Romanian WordNet: New Developments and Applications

Dan Tufiş and **Verginica Barbu Mititelu** and **Luigi Bozianu** and **Cătălin Mihăilă**
Romanian Academy Research Institute for Artificial Intelligence
13, Calea 13 Septembrie, 050711, Bucharest 5, Romania
{tufis, vergi, bozi, cata}@racai.ro

Abstract

Among the existing ontologies, the multilingual lexical ontologies have a special status. Structured in a similar way to standard ontologies, the lexical ones are distinguished by the fundamental requirement that each conceptualized entity is lexicalized by one or more synonymous words (a synset) of the natural language vocabulary. Multilingually aligned wordnets, such as EuroWordNet or BalkaNet, represent one step further with great promises in the domain of multilingual processing. This paper gives an account for the development and current status of the Romanian wordnet, aligned to the Princeton WordNet 2.0 (PWN2.0), and discusses some of its applications.

1 Introduction

Semantic lexicons are one of the most valuable resources for a plethora of natural language applications. Incorporating WordNet or its monolingual followers in modern NLP-based systems already represents a general trend supported by numerous reports showing significant improvements in the overall performances of these systems. Multilingual wordnets, such as EuroWordNet (Vossen 1998) and BalkaNet (Tufiş et al. 2004a), which adopted the Princeton WordNet (Fellbaum 1998, Miller et al. 1990) as an interlingual linking device, represent one step further with great promises in the domain of multilingual processing.

The BalkaNet multilingual lexical ontology has been developed within the European project with the same name (September 2001-August 2004) and includes five languages from the Balkan area (Bulgarian, Greek, Romanian, Serbian and Turkish), plus Czech, whose wordnet, initially developed in EuroWordNet, has been significantly extended. As in EuroWordNet, English is the hub language.

By observing the interlingual synset mapping principle and incorporating most of the conceptual extensions proposed by EuroWordNet (e.g. upper-level ontological descriptions for the 1208 base concepts (Rodriguez et al. 1998)), the BalkaNet wordnets can be easily combined with any of the other semantic networks of the EuroWordNet, thus one may speak about a really pan-European multilingual lexical ontology, covering at least 15 languages[1].

The BalkaNet multilingual environment took advantage of the latest developments in the PWN, which was adopted itself as an interlingual index (ILI). This is a major difference with respect to the EuroWordNet's ILI. As the SUMO/MILO (Niles and Pease 2001) and DOMAINS (Bentivogli et al. 2000, Bentivogli et al. 2004, Magnini and Cavaglia 2000) have been aligned with PWN,they automatically became available in each monolingual wordnet of the BalkaNet. To allow the representation of language idiosyncratic properties, the structural knowledge present in the monolingual wordnets has precedence over the structural knowledge imported from the ILI. As the Romanian wordnet imported SUMO/MILO and DOMAINS labels and the synsets unique identifiers are the same as in the PWN, it is self-contained, but at the same time unambiguously integratable in a PWN-centered multilingual wordnet infrastructure.

The rest of the paper is organized as follows: in section 2 we present the methodology, briefly mention the tools we use, comment on the main assumptions we relied on, and discuss some issues related to the interlingual equivalence of the synsets; section 3 gives a statistical account of the current[2] Romanian wordnet; section 4 briefly mentions a few applications we developed using the Romanian wordnet. Section 5 describes future work and draws conclusions.

2 Methodology, Assumptions and Development Environment

During the BalkaNet project, the responsibility for creating the Romanian wordnet (RoWN) and for its aligning to ILI belonged to the Romanian Academy Research Institute for Artificial Intelligence from Bucharest and to the Faculty of Informatics of the University "Al. I. Cuza" from Iaşi (Cristea et al. 2004, Tufiş et al. 2004c). After the end of the project, only the former continued the development of the Romanian wordnet.

From the two distinct ways of creating a wordnet (the expend model and the merge model, cf. Tufiş et al. 2004a), the Romanian team preferred the latter. The choice was facilitated by the existence of some software instruments and linguistic resources previously built and adequate to the task.

[1] Basque, Bulgarian, Catalan, Czech, Dutch, English, Estonian, French, German, Greek, Italian, Romanian, Serbian, Spanish, and Turkish.

[2] At the time of this writing, October 10, 2005; as this work is continuing, the statistics provided in the current paper are temporary and just orientative.

Petr Sojka, Key-Sun Choi, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2006, Proceedings, pp. 337–344.

2.1 Language Resources for the RoWN

In building the Romanian wordnet we adopted a language centric approach (as opposed to a simpler method based on the translation of the literals in PWN), relying on two reference lexicographic resources, the Explanatory Dictionary of Romanian (EXPD) and The Dictionary of Synonyms (SYND), as well as on an in-house Romanian-English dictionary.

