ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY
Volume 7, Numbers 1–2, 2004, 107–124

The Romanian Wordnet

Dan TUFIŞ, Eduard BARBU, Verginica BARBU MITITELU,
Radu ION, Luigi BOZIANU
Research Institute for Artificial Intelligence,
Romanian Academy, Bucharest
E-mail: {tufis, eduard, vergi, radu, bozi}@racai.ro

Abstract. This paper presents the steps followed in the development of the Romanian wordnet. First we describe an inventory of the resources we used for this purpose and the methodology we adopted. Then we present the tools we used for the construction and the syntactic validation of the Romanian wordnet. In the end we provide quantitative data about current status of our wordnet.

1. Introduction

The paper describes the methodology and the tools we developed for the purpose of building a Romanian wordnet, part of the BalkaNet multilingual [10] lexical semantic network which, via an inter-lingual index (ILI), allows for navigation from the words in one language to the words that express similar meanings in the other languages. The inter-lingual index is a repository of conceptual knowledge presumably lexicalized in most natural languages. The details on ILI structuring are provided elsewhere in this volume [11], but for the sake of this presentation it is convenient to regard it as a semantic network whose nodes represent language independent concepts linked by labeled arcs representing semantic relations among the respective concepts. The requirements in building a multilingual ontology of the EuroWordNet kind are frequently conflicting [8] and if not considered in the early stages of the project, later harmonizing might be extremely difficult, if possible at all. Specifically, there are two main compatibility issues to be considered: first, structuring principles and the interpretation of the defined relations should be the same in any monolingual Wordnet considered by the multilingual harmonized ontology, and second, there should be a significant cross-language conceptual coverage, meaning that each monolingual semantic lexicon should globally deal with the same conceptual areas or domains. The first aspect was easy to solve as the BalkaNet consortium programmatically adopted the EuroWordNet principles and the relations inventory. The second issue required

special attention and significant efforts were invested in making the choice for the ILI concepts to be cross-lingually covered by all BalkaNet wordnets. The choice of these concepts (BCS: BalkaNet Concept Set) has been made in three steps as described in [12]. Step 1 included all the concepts called Base Concepts in EuroWordNet. Steps 2 and 3 were based on the concepts implemented in most EWN wordnets plus the common concepts proposed by each partner. Thus we aimed at a significant cross-lingual coverage not only among the consortium's wordnets, but also among the BalkaNet wordnets and the EWN wordnets. The concept sets proposed by the BalkaNet partners were selected based on language specific criteria, supported by various linguistic resources and tools, and as such, the proposals served the purpose of finding the common set to be adopted by everybody and also defined a conceptual stock to be taken care by wordnets developers after the project ended.

The rest of the paper is organized as follows: in section 2 we present the basic language resources we used for the Romanian wordnet; in section 3 we describe the procedure for identifying the most relevant concepts in ILI to be implemented in the Romanian wordnet; in section 4 we present the methodology we adopted and the tools implementing it; section 5 gives a statistical account of the current[1] Romanian wordnet; section 6 presents future work and conclusions.

2. The Language Resources Used for Building the Romanian Wordnet

Due to the general concern of several lexicographers, according to whom translating the Princeton WordNet synsets would not result in a semantic dictionary representative for the target language (it would be an excellent dictionary for understanding, in one's own language, the semantic subtleties of American English lexical stock), we adopted a language centric approach (as opposed to a simpler method based on the translation of the literals in the Princeton WordNet), relying on reference lexicographic resources: the Explanatory Dictionary of Romanian, The Dictionary of Synonyms, as well as an in-house Romanian-English dictionary.

The Explanatory Dictionary of Romanian (EXPD) is a general dictionary of modern Romanian authored by the Institute of Linguistics of the Romanian Academy and contains about 56 000 entries[2]. We further extended this dictionary so that our current version contains almost 70 000 entries. EXPD is XML heavily annotated according to the encoding schema developed in the previous CONCEDE project[3] [17], [18]. A simple entry, exemplifying the CONCEDE encoding, is shown in Figure 1.

The tools (see next section) we use for the Romanian wordnet development require a simplified XML encoding. For the dictionary entry shown in Figure 1, the simplified format is given in Figure 2.

[1] We are currently five months before the end of the project, but the Romanian wordnet development is a continuous process, so that by the end of the project (August, 2004) the statistics will certainly change.

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

[3] For information on the CONCEDE encoding schema see: `http://www.itri.brighton.ac.uk/projects/concede/`

```
<entry id="ABANDONAT">
   <hw>ABANDONAT</hw><stress>ABANDON‘AT</stress>
   <alt><brack><gram>nominativ_masculin_singular_indefinit</gram>
               <orth extent="full">ABANDONAT</orth>
         </brack>
         <brack><gram>nominativ_feminin_singular_indefinit</gram>
                <orth extent="full">-A</orth>
         </brack>
         <brack><gram>nominativ_masculin_plural_indefinit</gram>
                <orth extent="full">abandonaţi</orth>
         </brack>
         <brack><gram>nominativ_feminin_plural_indefinit</gram>
                <orth extent="full">-te</orth>
         </brack>
   </alt>
   <pos>adjectiv</pos>
   <struc><def>Care a fost părăsit.</def>
          <struc type="Sec">
                <usg type="other">Despre copii nou-născuţi</usg>
                <def>Lepădat <xptr target="Lepădat.2" targOrder="u" /> </def>
          </struc>
   </struc>
   <etym>Vezi abandona <xptr target="abandona"targOrder="u" /> </etym>
</entry>
```

Fig. 1. The XML encoding of an adjective headword in XML-EXPD.

```
<entry>
    <word>abandonat</WORD>
    <pos>adjectiv</pos>
    <def>1. Care a fost părăsit.</def>
    <def>2. <usg>Despre copii nou-născuţi</usg> Lepădat. </def>
    <etym>Vezi abandona</etym>
</entry>
```

Fig. 2. The simplified XML encoding of an adjective headword in XML-EXPD.

