Wordnets: State of the Art and Perspectives.

Case study: the Romanian Wordnet

Verginica Barbu Mititelu
Romanian Academy
vergi@racai.ro

Abstract

During a quarter of a century of existence and in spite of much criticism, wordnets have thoroughly proved their appropriateness as repositories of linguistic knowledge and their usefulness in various applications. In this paper we present the methodology of creating the Romanian wordnet (RoWN), with special emphasis on the strategies adopted during ten years of ceaseless implementation and which highlight the efforts invested, the way we dealt with the alignment of the RoWN (previously aligned to PWN 2.0) to the PWN 3.0, as well as the future work we envisage for enriching and extending this resource.

1 Generalities on wordnets

Language is a system of signs. This structuralist perspective on language serves extremely well the description of natural languages by both theoretical linguists and specialists in the formal representation of language. Notions like *paradigm* (i.e. class of similar elements), *syntagm* (i.e. a linguistic environment in which the elements of a paradigm can occur), *value* (the distinguishable functional role of an element in a syntagm) are modeled to serve the formal representation of language as a whole.

Among the different knowledge representation formalisms, we focus here on wordnets, a special kind of semantic networks. While semantic networks have words in nodes and the arcs are semantic relations, wordnets are definitely more than that: they are:

- monolingual dictionaries: they contain words with definitions for each of their senses;
- multilingual dictionaries: via the Inter-Lingual Index, ILI, access from one language-specific network to all the others is facilitated; thus, it is possible to compare the organization of the lexical material of various languages, to find examples supporting the thesis of semantic specificity of languages, to introduce the multilingual dimension in various applications relying on wordnets;
- thesauri: lexical information is organized in terms of word meanings, not word forms;
- lexical ontologies: wordnets contain concepts lexicalizations from various domains and the relations between these concepts lexicalizations. There have been more projects enriching wordnet with ontological information: WordNet Domains (Pianta et al. 2002), SUMO (Niles and Pease 2003).

2 Development

There are more ways of creating a wordnet. The most accurate is the manual one. Used for developing the Princeton WordNet, it stands up most criticisms. However, the high costs involved in terms of money and time 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 Aggire et al (2002) for Basque, Barbu and Barbu Mititelu (2005) for Romanian, Fišer and Sagot (2008) for Slovene and French, 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 made like this. Many projects used a com-

Proceedings of Recent Advances in Natural Language Processing, pages 672–677,
Hissar, Bulgaria, 12-14 September 2011.

bined top-down method: a core wordnet was first developed (usually by translation of the English synsets) and then it was enriched in various ways: the EuroWordNet (Vossen et al. 1999), the BalkaNet (Tufiş 2004) projects. All these approaches assume a close conceptual similarity among languages, due to which the PWN structure is tranferable 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 as 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 74000 translation pairs). For each such translation, some sets of synonyms are proposed from an electronic synonyms dictionary (containing around 26000 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, if none of the proposed ones is correct. Each literal is assigned a sense number from the electronic explanatory dictionary (containing around 70000 entries).

More than 400 synsets were added to the RoWN during the BalkaNet project (as well as to the other wordnets created meanwhile), synsets that are considered specific to the culture and civilization of our geographical region. They are not dangling nodes, but were assigned the appropriate relations in the network.

The structure of the RoWN is imported from the PWN. Most of the relations it contains are conceptual, so they are transferable from one language to another. Thus the hierarchy of the PWN is preserved. Nevertheless, this does not contradict the thesis of semantic specificity of languages, since we mark the concepts that lack a Romanian lexicalization with the notation NL (for non-lexicalized). The differences on the horizontal and on the vertical axes are easily found in the wordnets aligned to the ILI.

During our implementation, we noticed that antonymy is transferrable into our network.

