Institute for Artificial Intelligence of the Romanian Academy undertook the task of maintaining and further developing the RoWN.

There are multiple ways of creating a wordnet. The most accurate is the manual one. Used for developing the PWN, it stands up most criticisms. However, the high costs involved prevent other teams to undertake a similar enterprise. A rather cheap approach is to automatically extract the synsets and the relations between them from various resources available: such experiments are presented in (Agirre et al. 2002) for Basque, in (Barbu, Barbu Mititelu 2007) for Romanian, in (Fišer, Sagot 2008) for Slovene and French, in (Isahara et al. 2008) for Japanese. Translation of the PWN synsets and transfer of its structure into the newly created wordnet is a fast way of creating a wordnet: the Finnish (Linden and Carlson 2010) and the Thai (Leenoi et al. 2009) wordnets have been created like this. Many projects used a combined top-down method: a core wordnet was first developed (usually by translating the English synsets) and then it was enriched in various ways. EuroWordNet (Vossen 1998) and BalkaNet (Tufiş 2004) projects followed this top-down method. All these approaches assume a close conceptual similarity among languages, due to which the PWN structure is transferable to other wordnets (this is also the assumption behind MultiWordNet, Pianta et al. 2002). Further manual revision is mentioned by most of the authors. Unlike the expand model used in all the above cases, in the merge approach a wordnet is developed for a certain language and then aligned to the PWN; this is the case of the Russian WordNet (Balkova et al. 2004).

The Romanian team undertook a methodology of development from scratch, combining the expand and merge models. Each English synset is considered a

part of the lexical network, it is viewed in the system of relations which it enters, it is contrasted with its hypernyms, hyponyms, co-hyponyms, troponyms, etc., so that the lexicographer can understand its exact meaning which needs to be expressed in Romanian. For each English synset, a list of possible Romanian translations is suggested to the lexicographer from an electronic English-Romanian dictionary (of 74 000 translation pairs). For each such translation, some sets of synonyms are proposed from an electronic dictionary of synonyms (containing around 26 000 sets of synonyms). The lexicographer can choose the correct one, can adapt it if necessary, by deleting or adding literals from/to it, can write a different synset in case none of the suggested ones is correct. Each literal is assigned a sense number from the electronic explanatory dictionary (containing around 70 000 entries).

During the BalkaNet project, the aim was the coverage of the set of Common Base Concepts established in the EuroWordNet project (Vossen 1998; 1 024 concepts) and of three sets of BalkaNet Base Concepts (8 516 concepts altogether). They were established starting from the lists of the most frequent words in the languages considered in the project.

Afterwards, the applications in which RoWN was used influenced the choice of synsets to be implemented. Thus, we aimed at covering lexically the following:

- the electronic version of the novel *1984* by George Orwell;
- the NAACL2003 corpus of newspaper articles;
- the JRC-AcquisCommunautaire corpus;
- the Eurovoc thesaurus;
- articles from Wikipedia in Romanian;
- Verbnet3.0.

Adding derivational relations to the Romanian Wordnet

There are three levels at which the importance of marking derivational relations is apparent:

- at the monolingual level – the density of relations in a wordnet increases between words with the same part of speech, but especially between words of different parts of speech. For example, the lexical family made up of *pădure* "forest", *pădurar* "forester", *pădurice* "grove", *păduros* "wooded", *împăduri* "afforest", *împădurire* "afforestation", *despăduri* "deforest", *despădurire* "deforestation", *reîmpăduri* "reafforest", *reîmpădurire* "reafforestation", there are four derivational links between words of the same part of speech (i.e. *pădure – pădurar*, *pădure – pădurice*, *împăduri – despăduri*, *împăduri – reîmpăduri*), and five derivational links between words of different parts of speech (noun – verb: *pădure – împăduri*, *împădurire – împăduri*, *reîmpădurire – reîmpăduri*, *despădurire – despăduri*; noun – adjective: *pădure – păduros*). From a theoretical linguistics perspective, we can conduct studies concerning the semantic aspects of affixation in Romanian.