The EXPD is a general dictionary of modern Romanian (Coteanu et al. 1996) and contains about 56.000 entries[3]. We further extended this dictionary so that our current version contains almost 75,000 entries. The rich XML annotation of EXPD was automatically generated by our DIC dictionary compiler (Tufiş 1999) which takes as input a MRD and a grammar of the dictionary entry and generates XML code, conformant with a specified DTD[4]. The SYND (Seche and Seche 1999) was keyboarded, XML encoded and completed with more than 4000 new synonymy sets extracted from EXPD. We filtered out the archaic and regional variants with provision for automatic inclusion if ever needed.

The Romanian-English dictionary was automatically extracted from a large parallel corpus using word alignment technology. We used our own word aligner which continuously improved since its initial versions (see section 4).

Beside these language specific lexical resources we also used the XML format of the PWN. The structure of a synset is the same for all wordnets developed in the BalkaNet project (see Horák and Smrž 2004).

2.2 Selection Criteria for the RoWN Synsets

For the practical utility of the Romanian wordnet at each stage of its development, as well as to facilitate its incremental extension, during the project period we conducted a series of statistical investigations on our corpora. We extracted a frequency list containing only the words of interest for the wordnet structuring (nouns, verbs, adjectives and adverbs), getting a candidate list of more than 50,000 words. We checked the nouns and verbs in the first frequency group against the Frequency Dictionary of Romanian (FDR[5]) (Juilland 1965). The comparison we made revealed that all 5,000 words in FDR were also in our list, although not with the same frequency ranges.

As frequency in running texts is a disputable criterion, we considered two additional criteria that were easy to implement as selection procedures. The first criterion is the number of senses a headword has in our reference dictionary, while the second considered the number of entries contained by a given headword (fertility).

As mentioned earlier, the construction of the RoWN was driven by data existing in our reference lexical resources and in the next section we briefly review the specifics of this approach.

2.3 Idiosyncratic Methodological Aspects

For the proper building of Romanian synsets, closest to the meaning of the concepts in the set of selected ILI records, the lexicographers chose one of the synonymic series in the SYND. They also attached sense numbers according to the EXPD numbering and used only definitions from EXPD. However, under special conditions, and providing motivations, they were allowed to modify an initial synonymy set from SYND, to add a special sense number (non-existent in EXPD) or to change an EXPD definition. Such special conditions were: the synonymic set was too long and as such did not match the meaning of the target concepts; the sense number of a Romanian literal which would fit a target concept was not listed in EXPD (although the lexicographers considered it should have been); some sense definitions in EXPD were too coarse grained and had to be refined, etc.

Concerning the sense labeling based on PWN, one general criticism is that the senses of a given literal are described in a flat manner, although some senses are arguably semantically related. As we have this information represented in the EXPD by means of a sense labeling notation, we kept it in our wordnet with the same interpretation (see Tufiş et al. 2004b).

After implementing the targeted ILI concepts in Romanian we made a thorough investigation of the nature of the relations that link the synsets in PWN for seeing which of them can be safely transferred into the Romanian wordnet. As a result of this investigation, Tufiş and Cristea (2002) conjectured the *Hierarchy Preservation Principle* (HPP) which is the basic motivation for automating the import of most of the semantic relations (hypernym, holo_part, holo_portion, holo_member, subevent, causes, verb_group, be_in_state, similar_to, also_see, category_domain) from PWN into our wordnet. As one would expect, lexical relations (such as derivative, participle, region domain, usage domain, direct antonymy, etc.) are in general not valid cross-lingually, so they were not subject to automatic import. However, observing various language specific lexical relations (especially in agglutinative languages) one could derive in his/her own language useful syntagmatic relations (Bilgin et al. 2004).

By virtue of the HPP mentioned earlier, all the semantic relations were automatically imported as follows: if the two source synsets $S_{1SOURCE}$ and $S_{2SOURCE}$ are linked by a semantic relation R and if the $S_{1TARGET}$ and $S_{2TARGET}$ are the correspondingly aligned synsets in the target wordnet, then they will be linked by the relation R. If in the source wordnet there are intervening synsets between $S_{1TARGET}$ and $S_{2TARGET}$ then, we will set the relation R between the corresponding target synsets only if R is declared as transitive (R^+, unlimited number of compositions, e.g. hypernym) or partially transitive relation (R^k with k a user-specified maximum number of compositions, larger than the number of intervening synsets between $S_{1TARGET}$ and $S_{2TARGET}$). For instance, we defined all the holonymy relations as partially transitive ($k = 3$).

[3] This is the number of entries in the 1996 edition of the dictionary. The last version (2002) has almost 100,000 entries.