A rather sensitive topic is represented by derivational relations. Let e_1 and e_2 be two English literals, and, for instance, r_1 and r_2 their Romanian equivalents; if e_1 is derived from e_2 with a certain affix, it may be the case, but it is not obligatory so, that r_1 is derived from r_2 with an affix. Thus, in English there are *drive – driver* and in Romanian *şofa* „drive" – *şofer* „driver"; in English there are *teach – teacher* but in Romanian there are *preda* „teach" – *profesor* „teacher"; in Romanian there are *bucătar – bucătărie*, while in English there are *cook – kitchen* respectively. Such examples can be found for any language pairs. Marking derivational relations is of great help: a base word and all the words derived from it belong to the same semantic field. Relating them can (at least partially) solve the famous "tennis problem" of wordnets (Fellbaum 1998: 10). Thus, derivation proved to be the third relation as importance for obtaining good quality lexical chains, after hypernymy and hyponymy (Novischi and Moldovan 2006). Lexical chains are then used in various tasks: improvement of QA systems (Novischi and Moldovan 2006), summarization (Barzilay and Elhadad 1997), document indexing (Stairmand 1996), detection of malapropism (Hirst and St-Onge 1997) and others.

In PWN some of the derivation relations are already marked (Fellbaum et al. 2009). Due to the lexical nature of these relations (i.e. they establish between two words, not between synsets), they cannot be automatically transferred into other wordnets. However, some other wordnets have derivation relations marked: the Czech one (Pala and Smrz 2004), the Bulgarian (Koeva 2008), the Russian (Azarova 2008) ones.

2.1 Sense numbering

In PWN polysemous words have sense numbers attributed in an artificial manner: the word senses are distributed in a decreasing order of their number of occurrences in tagged corpora.

Specific to the RoWN among all the existent wordnets is the way sense numbers are assigned to literals. Whenever a word is present in the EXPD, its sense number is preserved in the RoWN synset. However, in EXPD the organization of word meanings is hierarchical, highlighting their relatedness: many of them are derived from other meanings. Here are the meanings of the Romanian word *spart* "broken" in the EXPD:

1.1 Spargere. (En. "breaking into");

1.2 Sfârşit, încheiere a unei activități (En. "end of an activity");

1.3 Expresie: *A ajunge la spartul târgului (sau iarmarocului)*; a ajunge undeva prea târziu, când lucrurile sunt lichidate. (En. Expression *a ajunge la spartul târgului* "to arrive too late");

2.1.1 Prefăcut în bucăți, în cioburi (En. "turned into pieces");

2.1.2 plesnit, crăpat (En. "cracked");

2.1.3 găurit (En. "drilled");

2.2.1. Expresie: *a fi mână spartă*; a fi risipitor (En. Expression *a fi mână spartă* "to be easy money");

2.2.2. Expresie: *A mânca de parc-ar fi spart*; se spune cuiva sau despre cineva care mănâncă foarte mult şi cu lăcomie (En. Expression *A mânca de parc-ar fi spart* "to eat very much and with greed");

2.3. (Despre lemne) Tăiat în bucăți mici (potrivite pentru a fi arse în sobă) (En. (About woodsticks) to be chopped in small pieces (appropriate for burning in a stove));

2.4. (Despre pământ) Răscolit, plin de gropi (En. (about ground) embowelled);

2.5. (Rar, despre butoaie) Desfundat (En. (rare, about barrels) bilged);

2.6. (figurativ (Despre sunete)) Lipsit de sonoritate, răguşit, dogit (En. (fig. (about sounds)) lacking sonority, hoarse, jangle);

2.7. (Despre ziduri, clădiri) Stricat, dărăpănat, ruinat (En. (about walls, buildings) broken down, decaying);

2.8. (Despre obiecte de încălțăminte, de îmbrăcăminte) Rupt, uzat, tocit (En. (about footwear and clothes) worn out).

On the first hand, there is a clear distinction between homonyms (senses under 1 and senses under 2). On the other hand, senses under 1 are clearly distinguished one from the other, express activities. Senses under 2 express results and are grouped together as follows: those under 2.1 refer to objects, those under 2.2 are senses in expressions, while those under 2.3 to 2.8 refer to various entities that can undergo a disruption, a fracture of their wholeness; these are specific senses.

We decided to maintain these imbricated sense numbers for literals because they can be viewed as an extra "relation" in wordnet, which keeps track of related meanings (and can help in clustering experiments). Linguists can also extract from the semantic network statistics of various kinds of semantic evolutions of word meanings.

A special case is represented by words that have meanings unattested in EXPD. The lexicographer carefully examines the attested ones in order to find the closest one; if it exists, the unattested meaning gets the same sense number as this one with ".x" added at its end. Thus, the hierarchical organization of meanings remains unaltered. If it does not exist, i.e. the meaning under consideration is not close to any of the recorded meanings in EXPD, then the "x" sense "number" is assigned to it, so it is treated as a distinct meaning.