The <etym> tag is optional and it can be used to derive some lexical relations (here, there is a link to the verb from which the adjective is derived). The <usg> is also optional and provides the typical context of use.

The Synonyms Dictionary (SYND), also authored by the Institute of Linguistics of the Romanian Academy [9], was keyboarded, XML encoded and completed with more than 4 000 new synonymy sets extracted from EXPD. The XML encoding of SYND is exemplified in Figure 3.

```
<synset-rec><number>10</number><pos>adj</pos>
  <synset><elem><lit>abandonat</lit></elem>
          <elem><lit>părăsit</lit></elem>
          <elem><usg>înv., pop.</usg><lit>oropsit</lit></elem>
          <elem><usg>înv., reg.</usg><lit>năpustit</lit></elem>
  </synset>
    <example>Copil ~.</example>
</synset-rec>
```

Fig. 3. The XML encoding of a synonymy set entry in XML-SYND.

The synonyms series in SYND contained both words used in modern language and archaisms and/or regionalisms (marked as such as content of the <usg> tag). We removed the archaic and regional variants with provision for automatic inclusion if ever needed. In its simplified form as needed here, the SYND is a text file with one synset per line and the literals separated by commas. The simplified form of the entry in Figure 3 (with archaisms removed) would be: `abandonat,părăsit.`

The Romanian-English dictionary was automatically extracted from a parallel corpus (see below) by our translation equivalence based TREQ-AL word aligner [13]. A pair of sentences which are reciprocal translations is called a translation unit (TU). The parallel corpus can be seen as a consecutive numbered sequence of TUs. Although translation equivalents extracted by TREQ-AL may have different POSes, for the needs of our wordnet development only translation equivalents preserving part of speech were retained. Below there are shown four entries (slightly edited for readability sake), one for each part of speech relevant in wordnets, displaying, in addition to the translation equivalents, their identical part of speech, the confidence score in their equivalence (log-likelihood score), examples of TUs where the translation pair has been observed, the average relative distance among the words in the translation pairs (the average difference between the offsets of the two words in the two sentences of a TU) and a cognate-score for the two words (an orthographic similarity distance). Further details on the translation equivalents and word alignment are provided in [16], [15], [13].

```
...
kitchen bucătărie       nc    LL-score: 102.186  TU-examples: 48 49 272...
relative-distance=0.6   cognates-score =0
pretend pretinde        vm    LL-score: 29.023   TU-examples: 3336 4515...
relative-distance=0.25  cognates-score =0.857
patient răbdător        a     LL-score: 14.077   TU-examples: 4194
relative-distance=0     cognates-score =0
often des               r     LL-score: 12.978   TU-examples :26 2197...
relative-distance=1     cognates-score =0
...
```

Fig. 4. Translation dictionary.

The bilingual dictionary was hand validated and extended so that in the current version it has 74 111 entries. For the wordnet development tool, we used a simplified version preserving only the translation equivalents and their common part of speech (the first three columns in Figure 4).

Beside these language specific lexical resources we also used the XML format of the PWN. An example of an XML encoded PWN synset is shown in Figure 5. The structure of a synset shown in Figure 5 is the same for all wordnets developed in the BalkaNet project (see [4] in this volume).

The lexical resources mentioned above were complemented by a large tokenized, tagged and lemmatized text corpus of contemporary Romanian (web-published newspapers, fiction and several technical reports, altogether containing more than 100 000 000 lexical tokens) and a parallel corpus of Romanian–English texts containing about 900 000 tokens per language. The Romanian–English bitexts included into the parallel corpus were one-year web issues of *Evenimentul Zilei* newspaper,

Orwell's novel *1984* and the *Romanian Constitution*. The parallel corpus is segmented, lemmatized, tagged, sentence and word aligned.

```
<SYNSET><ID>ENG20-01004767-a</ID><POS>a</POS>
     <SYNONYM><LITERAL>abandoned<SENSE>2</SENSE></LITERAL>
     <LITERAL>deserted<SENSE>1</SENSE></LITERAL> </SYNONYM>
     <ILR><TYPE>similar_to</TYPE>ENG20-01004545-a</ILR>
     <DEF>left desolate or empty</DEF>
     <USAGE>an abandoned child</USAGE>
     <USAGE>their deserted wives and children</USAGE>
     <USAGE>an abandoned shack</USAGE>
     <USAGE>deserted villages</USAGE>
</SYNSET>
```

Fig. 5. The XML encoding of a synset in XML-PWN2.0.