[4] The DTD we used was developed within the CONCEDE project: `http://www.itri.brighton.ac.uk/projects/concede/`

[5] The FDR was constructed based on a balanced corpus of 500,000 words of Romanian literature, legal texts, poetry and journalism and contains the list of most frequent 5,000 Romanian lemmas (in that corpus).

2.3.1 WNBuilder and WNCorrect

The WNBuilder is a configurable graphical interface (Tufiş and Barbu 2004) by means of which a lexicographer has access to all the language resources necessary in building an interlingually-aligned wordnet. The interface ensures the following main functions:

- Synset definition (sense assignment to the literals of the synonymy series and gloss attachment) and their mapping onto the ILI via a set of user defined equivalence relations. The default equivalence relations are those defined in EuroWordNet, but they can be modified according to the user's needs. Although WNBuilder allows linking by any used defined relation, VisDic (Horák and Smrž 2004), the BalkaNet standard development and viewing tool can only manage the EQ-SYN interlingual relation. This restriction was partly alleviated by the use of non-lexicalized synsets (see Tufiş et al. 2004b and section 2.3.3 below).
- Importing relations specified by the user from the source wordnet (PWN) into the target wordnet (RoWN).
- Validation functions. The most useful functions are: validating the syntax of the created synsets, search for sense assignment conflicts, duplicated literals in a synset, dangling nodes or relations, missing synsets, etc.

Sense assignment clustering are easy to spot (WNBuilder generates a detailed report on them), but not always easy to eliminate. At each major milestone of the BalkaNet project, the synsets implemented by each member of the development teams were merged and validated. The sense assignment clusters arising from putting together individually developed set of synsets were corrected on a centralized basis, by three trained linguists. For solving this very problem we developed another user-friendly interface called WNCorrect (Tufiş and Barbu 2004) which allows the lexicographer to correct sense assignment conflicts in a focused way. WNCorrect detects the necessity of sense clustering and therefore generates sense labels of the type (d) or (f). It offers the lexicographer the possibility to change the sense labels by dispaying all the implemented synsets containing the literal whose senses require clustering.

2.3.2 Assumptions and Limitations

One of the assumptions used in the creating a wordnet starting from an already existing one is that the two languages (for which wordnets are constructed) conceptualize the reality in the same manner. According to this assumption, only the EQ-SYNONYM interlingual relation would be necessary in this attempt.

However, the interlingual perspective on lexical data shows that the above assumption does not hold (see also Huang et al. 2002). We could identify three reasons for this:

- The concept exists in both languages, but in one of them it is not lexicalized. Such examples are represented by the English *handwear*, *headdress*, *drinking vessel*, etc., which lack Romanian lexicalizations, although the concepts exist for the Romanian speakers.
- The two languages conceptualize the reality slightly different, such that the two equivalent synsets should not be in EQ-SYN relation. If we consider the synset {jubilee:1} (gloss: a special anniversary (or the celebration of it)), its translation equivalent is the Romanian *jubileu*. However, this has the definition "an anniversary celebrating the passage of a number of years (usually fifty) since a certain event". So, we can say that the appropriate interlingual relation existing between the Romanian *jubileu* and the English *jubilee* is HAS-EQ-HYPERONYM. In this case, the Romanian equivalents (should they exist) of the hyponyms of *jubilee* (*diamond jubilee*, *silver jubilee*) would not be hyponyms of *jubileu*, but its co-hyponyms. So, the hierarchical structure should be different from one language to another.
- There are concepts that are not lexicalized in Romanian. Take for instance the English synset {carrel:2, carrell:1, cubicle:2, stall:5} (gloss: small individual study area in a library). The reality designated by it does not exist in the Romanians' life.

In order to overcome the limitations of using a single interlingual relation, we have made a study aimed to better understand the way synonymy is dealt with in the PWN. We started from a wide coverage of the literature dealing with lexical semantic relations and we identified the following criteria that two or more words have to obey to be considered synonyms:

1. substitution in the same context;
2. identity of grammatical category, identity of syntactic characteristics;
3. identity of referent;
4. identity of the stylistic characteristics;
5. difference of form.

The definition given by the WordNet theoreticians to synonymy is reproduced below after Miller et al. (1990):

Let C be a linguistic context and e_1 and e_2 two linguistic expressions. If e_1 şi e_2 can be substituted in C without modifying its truth-value, then they are synonyms.

Defined as above, synonymy presupposes identity of meaning (so the semantic distance between the synonyms is 0), identity of occurrence (that is the contextual distance between the synonyms is 0). The criteria (1), (3), (4) above are respected. The fifth criterion is also valid, although tacitly assumed.