Sometimes, although extremely carefully examining two synsets, lexicographers realize that they simply cannot find any distinction between them. The solution in such a case is to appeal to a native speaker's knowledge of his/her mother tongue. If (s)he also cannot find any motivation for the existence of these two synsets, then we adopted a notation to manage these cases: we add ".c" after the sense number. A suggestive example in this case is the pair of synsets: {eclipse:3} (gloss: "cause an eclipse of; of celestial bodies") and {eclipse:2, occult:1} (gloss: "cause an eclipse of (a celestial body) by intervention"). A further proof of this impossibility to differentiate semantically between the two PWN 2.0 synsets is the fact that in PWN3.0 the two different synsets have been merged into one: the latter. (In other words, the former has been eliminated from the wordnet.) The Romanian synsets corresponding to these two were identical: {eclipsa:1.c}. However, after aligning the RoWN to PWN 3.0, one of these identical synsets disappeared. Thus, the notation ".c" becomes void of significance. It can be automatically removed alongside with other such cases that are easily identified in the wordnet: if there is only one occurrence of a literal with a certain sense number ending in ".c", then we can safely remove this ending without losing any information.

Another case in which this notation proves useful is represented by pairs of synsets such as: {mister:1, Mr:1} (gloss: "a form of address for a man") and {sir:1} (gloss: "term of address for a man"). According to Cambridge Dictionary, the former is a title, although it is also "an informal and often rude form of address for a man whose name you do not know" (http://dictionary.cambridge.org/dictionary/british/mister), while the latter is "used as a formal and polite way of speaking to a man, especially one who you are providing a service to or who is in a position of authority" (http://dictionary.cambridge.org/dictionary/british/sir_1). Their Romanian equivalent is *domn:1.1* ("polite form of address for a man"). It is also used as a title. However, since such a title is used to address a man, there is no semantic reason to postulate the existence of another meaning for *domn*. That is why we implemented these two synsets with two synsets of the form {domn:1.1.c}.

So the sense numbers that literals can have in RoWN have any of the forms covered by the regular expression: \d+(\.\d+)*(\.[xc])?|x

2.2 Tools

Two tools were designed to help the lexicographers develop the synsets of the RoWN: WNBuilder and WNCorrect (Tufiş and Barbu 2004). The former is a configurable graphical interface, language independent (but resources dependent) that assists the lexicographer in the synsets development, imports the relations for the created Romanian synsets from the PWN xml file, performs validation of the created synsets: the lexicographer receives a message about the existent problem(s) and suggestions for solving it/them and generates the xml version of the file.

WNCorrect is designed for the semantic validation of the RoWN. After identifying the synsets with conflicting literals (i.e. synsets in which a literal occurs with the same sense number), their list is given to a lexicographer. Using the WNCorrect, (s)he can visualize the synsets in which each literal occurs and can perform the necessary corrections.

2.3 Valence frames

The syntactic frames in which a verb can occur are registered in PWN in a highly general way, using indefinite pronouns like *somebody*, *something* and indefinite adverbs like *somewhere*. More frames are given for the same synset, in an uneconomical way: for optional arguments a new frame is recorded. For instance, for the verb *inherit* with sense number 1 (gloss: "obtain from someone after their death"), there are two frames: "Somebody ----s something" and "Somebody ----s something from somebody".

RoWN also contains valence frames for some verbs. They are defined at the literal level. That is why, for one synset more than one valence frames can be found. They are the result of an experiment relying on parallel corpora, word alignment and word sense disambiguation technologies through which we imported syntactic-semantic information from one part of the bitext, richly annotated for the respective language, into the other part where the linguistic annotation is scarce or missing.

The resources used in this experiment were: the *1984* corpus (available in Czech and Romanian), the Czech wordnet and the RoWN. The Czech wordnet contains valence frames for many verbs (Pala and Smrž 2004). Via the interlingual equivalence relations among the Czech verbal synsets and Romanian synsets we imported about 600 valence frames. They were manually checked against the BalkaNet test-bed parallel corpus (*1984*) and more than 500 subcategorisation frames were valid as they were imported or minor modifications were operated.