3. Motivations and the Selection of the Target Inter-Lingual Concepts

In order to ensure practical utility for the core Romanian wordnet developed during the project period and to facilitate further extensions, we conducted a series of statistical investigations on both corpora mentioned before in order to obtain reliable frequency data on the modern use of Romanian and to make use of this data as one criterion for the selection of ILI concepts to be implemented in the Romanian wordnet. All lemmas contained in the corpus were sorted according to their frequency and grouped in three sets, corresponding to the first 10 000 most frequent lemmas (I), the next most frequent 10 000 lemmas (II) and rest of the lemmas (III). The word frequency in running texts is considered by many lexicographers to be a questionable criterion in deciding on what is the most important subset of a language lexical stock. Among the strongest arguments they would come up with are the volume of texts and how representative they are with respect to the general language description. With more and more texts available on the Internet, the size of the data is not a significant issue anymore, but the relevance remains a systematic complaint. The exact definition of what representative texts should be included into a language reference corpus for quantitative data analysis is a long-standing debate and we would not get into this. Given that our data consisted, almost exclusively, of journalistic texts, the relevance issue could certainly be raised. Therefore, we checked our list of nouns and verbs against The Frequency Dictionary of Romanian – FDR [6], which, although published long time ago, and rather contested, is still used by many Romanian linguists as a reference. 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 the most frequent 5 000 Romanian lemmas (in that corpus). The comparison we made revealed that all 5 000 words in FDR were also in our list, although not with the same frequency ranks. From the three frequency lists we retained only the words of interest for the wordnet structuring (nouns, verbs, adjectives and adverbs).

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 (NS) a headword has in a reference dictionary. The second one is definitional productiveness (DP), that is the number of sense definitions a word participates in. Considering only the first two frequency ranges described above we extracted from our Explanatory dictionary a list of 8 000 nouns and nominal compounds and more than 6 000 verbs (accounting for about 50 000 senses) so that the DP was at least 3 (Table 1). Via the bilingual dictionary, we obtained from the English literals the set of ILI records that might represent the projections of the senses for our selection of the most representative nouns, nominal compounds and verbs. The set of ILI candidates was sorted according to the sorted list of Romanian equivalents. This sorted list is much larger than the set needed for the BalkaNet final deliverables, but it is very useful for the priority concepts in the future development.

Table 1. The selection parameters
for the most relevant 8 000 Romanian nouns

Word (noun)	**DP**	**NS**	$\mathbf{f}_{range}$
acţiune	2 279	14	I
persoană	1 979	9	I
parte	1 882	94	I
formă	1 286	21	I
obiect	1 204	16	I
fapt	1 044	11	I
apă	743	29	I
. . .	. . .	. . .	. . .

The BalkaNet Common Set of concepts is conceptually dense, that is for any concept in this set, all its hyperonyms, up to the top level concepts, are also in the set (for further details, see [12] in this volume).

More often than not, the conceptual density criterion (adopted for ensuring and controlling the cross-lingual coverage) requires the implementation of only some of the senses of the literals represented in our wordnet. For instance, the word *acţiune,* has in EXPD 14 senses, but in the current version of the Romanian wordnet only seven are implemented. Ensuring the implementation in a given wordnet of all senses a reference dictionary (in our case EXPD) describes is what we call *lexicographic density.* This property is obviously language dependent both on the different lexicalizations of the concepts represented in the interlingual index and on the explanatory dictionary taken as reference. The lexicographic density issue was outside the scope of the BalkaNet project and it should be dealt with by each partner at a later stage.

For the task of choosing the adjectives and adverbs we used the parallel Romanian–English corpus. Selection of the adjectives and the adverbs was done in the following three steps:

1. We computed the frequency of English adjectives and adverbs in the corpus.
2. All the PWN synsets which contain the words listed at the previous step were extracted as selection candidates.
3. We computed the score for the whole synset as the sum of frequency of each member of the synset; the literals occurring in more synsets, contributed with their frequency to each distinct synset.

From the sorted list of English synsets (equivalent to ILI concepts) we selected the top scored 900 adjectival and 800 adverbial concepts for implementation in our wordnet.

4. The Development Methodology and the Associated Tools

The Romanian wordnet is developed by two teams[4] of experienced computer scientists and linguists that work in close collaboration. For the process of the proper building of Romanian synsets, closest to the meaning of the concepts in the set of selected ILI records, the lexicographers were explicitly instructed to choose one of the synonymic series in the SYND. They were also instructed to attach sense labels according to the EXPD labelling and to use 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 label (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 targeted concepts; the sense 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 labelling 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 Explanatory Dictionary of Romanian by means of a sense labelling notation, we kept it in our wordnet with the same interpretation. More precisely, the sense labelling in the Romanian wordnet conforms to the BNF notation in Figure 6.

A sense-identifier of the **type (a)** is the usual case and the integer is the sense number found in the Explanatory Dictionary of Romanian, our lexicographic reference. A sense-identifier of **type (b)** is also the labelling used in the Explanatory Dictionary of Romanian and we kept it as it represents information that we do not want to loose, i.e. semantic relatedness of senses. It stands for the {<integer3>th sub-sense of the} <integer2>th sub-sense of the <integer1>th sense of the current literal. A sense identifier of **type (c)** defines a sub-sense of <integer>th sense which due to the coarser granularity of our reference dictionary is not explicitly mentioned in the Explanatory Dictionary of Romanian.

```
<sense-identifier>::=<integer>|                                (a)
                     <integer1>.<integer2>{.<integer3>}|       (b)
                     <integer>.<letter>|                       (c)
                     <integer1>.c                              (d)
                     <letter>                                  (e)
```

Fig. 6. The sense labelling in the Romanian wordnet.