Under these circumstances, wordnets should reflect perfect synonymy among the senses of the words included in the same synset. However, Fellbaum (1998) draws the attention on the fact that sometimes this does not hold: not all members of the same synset can occur in the same contexts; in such cases, in order to make the meaning clear enough,

different contexts are provided for (almost) each member of the synset. This can be interpreted in three ways:

- The context is syntactic in nature: ex.: {instantaneous1, instant1}: the second element of the synset can appear only in prenominal position, while the first one has no occurrence restrictions. The PWN lexicographers have provided two contexts for this synset: *relief was instantaneous*, *instant gratification*.
- The context is lexical/semantic in nature: {acme:1, height:2, elevation:2, peak:3, pinnacle:2, summit:1, superlative:2, top:5} (gloss: "the highest level or degree attainable"). Examples such as *the peak/acme of perfection*, *the peak/acme of the mountain*, *the summer/winter was at its peak*, **the summer/winter was at its acme* show that the synonyms do not combine with the same words.
- The context is stylistic in nature. Comments seem useless in the case of the synset: {stool:4, defecate:1, shit:2, take a shit:1, take a crap:1, ca-ca:1, crap:1, make:31}.

It is common knowledge that words that are synonyms for all their senses are rare in any language. We wanted to test this on PWN2.0. For this, we extracted those pairs (or sets) of words fulfilling three conditions: belong to the same part of speech; have an identical number of senses; appear in the same synsets. We obtained 1083 such synsets. This number is not to scare us and is not evidence to contradict what linguists have known for a very long time. A great number of these sets are represented by graphical variants of the same word (for instance *barbarize* and *barbarise*, *mouse-wood* and *moosewood*, *Saint Louis* and *St. Louis*, etc.). Still, there are also (few) polysemous words which are synonyms for all their senses: *rapacious* and *ravening*, for instance. They appear in the synsets: {predatory:2, rapacious:1, raptorial:2, ravening:1, vulturine:1, vulturous:1} (gloss: "living by preying on other animals especially by catching living prey"), {rapacious:2, ravening:2, voracious:1} (gloss: "excessively greedy and grasping"), {edacious:1, esurient:3, rapacious:3, ravening:3, ravenous:2, voracious:2, wolfish:2} (gloss: "devouring or craving food in great quantities"). Another example is *limpidity* and *pellucidity*, that appear in the synsets: {clarity:1, lucidity:1, pellucidity:1, clearness:1, limpidity:1} (gloss: "free from obscurity and easy to understand; the comprehensibility of clear expression"), {pellucidness:1, pellucidity:2, limpidity:2} (gloss: "passing light without diffusion or distortion").

The conclusion of the study was that one could overcome to a large extent the limitation of using only one interlingual equivalence relation (EQ-SYN) by making use of non-lexicalized (NL) synsets in the monolingual wordnet. For instance, if the synset S_1 in language L_1 should be properly linked to ILI_k by an EQ-HOLO_PART relation, this relationship can be modeled by using the internal relation HOLO_PART among S_1 and a non-lexicalized synset NL_1 plus an EQ-SYN between NL_1 and ILI_k:

EQ-HOLO_PART(S_1,ILI_k) $\cong$HOLO_PART(S_1,NL_1) & EQ-SYN(NL_1, ILI_k)

The NL synsets were used not only to compensate for a larger set of interlingual relations, but also to represent lexical gaps in our wordnet. One can show that other interlingual equivalence relations (EQ-HYPER, EQ-HYPO, EQ-MERO, etc.) can be easily modeled in terms of non-lexicalized synsets and EQ-SYN. The ambivalence of a non-lexicalized synset (proper non-lexicalized synset versus modeling a non-existent interlingual relation) is one of the main drawbacks of this solution. The other one is populating the monolingual wordnets with artificial synsets. However, the great benefit is that VisDic browser can handle the interlingual links and synchronize the multiple aligned wordnets in an extremely efficient way.

3 Current Status of the RoWN

The post-BalkaNet development of the RoWN continued, observing the conceptual density criterion, but some applications we were interested in (such as word alignment, word sense disambiguation, annotation import) made us pay attention to the lexical density criterion as well. Therefore, although the number of synsets significantly increased, the number of literals contained by our wordnet did not increase, as one would have expected, if only the conceptual density principle were observed. The quantitative data pertaining to the Romanian wordnet are summarized in the tables below.

Table 2: Internal relations used in the Romanian wordnet.

hypernym	23810	category_domain	1396
near_antonym	1810	also_see	453
holo_part	1710	subevent	213
similar_to	899	holo_portion	127
verb_group	1158	causes	143
holo_member	964	be_in_state	562

4 Applications

4.1 Word Alignment

Word Alignment is an extremely useful task in parallel corpora processing and it can be briefly stated as explicitly representing for each lexical item W_{1i} in one part of a bitext, which lexical item W_{2j} is translating it in the other part of the bitext. There is a large spectrum of applications for word alignment which covers bilingual lexicography, multilingual terminology, cross-lingual information extraction and retrieval, word sense disambiguation, machine translation.