Very similar to the frames used in the FrameNet project (www.icsi.berkeley.edu/~framenet), the valence frames are attached to verbs only in our wordnet (although other words that can be logical predicates can also have such frames) and specify syntactic and semantic restrictions for the arguments of the predicate denoting the meaning of a given synset. They also specify the case roles of the arguments. The nice property of the Czech valence frames is that the semantic restrictions are endogenous, i.e. they are specified in terms of other synsets of the same wordnet. Let us consider, for instance, the verbal synset ENG20-02609765-v (se_afla:3.1, se_găsi:9.1, fi:3.1) with the gloss "be located or situated somewhere; occupy a certain position". Its valence frame is described by the following expression: (nom*AG(fiinţă:1.1)|nom*PAT(obiect_fizic:1)) = prep-acc*LOC(loc:1), where Ro fiinţă:1.1 means En being:2, Ro obiect_fizic:1 means En physical_object:1, and Ro loc:1 means En location:1.

The specified meaning of this synset is: an action the logical subject of which is either a *fiinţă* (sense 1.1) with the AGENT role(AG), or a *obiect_fizic* (sense 1) with the PATIENT role (PAT). The logical subject is realized as a noun/NP in the nominative case (nom). The second argument is a *loc* (sense 1) and it is realized by a prepositional phrase with the noun/NP in the accusative case (prep-acc).

A verbal synset can have two different frames, thus proving that the synonymy between words in the same synset is not very strictly defined in wordnet.

3 Aligning RoWN to PWN 3.0

Wordnets for various languages are useful in multilingual tasks if aligned to the same versionof PWN. We have recently aligned the RoWN to PWN version 3.0 via a mapping from PWN2.0 (to which the RoWN was aligned) to PWN 3.0. The main problems encountered in this process are of three types:

- there were 457 cases in which two or more Romanian synsets were aligned to one PWN 3.0 synset: in this case we had to decide which of the Romanian synsets is the best equivalent of the English one; necessary modifications in the synsets structure and in the gloss were operated;
- there were 56 cases when one Romanian synset aligned to two PWN 3.0 synsets: in their case we decided which of the two PWN 3.0 synsets is the correct equivalent of the Romanian synset and we also implemented an equivalent for the other PWN synset;
- 210 Romanian synsets disappeared through this mapping: their equivalent English synsets were eliminated: this is the case of many participial adjectives, of obsolete, euphemistic and slang meanings; some meanings were merged due to their identity; some compound literals were morphologically reanalyzed in simple words that were already in the network (e.g. *well endowed*), etc.

At present, the Romanian wordnet aligned to the PWN 3.0 contains 51986 literals in 57895 synsets. Its version aligned to the PWN 2.0 contained 52357 literals in 58725 synsets. Around 400 literals and 900 synsets were lost in the mapping process.

4 Conclusions and further work

In spite of the criticism against various aspects of the wordnet (treatment of various relations, sense granularity), there is a worldwide proliferation of the projects in which such a resource is created by various methods, either automatic or manual. To catch up with the PWN, many teams appeal to fast and cheap strategies, such as the automatic translation of the PWN and the import of its structure, sometimes leaving the glosses not translated, thus making it impossible to talk about that wordnet as a monolingual dictionary. However, the richness of relations is aimed by many developers as they can facilitate the extraction of valuable information for various applications. Such efforts are a proof that lexical resources in the form of wordnets are a must for natural languages in the electronic era, although there are still unsettled matters about wordnets. Further proof of their utility can be found in the applications relying on wordnets: summarization, question answering, word sense disambiguation, machine translation, information extraction and so on.

The ongoing development of the RoWN in the last decade followed three directions of research: implementation of new concepts and associated relations in the RoWN, with the aim of attaining a huge coverage of the Romanian lexicon, extensions to the RoWN and its using in applications (Word Sense Disambiguation see Ion and Tufiş 2004, Question Answering see Barbu Mititelu et al. 2009). The extensions to RoWN are the description of literals in terms of paradigmatic morphology (thus offering the great facility of searching for a word by its inflected forms, which is of extreme help in various applications using RoWN, especially as Romanian has a rich inflectional system, see Irimia 2007 for details) and the subjective mark-up of synsets (with the aim of mining opinions in text, see Tufiş 2009).

As other teams of researchers have already started to do, we also envisage the marking of derivational relations between words in RoWN, as well as enrichment of these relations with semantic information about the semantic type of the derived nominal, which could be of great help in various applications in which our wordnet will be used.

Acknowledgments

Part of the work reported here is suported by the Sectorial Operational Programme 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.