- at the multilingual level – the semantic labels associated with the derivational relations are established at the synset level, so they hold among concepts and could be transferred from one wordnet into another, provided that they are aligned with each other. The more wordnets with such relations, the more numerous and interesting comparative studies can be made: one can analyze how a certain semantic relation is morphologically realized in various languages: if it has a morphologic counterpart or not, what affixes express it, etc.
- at the applications level – a wordnet enriched with morpho-semantic relations turns into a knowledge base useful for various tasks such as question answering, information retrieval, and others.

Identifying derivationally related literals in Romanian[1]

Given the literals in the synsets of the RoWN, our aim is to find pairs of words made up of a derived word and its base. In order to render the steps of the derivational process, we are interested in finding for each derived word its stem, not its root (when the stem is different from the root). For instance, we want to mark *reîmpădurire* derived from *reîmpăduri* (by means of the suffix *-re*), derived, in its turn, from *împăduri* (by means of the prefix *re-*), which, in its turn, is derived from *pădure* by the prefix *îm-* (variant of *în-*) and the suffix *-i*. So, there are direct derivational links between stems and the words derived from them, but there are also indirect derivational links, like the one between *reîmpădurire* and *pădure*, which is reconstructed from the direct derivational links. Choosing this working method is most appropriate for the derivational process that takes place in steps: affixes are usually attached one after the other. We chose not to use directed links, because we aim at a similar treatment of both proper derivation and back formation.

We extracted from RoWN all simple literals, irrespective of their part of speech. We did not deal with proper nouns. Given this list of literals and the list of prefixes and that of suffixes, we made combinations of one literal and either a prefix or a suffix. When the resulted form could be found in the list of literals extracted from the RoWN, we retained the pair initial literal – obtained literal as a candidate pair of a base – derived words. Adding prefixes, we obtained 2 862 such pairs. Adding suffixes, we obtained 13 556 pairs. The explanation comes from the fact that Romanian has a larger number of suffixes than of prefixes; suffixation is a highly productive linguistic phenomenon, unlike prefixation.

A further step was to validate these candidates. For that, we tried some automatic heuristics. For prefixes, we used a morphologic validation method: the base and the derived word must have the same part of speech. The assumption is that prefixation does not change the part of speech of the stem it attaches to. Out of the 2 862 pairs only 2 621 met this constraint. Analyzing the eliminated pairs, we noticed that we could validate 83 of them. There are three types of examples among them: (a) 71 cases are due to RoWN incompleteness: e.g. *mulțumit* “pleased” – *nemulțumit* “displeased”: these words can be adjectives, adverbs, or nouns in Ro-

manian; however, in the RoWN the former is implemented only as an adjective, while the latter as an adverb and a noun; the other values will be implemented in RoWN in the future; (b) 1 case when prefixation does change the part of speech of the base word: e.g.: *cancer* "cancer" (noun) – *anticancer* "anticancer" (adjective); this is an exception to the rule making our assumption; (c) 11 cases when the literals have a wrong part of speech tag in RoWN and require correction. Through manual inspection of the 2 862 pairs of words with the same part of speech, we validated a set of 1 907 base-prefixed word pairs, so almost 67%. This means that in around 33% of all cases the beginning of words is a false prefix: e.g.: *curs* "course" – *excurs* "excursus": although both words are nouns and the latter has the first two letters *ex-*, which is a prefix attachable to nouns to create another noun, the semantic condition is not met: the two words have no overlap of meaning. There are also some cases when compounding is mistaken for prefixation: *casă* "house" – *acasă* "at home" (< preposition *a* + noun *casă*).

For automatically validating the suffixed words, we exploited the morphologic information about them. Suffixes combine with words of certain parts of speech to create words with certain parts of speech. For example, the suffix *-eală* attaches to verbs to create nouns as in: *plictisi* "get bored" + suffix *–eală* > *plictiseală* "boredom".

For the suffixes occurring in our list of 13 556 pairs, relying on the literature dedicated to them (mostly Graur & Avram 1970, 1978, 1989; SMFCLR 1959, 1967, 1969), we established the parts of speech with which they combine and the part of speech of the resulting words. Exploiting this information, we numerically reduced the list to 9 123 pairs. In order to verify how correct these are, we manually validated them. We found that 8 452 pairs were correct.