Our main motivation for the development of the word alignment application was generated by the BalkaNet explicit goal to ensure high quality multilingual wordnets. The first word alignment system, TREQ (Tufiş 2002, Tufiş and Barbu 2002) was used for extracting an accurate bilingual (English-Romanian) lexicon, to be used as a supporting resource for WNBuilder (see section 2.3.1). At a later stage, it was used for building a WSD system for validating the

Table 1: POS Distribution of the Synsets.

Noun synsets	Verb synsets	Adj. synsets	Adv. synsets	Total
18158	5563	851	834	25406

monolingual wordnets alignments (Tufiş et al. 2004c, Tufiş and Ion 2004, Tufiş et al. 2004d). The most recent word aligner, called COWAL and rated the best in the word alignment shared task at ACL2005[6] (Tufiş et al. 2005), became the underlying module of our word sense disambiguation system and also the translation model generator for an MT under development.

COWAL is a complex integrated platform that takes as input two parallel raw texts and generates their alignment. The alignment platform includes language independent modules (sentence aligner, tokenizer, collocation detector, part of speech tagger, chunker, dependency linker, SVM classifier for filtering out improbable alignment links, alignment viewer and editor, etc.) and various language resources (language specific rules (regular expressions) for tokenization, chunking and dependency linking, language models for tagging, aligned wordnets for the languages concerned).

The wordnets for the languages of the texts to be aligned are used to reinforce the statistically established alignment links. On the other hand, the alignment links are used to extend the synsets (see below) with new lexical items. As the word alignment identifies contextual translation equivalents in arbitrary texts it is obvious that the aligned wordnets for the languages of the aligned bitext should explicitly represent these equivalences. If not, more often than not the explanation resides in the incompleteness of the synsets. Word alignment easily detects such a case, while the word sense disambiguation brings evidence on the very synsets which should be extended with new literals.

4.2 Word Sense Disambiguation

Word Sense Disambiguation (WSD) is well known as one of the most difficult problems in the field of natural language processing, as noted in (Gale et al. 1992) and others. The difficulties stem from several sources, including the lack of means to formalize the properties of context that characterize the use of an ambiguous word in a given sense, lack of a standard (and possibly exhaustive) sense inventory, and the subjectivity of the human evaluation of such algorithms.

We addressed these questions in several experiments involving sense clustering based on translation equivalents extracted from parallel corpora (Ide 1999, Ide et al. 2001, Ide et al. 2002). In Tufiş et al. (2005) we brought evidence that the granularity of the sense inventory used by a WSD program crucially influences the accuracy of the task and pleaded in favour of using, for whatever critical analysis of state-of-the-art in WSD and systems comparisons, a common sense inventory, publicly available. As such a public available sense inventory is the PWN our plea was in favour of using the original or whatever deterministic generalization of the PWN senses. Our multilingual experiments with the WSDtool system (Ion and Tufiş 2004, Tufiş et al. 2004c, Tufiş and Ion 2004) confirm that the accuracy of word sense clustering based on translation equivalents is heavily dependent on the number and diversity of the languages in the parallel corpus and the language register of the parallel text. The use of parallel texts and aligned wordnets, as those developed in the EuroWordNet or BalkaNet projects, demonstrated that the WSD accuracy can be increased beyond the performances of any monolingual WSD system. The underlying hypothesis in this approach exploits the common intuition that reciprocal translations in parallel texts should have the same (or closely related) interlingual meanings (in terms of BalkaNet, ILI record-projections or simply ILI codes). However, this hypothesis is reasonable if the monolingual wordnets are reliable and correctly linked to the ILI. Quality assurance of the wordnets was a primary concern in the BalkaNet project, and this was the primary motivation for the development of the WSDtool.

The methodology for the WSD based on parallel corpora and interlingually aligned wordnets assumes the following basic steps:

A) given a bitext T_{L1L2} in languages $L1$ and $L2$ for which there are aligned wordnets, one extracts the pairs of lexical items that are reciprocal translations: $\{<W^i_{L1}W^j_{L2}>^+\}$.