References

Eneko Agirre, Olatz Ansa, Xabier Arregi, José Mari Arriola, Arantza Diaz de Ilarraza, Eli Pociello, Larraitz Uria. 2002. Methodological Issues in the building of the Basque WordNet: quantitative and qualitative analysis. *Proceedings of the first International Conference of Global WordNet Association.*

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

Eduard Barbu and Verginica Barbu Mititelu. 2005. Automatic Building of Wordnets. *Proceedings of the International Conference Recent Advances in Natural Language Processing*:99-106

Verginica Barbu Mititelu, Alexandru Ceauşu, Radu Ion, Elena Irimia, Dan Ştefănescu, Dan Tufiş. 2009. Resurse lingvistice pentru un sistem de întrebare-răspuns pentru limba română. *Revista Română de Interacţiune Om-Calculator* 2:1-17.

Regina Barzilay and Michael Elhadad. 1997. Using lexical chains for text summarization. *Proceedings of the Intelligent Scalable Text Summarization Workshop*.

Darja Fišer and Benoît Sagot. 2008. Combining multiple resources to build reliable wordnets. *TSD*.

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

Christiane Fellbaum. 2009. Putting Semantics into WordNet's "Morphosemantic" Links. In Z. Vetulani, H. Uszhoreit (Eds.), *Human Language Technology*, Springer:350-358.

Graeme Hirst and D. St-Onge. 1998. Lexical Chains as Representation of Context for Detection and Correction of Malapropisms. In Ch. Fellbaum (Ed.)

Radu Ion and Dan Tufiş. 2004. Multilingual Word Sense Disambiguation Using Aligned Wordnets. *Romanian Journal of Information Science and technology*, vol. 7, no. 1-2:183-200.

Elena Irimia. 2007. ROG - A Paradigmatic Morphological Generator for Romanian. In Z. Vetulani (Ed.), *Proceedings of the 3rd Language & Technology Conference: Human Language Technologies as a Challenge for Computer Science and Linguistics*:408-412.

Hitoshi Isahara, Francis Bond, Kiyotaka Uchimoto, Masao Utiyama, Kyoko Kanzaki. 2008. Development of the Japanese WordNet. *Proceedings of LREC'2008*.

Svetla Koeva. 2008. Derivational and Morpho-Semantic Relations in Bulgarian Wordnet. *Intelligent Information Systems*, XVI: 359-369, Academic Publishing House.

Krister Lindén and Lauri Carlson. 2010. Finn-WordNet - WordNet på finska via översättning. *LexicoNordica - Nordic Journal of Lexicography*, vol 17.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, Katherine Miller. 1993 Introduction to WordNet: An On-line Lexical Database. *Special Issue of the International Journal of Lexicography*, 3 (4), initial version 1990.

Karel Pala and Pavel Smrž. 2004. Building Czech WordNet. *Romanian Journal of Information Science and technology*, vol 7, numbers 1-2:79-88.

Ian Niles and Adam Pease. 2001 Towards a Standard Upper Ontology. *Proceedings of the 2nd International Conference on Formal Ontology in Information* Systems:2-9.

Adrian Novischi and Dan Moldovan. 2006. Question Answering with Lexical Chains Propagating Verb Argument. *ACL*.

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

M.A. Stairmand. 1996. *A Computational Analysis of Lexical Cohesion with applications in Information Retrieval*, Ph.D Thesis, UMIST.

Dan Tufiş. 2009. Paradigmatic Morphology and Subjectivity Mark-up in the RO-WordNet Lexical Ontology. In H.N. Teodorescu, J. Watada, L.C. Jains (Eds.), *Intelligent Systems and Technologies – Methods and Applications*, Springer:161-179.

Dan Tufiş and Eduard Barbu. 2004. A Methodology and Associated Tools for Building Interlingual Wordnets. *Proceedings of the 4th International Conference on Language Resources and Evaluation LREC 2004*:1067-1070.

Dan Tufiş, Dan Cristea, Sofia Stamou. 2004. BalkaNet: Aims, Methods, Results and Perspectives. *Special Issue of the Romanian Journal of Information Science and Technology*, vol. 7, no. 1-2:9-43.

Piek Vossen, Wim Peters, Julio Gonzalo. 1999 Towards a universal index of meaning. *Proceedings of the ACL-99 Siglex workshop*:81-90.