Such morphological relations are valid only within a language and they are established at the word level, more exactly at the word sense level. Consequently, derivational relations need to be marked among literals, not at the synset level. According to wordnet terminology, these are lexical, not semantic relations. They have the following properties:

- symmetry: if word w_1 is in derivational relation with word w_2, then w_2 is also in derivational relation with w_1;
- transitivity: if word w_1 is in derivational relation with word w_2 and w_2 is in derivational relation with word w_3, then w_1 is also in derivational relation with w_3;
- non-reflexivity: word w_1 is not in derivational relation with itself, which means that we do not treat conversion as a type of derivation ("zero-derivation" as it is called in various books), as they do in PWN.

Usually, affixes have meanings which can be rendered in terms of semantic labels. They can be represented at the synset level.

From the pairs of stem-derived words that we identified we extracted those whose members occur in only one synset each and searched for the semantic relations marked in RoWN for those synsets. We found antonymy, hypo- and hyper-

nymy, meronymy and holonymy, pertainymy. We consider that in such cases the semantic relations are morphologically motivated and there is no need for further semantic labeling of the links.

For those that are polysemantic we performed a manual annotation, using the following semantic labels[2]: *colour*, *make-become*, *dim*, *agent*, *manner*, *similitude*, *event*, *result*, *gender*, *location*, *of-origin*, *job*, *object-made-by*, *abstract*, *action*, *place*, *state*, *period*, *undergoer*, *tax*, *instrument*, *by-means-of*, *part*, *make-acquire*, *disease*, *sound*, *cause*, *container*, *aug, clothes*, *wife*, *fruit*, *animal*.

Semantic labels associated with derivational relations are valid cross-lingually, even if the morphological relation is not present in all languages. For instance, in Romanian there is a morphological relation between *bucătar* "cook" and *bucătărie* "kitchen": the latter is derived from the former with the help of the suffix *-ie* (and the vowel mutation *a:ă* which is common when *a* loses stress). The semantic relations or labels involved in this case are those of job and place. However, the same semantic relations exist for the English *cook* and *kitchen*, although they are morphologically unrelated. When wordnets are aligned, the semantic labels existing in one language can be transferred into the other language(s) or can be checked cross-lingually. Whenever discrepancies occur, they signal a mistake in annotation.

Transfer of derivational relations from PWN

Within METANET4U project we experimented with the transfer of the semantic labels from the standoff file of PWN. We went through 4 384 pairs of synsets and for 1 475 of them we found that they have equivalent morphologically related translations in Romanian. For instance, in English the verb *weed* in its second meaning "clear of weeds" is derivationally related with the first meaning of the noun *weeder* "a farmhand hired to remove weeds". Their equivalents in Romanian, *plivi* and *plivitor*, respectively, are also derivationally related and in both languages the noun expresses an agent. The other pairs (2 909) either are not implemented in RoWN or the literals implementing them are not derivationally related. Consider the English pair: the verb *dry* in its second sense "become dry or drier" and the noun *drier* in its first meaning "a substance that promotes drying". Their respective equivalents in Romanian are *usca* and *sicativ*, which are morphologically unrelated.

Within a bilateral project of the Institute for the Bulgarian Language of the Bulgarian Academy of Sciences and our Institute, we are interested in digging for derivational relations valid among literals in correspondent synsets in our wordnets.

Conclusions and further work

The RoWN is a mature resource. Here are some quantitative data about it:

	Nouns	Verbs	Adjectives	Adverbs	TOTAL
literals	38912	6889	4469	2799	53069
synsets	41063	10397	4822	3066	59348
senses	56532	16484	8203	4019	85238

Fig. 1. Statistical data about RoWN.

Our main concern now is to make it more appropriate for further use in various applications. Thus, we are primarily preoccupied with enriching it with morphologic, semantic and syntactic information. As far as the first two aspects are concerned, adding these derivational relations and some associated semantic links contribute a lot to the density of relations in our resource, especially cross-part of speech links.