B) for each lexical alignment of interest, $<W^i_{L1}W^j_{L2}>$, one extracts, for each language, the ILI codes for the synsets that contain *literal*(W^i_{L1}) and *literal*(W^j_{L2}) respectively; thus, one gets two lists of ILI codes, $L^1_{LI}(W^i_{L1})$and $L^2_{ILI}(W^j_{L2})$, one for each language. The WSD of the lexical items under consideration comes to identify one ILI code common to the intersection $L^1_{ILI}(W^i_{L1})\cap\ L^2_{ILI}(W^j_{L2})$or a pair of ILI codes $\text{ILI}_1 \in L^1_{ILI}(W^i_{L1})$ and $\text{ILI}_2 \in L^2_{ILI}(W^j_{L2})$ so that ILI_1 and $\text{ILI}_2$are the codes of the most similar ILI concepts (below we elaborate on this issue) among the candidate pairs $(L^1_{ILI}(W^i_{L1})\otimes L^2_{ILI}(W^j_{L2})$ with $\otimes$ representing the Cartesian product among the two sets).

Step A) is crucial and its accuracy is essential for the success of the validation method.

Step B) is where the aligned wordnets come to work. The correctness of the interlingual alignment is essential in finding a pair of ILI codes that should label the translation equivalents. In the context of this research, we assume that the HPP is sound. Under this assumption, we take the *similarity* of two ILI codes $R1$ and $R2$ as a measure for the *semantic-similarity* between the synsets *Syn1* and *Syn2* in PWN2.0 that correspond to $R1$ and $R2$. We used a very simple definition of the semantic similarity (*sem-sim*)

[6] See also: Joel Martin, Rada Mihalcea, Ted Pedersen "Word Alignment for Languages with Scarce Resources" `http://acl.ldc.upenn.edu/W/W05/W05-0809.pdf`

between two synsets:

$$sem - sim(Syn1, Syn2) = 1/(1 + N) \quad (1)$$

where N is the number of oriented links from one synset to another or from the two synsets to the nearest common ancestor. The score is 1 when the two synsets are identical (or, equivalently said, they have the same ILI code), it is 0.33 for two sister synsets and 0.5 for mother/daughter or whole/part or any single link related synsets. Two ILI records $R1$ and $R2$ will be considered similar if

$$sim(R1, R2) = sem - sim(Syn1, Syn2) \geq t \quad (2)$$

where t is an empirical threshold. In our experiments we considered it 0.33 (i.e. we allowed at most two link traversal between what we consider two closely related synsets).

We should note at this point that *similarity* is meant as a language independent score which is approximated by an English specific score. This is justified, irrespective of ILI being structured or not, because the general model for all the wordnets was the PWN. Yet, since in BalkaNet ILI is PNW2.0 (thus, structured) the two measures are identical and, when the WSD task is considered among a pair of languages that includes English, the distinction we made seems a useless complication. However, if we consider the WSD task on a Czech-Romanian bitext, for instance, it is very likely that the topologies of the Czech and the Romanian wordnets differ. Therefore the *sem-sim* score would have different values, depending whether it was computed between the Czech synsets that correspond to R1 and R2 or between the Romanian synsets that correspond to the same ILI codes.

The PWN-based *sem-sim* mediates among different (but similar) wordnet topologies. The similarity of the wordnets topologies is a direct consequence of the HPP.

4.3 Annotation Import

One very promising research area on parallel corpora concerns the use of word alignment and word sense disambiguation technologies for importing syntactic-semantic information from one part of the bitext, richly annotated for the respective language, into the other part of the bitext where the linguistic annotation is scarce or simply missing. We investigated the feasibility of this enterprise by trying to import the valency frames defined in the Czech WordNet[7] into the Romanian wordnet. Very similar to the frames used in the FrameNet project[8], the valency frames are attached to the verbs[9] and specify syntactic and semantic restrictions for the arguments of the predicate denoting the meaning of a given synset. The valency frames also specify the case roles of the arguments. The nice property of the Czech valency 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 (a_se_afla:3.1, a_se_găsi:9.1, a_fi:3.1) with the gloss "be located or situated somewhere; occupy a certain position". Its valency frame is described by the following expression:(nom*AG(fiinţă:1.1)| nom*PAT(obiect_fizic:1)) = prep-acc*LOC(loc:1)[10].

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).

Via the interlingual equivalence relations among the Czech verbal synsets and Romanian synsets we imported about 600 valency 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 with minor modifications. This result, maybe previsied by lexical semanticists, supported by real data, motivated our further investigations on automatically acquiring FrameNet structures for Romanian and associating them with wordnet synsets. In cooperation with our partners at "Al.I.Cuza" of Iaşi, the project started with the translation of several English texts annotated by the members of the FrameNet project and their word alignment to the originals. This work developed within a larger international framework, under the name of the Romance FrameNet initiative (see `http://ic2.epfl.ch/~pallotta/rfn/`). The preliminary results for Romanian language were presented at the Romance FrameNet workshop and the kick of meeting, in Cluj-Napoca, on the occasion of the 7th EUROLAN International Summer School (`http://www.cs.ubbcluj.ro/eurolan2005/`).