[4]The Romanian parts involved in this project are: the Research Institute for Artificial Intelligence of the Romanian Academy (coordinator: Dan Tufiş) and the Faculty of Informatics of the University "A. I. Cuza" of Iaşi (coordinator: Dan Cristea).

Multiple sub-senses of a given sense should be numbered according to the frequency of use; when we will be able to evaluate sense frequencies, the notation of type (c) will be turned into a notation of type (b). A sense identifier of **type (d)** defines a sense clustering. Due to the difference in granularities among PWN and our reference explanatory dictionary, some sense distinctions in PWN are not naturally justified (at the lexical level) in Romanian. The need for sense clustering became apparent while we tried to solve sense assignment conflict (see below). Such cases could be easily represented by an EQ-NEAR-SYN (n-m) relation. However, as in the BalkaNet standard browser (VisDic) only EQ-SYN (1-1 relation) is available for interlingual mapping, we simulated the EQ-NEAR-SYNrelation by duplicating the Romanian synsets and using the **type (d)** sense labelling. We achieve this way a simple representation for the clustering (the final "c" in a sense label stands for "clustering") of the English meanings which are not lexically distinguishable in Romanian. A sense identifier of **type (e)** represents a sense that is not listed in the Explanatory Dictionary of Romanian but we considered it as a legitimate distinct one. In this case, the gloss represents simply the translation of the corresponding sense in PWN (adjusted, if necessary). Instead of a letter we could have used one integer larger than the one of the last definition listed in the reference dictionary. However, because the lexical density of the Romanian wordnet was not yet addressed (meaning that currently not all the senses of every word are included in the wordnet) we don't have enough information to order them. When sense frequency can be estimated (automatically or by professional introspection) this type of sense labeling should be turned into a type (a) with possible relocation of the other sense numbers.

With a large team of lexicographers working in parallel and the very fine-grained sense inventory of the PWN, sense assignment conflicts (literals with the same sense labels occurring in more than a single synset) are not surprising in our merge approach. Detecting sense assignment conflicts is simple, but eliminating them requires significant efforts. There were five types of sense assignment conflicts, generated by the much finer granularity of PWN as compared to EXPD & SYND:

- sense distinctions in PWM with a metonymic flavor (e.g. quality for the act) represent, by far (56%), the most frequent source of sense assignment conflicts in our wordnet: {dishonesty[2], knavery[1]} (GLOSS: lack of honesty; acts of lying or cheating or stealing) and {dishonesty[1]} (GLOSS: the quality of being dishonest);

- an English hyperonym and one of its hyponyms have as a Romanian equivalent the same literal with the same sense identifier: the synset {end[2], ending[3]} (GLOSS: the point in time at which something ends) and its hyponym {stopping point[1], finale[1], finis[1], finish[5], last[1], conclusion[3], close[1]} (GLOSS: the temporal end; the concluding time) are given sfârşit(1.1.3) as a Romanian equivalent;

- two English co-hyponyms were given the same equivalent in Romanian: for {mister[1], Mr[1]} (GLOSS: a form of address for a man) and {sir[1]} (GLOSS:

term of address for a man) the lexicographers provided {domn(1.1)} as the equivalent;

- the EXPD gloss of a Romanian literal covers the meaning of two English synsets, themselves not very well differentiated: ţâr(2.1) as compared to {herring[1]} (GLOSS: valuable flesh of fatty fish from shallow waters of northern Atlantic or Pacific; usually salted or pickled) and {kipper[1], kippered herring[1]} (GLOSS: salted and smoked herring);
- The inclusion of some metonymical senses in PWN, which lack from our EXPD, may generate conflicts: only for {cabinet minister[1]} (GLOSS: a person who is a member of the cabinet) there is a correspondent in EXPD, namely {ministru[1]} (GLOSS: înalt funcţionar de stat, membru al guvernului, care conduce un minister), while for its metonym, {cabinet minister[2]} (GLOSS: the job of a senior minister who is a member of the cabinet), there is none. A remark is worth being made at this point: inclusion of figurative senses in WordNet is a highly debatable problem. We agree that those senses which have become conventionalized deserved being included in the dictionary, but we aim at a systematic way of doing that: either all instances of a type of metonymy should be included in the wordnet, or they all should be left aside.

Solving these types of conflicts assumed either modifying the offending synsets or clustering them as previously mentioned. Beside these categories of "objective" sources of sense assignment conflicts, we discovered several errors due to lexicographers' wrong decisions in equivalence mappings. For instance, the Romanian synset {petală [1]} has been wrongly mapped on both {floral leaf [1]} and {petal [1]}, when only the second equivalence is valid.

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 to the Romanian wordnet. As a result of this investigation, in [14] it is conjectured the *Hierarchy Preservation Principle* (HPP) which is the basic motivation for automating the import of most of the semantic relations 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 [1].

In the table below we present the relations, encoded in the XML representation of the PWN 2.0 database, which we considered to be importable in our wordnet (and probably in the wordnets for many other languages). Their names are sometimes slightly different from the original names that are used in PWN, but the semantics of these relations is the same. In the table below, the first column lists the relations; the second shows the parts of speech of the synsets that the relation may link; the last column states to what degree the import of the respective relation in the Romanian wordnet was automatized.