We also take care that our synsets and glosses are accurate and every now and then we check their correctness and perform any necessary improvement.

The latest information about RoWN can be found at nlptools.racai.ro. RoWN can be browsed by a web interface implemented on our language web services platform (http://nlp.racai.ro/WnBrowser/).

NOTES

[1] The work described in this section was supported by the Sectorial Operational Program Human Resources Development (SOP HRD), financed from the European Social Fund and by the Romanian Government under the contract number SOP HRD/89/1.5/S/59758.

[2] The teams that added such labels to their wordnets worked with a different set of labels. In Princeton WordNet they are suggestive for the semantic type of the relationship between verbs and nouns. 14 labels are used: agent, material, instrument, location, by-means-of, undergoer, property, result, state, uses, destination, event, body-part, vehicle. In the Czech wordnet, they reflect the parts of speech involved in the relation rather than the semantic type of these relations: deriv-na, deriv-ger, deriv-dvrb, deriv-pos, deriv-pas, deriv-aad, deriv-an, deriv-g, deriv-ag, deriv-dem. For Turkish they were chosen so that they have a higher degree of generality: become, acquire, be-in-state, someone-with, something-with, someone-from, someone-without, something-without, pertains-to, with, reciprocal, causes, is-caused-by, cat-of, manner.

REFERENCES

Agirre et al. 2002: Agirre, E., O. Ansa, X. Arregi, J. M. Arriola, A. D. DeIlarraza, E. Pociello, L. Uria. Methodological Issues in the building of the Basque WordNet: quantitative and qualitative analysis. // *Proceedings of the first International Conference of Global WordNet Association.*

Balkova et al. 2004: Balkova, V., A. S. Suhonogov, A. Yablonsky. Russia WordNet. From UML-notation to Internet / Intranet Database Implementation. // *Proceedings of the Second International WordNet Conference*, pp. 31–38.

Barbu, Barbu Mititelu 2007: Barbu, E., V. Barbu Mititelu. Automatic Building of Wordnets. // N. Nicolov, K. Bontcheva, G. Angelova and R. Mitkov (Eds.) *Recent Advances in Natural Language Processing IV.* Amsterdam: John Benjamins, pp. 217–226.

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

Fišer, Sagot 2008: Fišer, D., B. Sagot. Combining multiple resources to build reliable wordnets. // *Text, Speech and Dialogue Conference*. Berlin: Heidelberg, Springer, pp. 61–68.

Grau, Avram 1970, 1978, 1989: Graur, A., M. Avram. (Eds.) *Formareacuvintelorînlimb aromână*. Vol. I, 1970. Vol. II, 1978. Vol. III, 1989. Romanian Academy Publishing House.

Isahara et al. 2008: Isahara, H., F. Bond, K. Uchimoto, M. Utiyama, K. Kanzaki. Development of the Japanese WordNet. // *Proceedings of LREC'2008,* pp. 2420–2423.

Leenoi et al. 2009: Leenoi, D., T. Supnithi, W. Aroonmanakun. Building Thai WordNet with a Bi-directional Translation Method. // *Asian Language Processing* (IALP 2009), pp. 48–52.

Lindén, Carlson 2010: Lindén, K., L. Carlson. FinnWordNet – WordNetpåfinska via översättning. // *LexicoNordica – Nordic Journal of Lexicography*, No. 17, pp. 119–140.

Miller 1995: Miller, G. A. WordNet: A Lexical Database for English. // *Communications of the ACM*, Vol. 38, No. 11, pp. 39–41.

Pianta et al. 2002: Pianta, E., Bentivogli, L., Girardi, C. MultiWordNet: developing an aligned multilingual database. // *Proceedings of the First International Conference on Global WordNet*, pp. 293–302.

SMFCLR 1959, 1967, 1969: *Studiişimaterialereferitoare la formareacuvintelorînlimbaro mână.* Romanian Academy Publishing House.

Tufiş 2004: Tufiş, D. (Ed.) *Romanian Journal on Information Science and Technology. Special Issue on BalkaNet.* Romanian Academy, **7**, No. 2–3.

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