Conclusions

We presented the further extension of the RoWN, after the end of the BalkaNet project, observing the conceptual density criterion and the lexical density criterion. In cooperation with IRST, the Romanian WordNet has been incorporated into the MultiWordnet (`http://multiwordnet.itc.it/online/`) as well as in the Memodata's Alexandria multilingual semantic lexicon (`http://www.memodata.com/2004/fr/Alexandria/`). Several multilingual basic applications (word alignment, word sense disambiguation, annotation import) heavily relying on the quality of the wordnet are under development for Romanian language. Among them, a MT system combining statistical methods and aligned wordnets is the most ambitious target.

Acknowledgements

The work reported here was initiated by the European project BalkaNet, no. IST-2000 29388 and was supported

[7] See Pala K., Smrž P. Building the Czech Wordnet. In *Romanian Journal on Information Science and Technology*, D. Tufiş (Ed.) Special Issue on BalkaNet, Romanian Academy, vol. 7, no. 2–3, 2004.

[8] `http://www.icsi.berkeley.edu/~framenet`

[9] Valency frames can be associated with deverbative nouns as well but we considered only verbs for this experiment

[10] fiinţă:1.1 = being:2;
obiect_fizic:1 = physical object:1;
loc:1 = location:1.

by the Romanian Ministry of Education and Research under the CORINT programme. The recent developments were achieved within the Romanian Academy program "Multilingual Acquisition and Use of Lexical Knowledge". The Romance FrameNet initiative got support from too numerous people to mention here, but special acknowledgements are due to Collins Baker, Charles Fillmore, Vincenzo Pallota, Emanuelle Pianta, Christian Girardi and Dan Cristea.

References

Bentivogli, L., Pianta, E. and Pianesi, F. Coping with lexical gaps when building aligned multilingual wordnets, in *Proceedings of LREC 2000*, Athens, Greece.

Bentivogli, L., Forner, P., Magnini, B. and Pianta, E. Revising WordNet Domains Hierarchy: Semantics, Coverage, and Balancing. In *Proceedings of COLING 2004 Workshop on "Multilingual Linguistic Resources"*, Geneva, Switzerland, August 28, 2004, pp. 101–108.

Bilgin, O., Cetinoglu, O., Oflazer, K. Morpho-semantic Relations in and Across Wordnets. In *Proceedings of the Global Wordnet Conference*, Brno, 2004, pp. 60–66.

Coteanu, I., Seche, L., Seche, M. (coord.). *DEX. Dicţionarul Explicativ al Limbii Române*, Second edition, Univers Enciclopedic, Bucureşti, 1996.

Cristea, D., Mihăilă, C., Forăscu, C., Trandabat, D., Husarciuc, M., Haja, G., Postolache, O. Mapping Princeton WordNet Synsets onto Romanian WordNet Synsets. In *Romanian Journal on Information Science and Technology*, D. Tufiş (Ed.) Special Issue on BalkaNet, Romanian Academy, vol. 7, no. 2–3, 2004.

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

Gale, W., Ward, Church K. and Yarowsky, D. Estimating upper and lower bounds on the performance of wordsense disambiguation programs. In *Proceedings of the 30th Annual Meeting of the Association for Computational Linguistics*, 1992, pp. 249–256.

Horák, A. and Smrž, P. New Features of Wordnet Editor VisDic. In Dan Tufiş (Ed.), *Romanian Journal of Information Science and Technology*, vol. 7, no. 2–3, 2004.

Huang, C.-R., Tseng, I-J. E., Tsai, D.B.S. Translating Lexical Semantic Relations: The First Step Towards Multilingual Wordnets. In *Proceedings of COLING 2002 SemaNet Workshop on Building and Using Semantic Networks*, Taipei, Taiwan, 2002.

Kilgarriff, A. "I don't believe in word senses". In *Computers and the Humanities* 31 (2), pp. 91–113, 1997.

Ide, N., Véronis, J. Word Sense Disambiguation: The State of the Art. *Computational Linguistics*, 24:1, 1-40, 1998.

Ide, N. Parallel translations as sense discriminators. *SIGLEX99: Standardizing Lexical Resources*, ACL99 Workshop, College Park, Maryland, 52–61.

Ide, N., Erjavec, T., Tufiş, D. Automatic Sense Tagging Using Parallel Corpora. In *Proceedings of the 6th Natural Language Processing Pacific Rim Symposium*, pp. 212–219, Tokyo, 2001.

Ide, N., Erjavec, T., Tufiş, D. Sense Discrimination with Parallel Corpora. In *Proceedings of the SIGLEX Workshop on Word Sense Disambiguation: Recent Successes and Future Directions*. ACL2002, July Philadelphia, pp. 56-60.