Hyperonymy, denoted by *hypernym*, is a semantic relation which establishes a specific-generic relationship between the related meanings. When it is established

between meanings realized as noun synsets, the relation does not discriminate among individuals and classes (in a set-theoretic interpretation), or between instances and types (in a typed logic interpretation): (bush:4)→(president:2); (bush:1)→(woody plant:1).

Table 2. The relations in PWN 2.0 which are subject to import in the Romanian wordnet

Relation	Valid POSes for the relation	Imported
hypernym	<N, N>; <V, V>	yes
holo_part	<N, N>	yes
holo_portion	<N, N>	yes
holo_member	<N, N>	yes
subevent	<V, V>	yes
causes	<V, V>	yes
verb_group	<V, V>	yes
be_in_state	<A, N>	yes
similar_to	<A, A>	yes
also_see	<V, V>; <A, A>	yes
category_domain	<N, N>; <V, N>; <A, N>; <B, N>	yes
near_antonym	<N, N>; <V, V>; <A, A>; <B, B>	yes but with restrictions
derived	<A, A>; <B, A>; <A, N>	partially

In the case of verbs the relation is a particular kind of lexical entailment with the activity denoted by the more specific verbal meaning being temporally coextensive with the activity denoted by the more general verbal meaning. One can say that the activity denoted by the more specific argument (synset, meaning) is a manner elaboration of the activity denoted by the more general argument (synset, meaning): (run:1)→(travel rapidly:1).

Holonymy (part-whole relationship). PWN distinguishes three kinds of holonymic relations:

- Member parts which in the BalkaNet XML version of PWN is encoded as *holo_part* (PART-OF in PWN20): (finger:1)→(hand:1)
- Substantive parts which in the BalkaNet XML version of PWN is encoded as *holo_portion* (SUBSTANCE-OF in PWN20): (wood:1)→(timber:1)
- Component parts which in the BalkaNet XML version of PWN is encoded as *holo_member* (MEMBER-OF in PWN20): (tree:1) →(forest:1)

For verbs there are two kinds of entailment denoted as:

- *subevent* – when the activity denoted by one argument of the relation is temporally properly included in the activity denoted the by the other argument: (snore:1)→(sleep:1).

- *causes* – when the verb concepts are one causative and the other resultant: (kill:1)→ (die:1).

verb_group. This relation groups several similar overlapping meanings of the verbs: (act:2 behave:1 do:9)→((act:5 play:8 act-as:2)(dissemble:3 pretend:2 act:9))

be_in_state. This is a relation that specifies a value for a property. The values related by ***be_in_state*** are represented by descriptive adjectival synsets and the properties by nominal synsets: (tall:1)→(stature:2 height:3).

similar_to. This is a relation between adjectival meanings analogous to the ***verb_group*** relation for verbal meanings. It relates an adjectival meaning (the head) to a set of similar adjectival meanings (the head's satellites). The special status of the head is given by the fact that it has a direct antonym, which is inherited (as an indirect antonym) by all the adjectival meanings in the satellites:

((tall:1) vs. (short:3))→((full-length:1)(gangling:1)....(tallish:1)).

It is worth mentioning that although the mapping of the Romanian head adjectives to the PWN 2.0 equivalents did not raise significant problems[5], this was not the case with their respective satellites and for the present moment several English adjective satellites have no equivalents in the Romanian wordnet.

also_see. It is a relation which links semantically related verbs and, via their heads, similar adjectival clusters:

((tall:1 vs. short:3))→(((high:2) vs. (low:2))((large:1 big:1) vs. (small:1 little:1))).

category_domain. This is a special relation that allows topical classifications of the meanings represented by the synsets; the target synset of the relation is always a nominal synset, but any synset, irrespective of its POS can be source of this relation (put it otherwise, can be topical classified). It is always imported: diplomatic immunity:1)→(law:2 jurisprudence:2)

near_antonym. Because this is a lexical relation between word forms, antonymy is not granted for fully automatic import. Nevertheless, we found that this relation holds in most cases. We imported it, but manually checked it and whenever necessary modified it or its synset arguments.

derived. This is also a lexical relation that links derivatives to their stems (adverb&adjective: quickly→quick), (adjective&adjective: astomatal→stomatal; adjective&noun: abbatical→abbey). We found that when the relation was established between adjectives and adjectives or nouns they pertain to (<a a>; <a n>), more often than not, the relation holding in PWN could be imported in the Romanian wordnet. Since the -ly suffixation mechanism which underlies the ***derived*** relation between adverbs and their stem adjectives in English has no equivalent in Romanian, the <b a> case of the ***derived*** relation was excluded from import and further validation.

By virtue of the hierarchy preservation principle mentioned earlier and the BalkaNet consortium agreement on non-lexicalized synsets, all the semantic relations described above were automatically imported as follows: if the two source synsets

[5]One of the problems encountered was the presence in head position of adjectives at synthetic degrees of comparison (*better*, *worse*, *best*, *worst*, etc.). We decided not to implement such adjectives (not even in cases of mere satellites) but we faced a new problem, namely that of reorganizing the clusters headed by such antonymic pairs.

$S_{1SOURCE}$ 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 partial transitive ($k = 3$). The lexical relations mentioned in Table 2 were imported according to the same algorithm, but each relation import was manually checked.

4.1. WNBuilder

The WNBuilder is a configurable graphical interface, click controlled, 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 interlingual index 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 needs.
- Importing relations specified by the user from the source wordnet (PWN) into the target wordnet.
- 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.