Ion, R., Tufiş, D. Multilingual Word Sense Disambiguation Using Aligned Wordnets. In *Romanian Journal on Information Science and Technology*, D. Tufiş (Ed.) Special Issue on BalkaNet, Romanian Academy, vol. 7, no. 2-3, 2004, pp. 198-214.

Julliard, A. *The Frequency Dictionary of Romanian*. Massachusetts: MIT Press, 1965.

Magnini, B. and Cavaglià, G. Integrating Subject Field Codes into WordNet. In Gavrilidou, M., Crayannis, G., Markantonatu, S., Piperidis, S. and Stainhaouer, G. (Eds.) *Proceedings of LREC 2000, Second International Conference on Language Resources and Evaluation*, Athens, Greece, 31 May – 2 June, 2000, pp. 1413–1418.

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

Niles, I., and Pease, A. Towards a Standard Upper Ontology. In *Proceedings of the 2nd International Conference on Formal Ontology in Information Systems (FOIS-2001)*, Ogunquit, Maine, October 17–19, 2001.

Rodriguez, H., Climent, S., Vossen, P., Bloksma, L., Peters, W., Alonge, A., Bertagna, F., Roventini, A. The Top-Down Strategy for Building EuroWordNet: Vocabulary Coverage, Base Concepts and Top Ontology. In *Computers and the Humanities*, 32 (2–3), 1998, 117–152.

Seche, L. and Seche, M. *Dicţionarul de Sinonime al Limbii Române*, second edition, Univers Enciclopedic, Bucharest, 1999.

Tufiş, D. *Blurring the distinction between machine readable dictionaries and lexical databases*. Research Report, RACAI-RR56, 1999.

Tufiş, D. A cheap and fast way to build useful translation lexicons. In *Proceedings of the 19th International Conference on Computational Linguistics, COLING 2002*, Taipei, 25–30 August, 2002, pp. 1030–1036.

Tufiş, D., Cristea, D., Methodological issues in building the Romanian Wordnet and consistency checks in Balkanet. In *Proceedings of LREC2002 Workshop on Wordnet Structures and Standardisation*, Las Palmas, Spain, May, 35–41.

Tufiş, D., Barbu, A.M., Revealing translators knowledge: statistical methods in constructing practical translation lexicons for language and speech processing. In *International Journal on Speech Technology*, Vol.5, No. 3, 2002 Kluwer Pbls.

Tufiş, D., Barbu, A.M., Ion, R. A word-alignment system with limited language resources. In *Proceedings of the NAACL 2003 Workshop on Building and Using Parallel Texts; Romanian-English Shared Task*, Edmonton, Canada, 36–39.

Tufiş, D., Cristea, D., Stamou, S. BalkaNet: Aims, Methods, Results and Perspectives: A General Overview. In

Romanian Journal on Information Science and Technology, D. Tufiş (Ed.) Special Issue on BalkaNet, Romanian Academy, vol. 7, no. 2–3, 2004a.

Tufiş, D., Barbu, E., Barbu Mititelu, V., Ion, R., Bozianu, L. The Romanian Wordnet. In *Romanian Journal on Information Science and Technology*, D. Tufiş (Ed.) Special Issue on BalkaNet, Romanian Academy, vol. 7, no. 2–3, 2004b.

Tufiş, D., Ion, R., Barbu, E., Barbu Mititelu, V. Cross-Lingual Validation of Wordnets. In *Proceedings of the 2nd International Wordnet Conference*, pp. 332–340, Brno, 2004c.

Tufiş, D., Ion, R. Interlingual wordnets validation and word-sense disambiguation. In *Proceedings of the Natural Language Understanding and Cognitive Systems Symposium*, Porto, 2004, pp. 97–105.

Tufiş, D., Barbu, E. A Methodology and Associated Tools for Building Interlingual Wordnets. In *Proceedings of the 5th LREC Conference*, Lisbon, 2004, pp. 1067–1070.

Tufiş, D., Ion, R., Ide, N. Fine-Grained Word Sense Disambiguation Based on Parallel Corpora, Word Alignment, Word Clustering and Aligned Wordnets. In *Proceedings of the 20th International Conference on Computational Linguistics, COLING2004*, Geneva, 2004d, pp. 1312–1318.

Tufiş, D., Ion R. Evaluating the word sense disambiguation accuracy with three different sense inventories. In *Proceedings of the Natural Language Understanding and Cognitive Systems Symposium*, Miami, Florida, May 2005, pp. 118–127.

Tufiş, D., Ion, R., Ceauşu, Al., Ştefănescu, D. Combined Aligners. In *Proceeding of the ACL2005 Workshop on "Building and Using Parallel Corpora: Data-driven Machine Translation and Beyond"*, Ann Arbor, Michigan, June, 2005, pp. 107–110.

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