Although WNBuilder can be used with any pair of Source/Target languages (provided the required language resources are available) we will exemplify for the English/Romanian languages. In Figure 7 there is a snapshot of the WNBuilder interface showing four frames that we will denote with: UL (upper left) frame, UR (upper right) frame, LL (lower left) frame, LR (lower right) frame. The UL frame displays the list of ILI codes (initially blue colored, signifying a non-visited ILI record) that are in the lexicographer's responsibility.

Clicking any ILI code will turn its color into red (signifying a visited ILI record, not yet implemented) and in the UR frame there will appear:

- the English synset (and its associated gloss) which is mapped onto the respective ILI record;
- a list of translation equivalents for the words in the English synset. The translation equivalents are taken from the bilingual dictionary. By selecting (clicking) one translation equivalent in this list, the interface will display the following information:

– the definitions of the selected translation (in the LL frame; they are extracted on the fly from EXPD);

– all the synonymy sets which the selected translation belongs to (in the LR frame; they are extracted on the fly from SYND). Each literal in a synonymy set is linked to a headword entry in EXPD so that the lexicographer has the possibility to see all the definitions for each word in the current synonymy set.

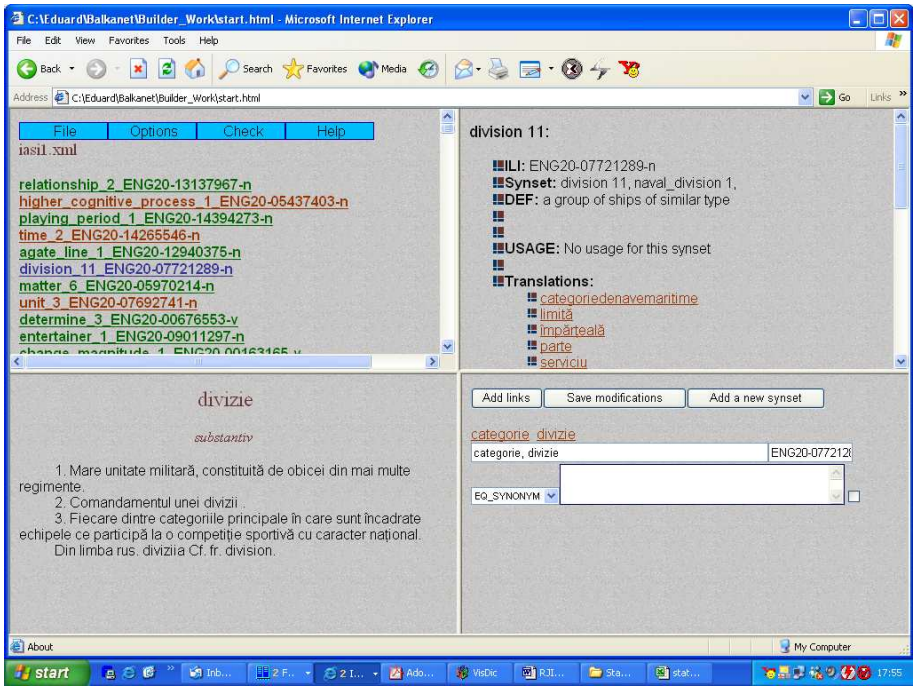

Fig. 7. The WNBuilder graphical interface.

With this information displayed in a friendly format, the lexicographer has to answer four main questions and make decisions that in the end would result in a target language synset, mapped to the starting ILI-record:

1. which are the best translation equivalents for the literals in the selected English synset; the lexicographer has the possibility to add new translation equivalents;

2. which of the synonymic sets best fits the English synset. The lexicographer can add or delete words from each of the synonym set, or can create his/her own synonym set if a relevant one is not present in SYND;

3. which of the definitions (if different) of the translation and its synonyms best fits the English gloss;

4. which is the interlingual relation between the English synset and the Romanian synset under construction; the interface gives the lexicographer the possibility to select among a set of interlingual relations.

After the lexicographer completed one or more target synsets and mapped them on the ILI records (via interlingual relations) s/he may launch the syntactic validation functions of WNBuilder[6]. The color of the ILI codes in the UL frame linked to the synsets that were completed and passed the validation tests turns green (signifying an already implemented ILI record). The ILI codes may be ordered according to their colors so that the unvisited or not yet implemented ILI records come to the top of the UL frame (the blue and red codes). The completed synsets (name stamped) may be at any moment saved in VisDic[7] compatible XML format [4]. If errors are still present in the generated semantic sub-network, they are recorded into a separate file for the subsequent correction.

We have shown in the previous section that sense assignment conflicts were 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 two Romanian wordnet development teams were merged and validated. The sense assignment conflicts 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 which allows the lexicographer to correct sense assignment conflicts in a focused way.

4.2. WNCorrect

WNCorrect is implemented in two functionally equivalent variants, but for the sake of lexicographers' validation preferences one is oriented on literals with sense assignment conflicts and the other one is centered on the synsets containing at least one literal with the same sense label. The graphical interfaces of the two variants of WNCorrect have similar designs as the one used by WNBuilder.

Working with WNCorrect1 assumes the following strategy:

- Identifying the literals with senses in conflict, i.e. the same literal appearing in two or more synsets with the same sense;
- Collecting all synsets containing those literals.

Each member of the validation team has a distinct set of sense conflicting literals; their list is displayed in the upper left frame; when clicking such a literal, in the upper right frame appears the list of synsets containing the conflicting literal. The lexicographer is supposed to change the sense identifier (and when assigning a sense not listed in the reference dictionary also to provide a gloss), or to delete the literal from the synset if it does not belong to that synset. The advantage of this procedure

[6] These validation functions are also automatically launched each time the work of the lexicographer is saved.

[7] VisDic is the standard browser and maintenance system for the BalkaNet multilingual wordnets.

is that, at the end of the validation task, there will be no conflicts left in the wordnet, as the interface does not allow saving the work if there still are conflicts to be solved. But deleting the literals from synsets may lead to some empty synsets. They may be legal empty synsets, representing non-lexicalized concepts in Romanian. In such a case, their encoding in the XML format is different. However, not all the synsets that became empty after conflicts removal correspond to non-lexicalized concepts. They have to be implemented again using the WNBuilder interface, and thus new conflicts may still appear. Another problem with this version is that the correction procedure does not allow the lexicographers to modify anything but the conflicting literals in synsets, leaving the others as they are (even if they are wrong). Moreover, the same synsets are possibly checked by several lexicographers.

Using the second variant, WNCorrect2, assumes the following strategy:

- Identifying the synsets with literals in conflict;
- Individual lexicographers are given disjoint sets of synsets with conflicting literals.
- As the lexicographer is now responsible for the correctness of the whole synset, s/he is allowed to modify the senses of the literals within the synset, to delete literals from the synset or add literals. That is the greatest advantage of this procedure: full control over the synset by a certain lexicographer.
- Checking on the fly the work of the lexicographer for new conflicts with the help of a function implemented in WNCorect2. If there are any, they will be solved by the same lexicographer.
- The corrected synsets replace the initial ones in the wordnet XML file and the procedure is repeated from the first step until there are no more conflicts left.

The team working on the sense assignment conflict resolution started to use WNCorrect1 but all three members soon found WNCorrect2 variant (see Figure 7) more convenient in converging faster towards a sense-conflict free wordnet.

The conflict resolution process is a cyclic one: extending the wordnet in the distributed regime we mentioned, is likely to introduce new conflicts as new synsets are added. However, as the extension progresses downwards, by specialization of the already defined synsets, we noticed that fewer and fewer conflicts appeared.

Both WNBuilder and WNCorrect(1, 2) are implemented in Jscript and Perl, run under IE 6.0 or higher and are freely available on demand.

5. Current Status of the Romanian Wordnet

The quantitative data pertaining to the Romanian wordnet[8] are summarized in the tables below.

Table 3 shows the number of validated synsets for each part of speech.

[8] At the time of this writing, May 17^{th}, 2004.

Table 3. POS Distribution of the Synsets

Noun synsets	Verb synsets	Adj.synsets	Adv. synsets	Total
10 725	4 164	844	833	16 566

Table 4. Internal relations used in the Romanian wordnet

hypernym	14 867	category_domain	579
near_antonym	1 576	also_see	394
holo_part	1 005	subevent	169
similar_to	896	holo_portion	107
verb_group	980	causes	122
holo_member	779	be_in_state	546

Table 5. A comparison between Romanian wordnet and PWN 2.0

Language	Synsets	Token literals	Type literals	Average synset length	Average senses/lit
Romanian	16 566	29 130	17 538	1.75	1.66
English	115 424	203 147	145 627	1.76	1.39

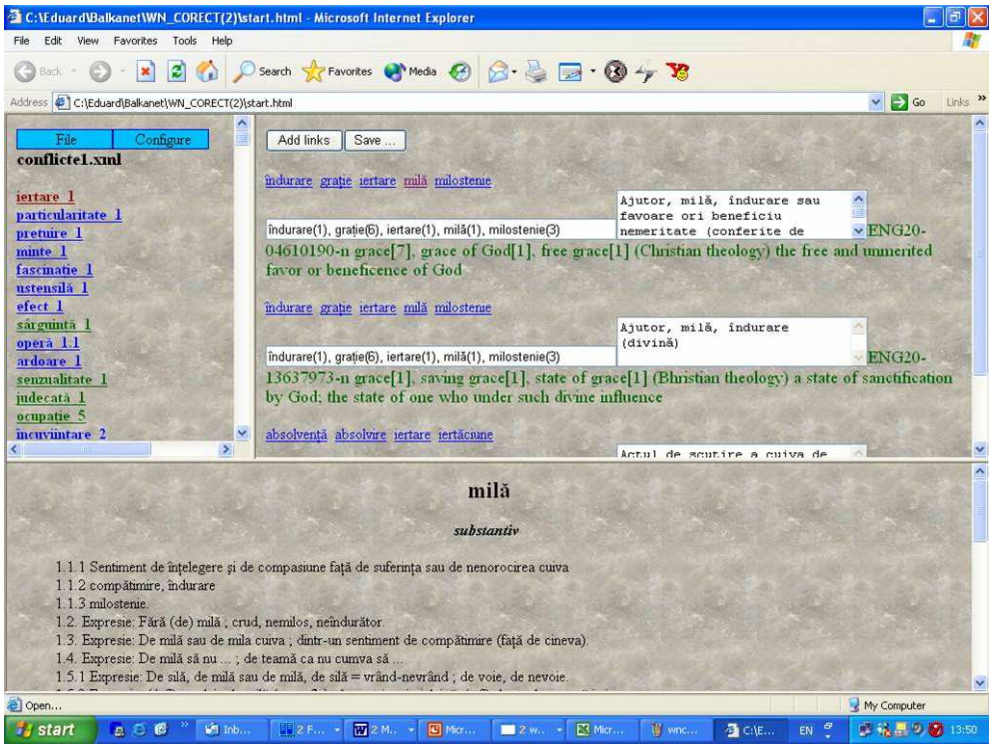

Fig. 8. The WNCorrect graphical interface.

Table 4 lists the internal relations used in our wordnet. Most relations also have a corresponding reverse relation, but these were not counted in the table above.

The comparison shown in Table 5 reveals similar average synset lengths in the two wordnets but a higher ambiguity degree per literal in Romanian. A simple explanation might be that on the lowest hierarchical levels in PWN there are many specialized terms which are more often than not unambiguous. Most of the ILI concepts BalkaNet wordnets implemented (thus, Romanian too) corresponds to upper levels PWN synsets. It is very likely that further extensions of our wordnet (downwards expansion of the hierarchies) will decrease the average ambiguity of the literals.

6. Conclusions

We presented a methodology and the associated software for the development of the ILI-based aligned Romanian wordnet. We believe this approach is general enough to be applicable to a large number of languages. We envisage continuing the extension of the Romanian lexical ontology after the project ends. We will continue to observe the conceptual density criterion but also we will pay attention to the lexical density criterion as well. Several applications for Romanian language heavily relying on the quality of the wordnet described here are under development. The word sense disambiguation system, based on aligned wordnets [11], [5] is just one of them.

Acknowledgement. The work reported here was carried with within the European project BalkaNet no. IST-2000 29388 and support from the Romanian Ministry of Education and Research under the CORINT programme.

References

[1] BILGIN, O., ÇETINOĞLU, Ö., OFLAZER, K., *Morpho-semantic Relations in and across Wordnets*, in *Proceedings of the Global WordNet Conference*, 60–66, Brno, 2004.

[2] BLOKSMA, L., DIEZ-ORZAS, P., VOSSEN, P., *The User Requirements and Functional Specification of the EuroWordNet project*, EWN-deliverable D.001, LE-4003, 1996.

[3] COTEANU, I., SECHE, L., SECHE, M. (Eds.), *Dicţionarul explicativ al limbii române*, Bucureşti, Univers Enciclopedic, 1996.

[4] HORÁK, A., SMRŽ, P., *Wordnet Browsing and Editing Tool*, in *Proceedings of the Global Wordnet Conference*, 136–141, Brno, 2004.

[5] ION, R., TUFIŞ, D., *Multilingual Word Sense Disambiguation using aligned Wordnets*, in this volume, 2004.

[6] JULLIARD, A., *The Frequency Dictionary of Romanian*, MIT Press, Massachusetts, 1965.

[7] MILLER, G.A., BECKWIDTH, R., FELLBAUM, C., GROSS, D., MILLER, K.J., *Introduction to WordNet: An On-Line Lexical Database*, International Journal of Lexicography, **3**, no. 4, winter 1990, 235–244.

[8] 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, Computers and the Humanities, **32**(2–3), 117–152, 1998.

[9] SECHE, L., SECHE, M., *Dicţionarul de sinonime al limbii române*, Univers Enciclopedic, Bucureşti, 1997.

[10] STAMOU, S., OFLAZER, K., PALA, K., CHRISTODOULAKIS, D., CRISTEA, D., TUFIŞ, D., KOEVA, S., TOTKOV, G., DUTOIT, D., GRIGORIADOU, M., BALKANET – *A Multilingual Semantic Network for the Balkan Languages*, in *Proceedings of the International WordNet Conference*, Mysore, India, January 21–25, 2002.

[11] TUFIŞ, D., ION, R., BARBU, E., MITITELU, V., *Cross-Lingual Validation of Multilingual Wordnets*, in *Proceedings of the Global WordNet Conference*, 332–340, Brno, 2004.

[12] TUFIŞ, D., CRISTEA, D., STAMOU, S., *BalkaNet: Aims, Methods, Results and Perspectives. A General Overview*, in this volume, 2004.

[13] 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*, 36–39, Edmonton, 2003.

[14] TUFIŞ, D., CRISTEA, D., *Methodological issues in building the Romanian Wordnet and consistency checks in BalkaNet*, in *Proceedings of the Workshop on Wordnets, LREC2002*, 35–41, Las Palmas, 2002.

[15] TUFIŞ, D., *A cheap and fast way to build useful translation lexicons*, in *Proceedings of the 19th International Conference on Computational Linguistics, COLING2002*, 1030–1036, Taipei, August 25–30, 2002.

[16] TUFIŞ, D., BARBU, A.M., *Computational bilingual lexicography: automatic extraction of translation dictionaries*, International Journal of Science and Technology of Information, **4**, no. 3, 312–325, 2001.

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

[18] TUFIŞ, D., ROTARIU, G., BARBU, A.M., *TEI-Encoding of a Core Explanatory Dictionary of Romanian*, in *Papers in Computational Lexicography*, 219–228, Kiefer, F., Pajzs J. (Eds.), Hungarian Academy of Sciences, 1999.

[19] VOSSEN, P. (Ed.), *A Multilingual Database with Lexical Networks*, Kluwer Academic Publishers, Dordrecht, 1998.