A Methodology and Associated Tools for Building Interlingual Wordnets

Dan Tufiş[1,2], Eduard Barbu[1]

[1]Research Institute for Artificial Intelligence, Bucharest; [2]University "A.I. Cuza", Iaşi
Calea „13 Septembrie", no. 13, PO 050711, Bucharest
{tufis, eduard}@racai.ro

Abstract

The paper reports on the ongoing effort towards the development of a Romanian wordnet aligned to the Princeton WordNet. The first part generically describes the methodology we used as well the language resources that supported our approach. In the second part we will describe the tools that implemented this methodology and a quantitative account for the content of the Romanian wordnet at the time of this writing. Both the methodology and the tools are language independent, provided the necessary supporting language resources are in the required format.

Introduction

The paper describes the methodology and the tools we developed for the purpose of building a Romanian wordnet. The work is carried out within the European project BalkaNet (IST-2000 29388) which started in September 2001 and will finish this year in August. The aim of the project is to build interlingually linked prototype wordnets for Bulgarian, Czech, Greek, Romanian, Serbian and Turkish As in the previous EuroWordNet (EWN) project, the interlingual index (ILI) of the BalkaNet is represented by the offsets in the databases of the Princeton Wordnet (PWN) synsets. This way, ILI, although unstructured, represents a one-to-one mapping to the PWN. Inspecting an ILI record shows the corresponding PWN synset plus its associated gloss. The ILI records represent the BalkaNet interlingual concepts. The lack of structure at the ILI level is motivated by the idea that besides meanings lexicalized in English, several language specific (or common to some languages) could be conceptually represented in the ILI. In order to ensure maximal cross-lingual lexical coverage, the consortium decided to implement a set of obligatory ILI concepts. They were selected starting with the set of 1310 concepts[1] which in EWN have been labelled as base concepts. The common set was extended up to 8516 concepts so that they should form a dense sub-network[2] of nominal and verbal concepts. In addition to the cluster criterion, for the selection of the descriptive adjectives we also imposed the condition that the nouns denoting the attributes they are value of (specified in PWN by the relation be-in state) are mapped over concepts in the common set. Two other selection criteria were followed:

a) the ILI records should have been implemented in most wordnets developed within the EuroWordNet project; this way we aimed at maximizing cross-lingual coverage not only for BalkaNet languages, but also for the languages represented in EuroWordNet.

b)the ILI records should represent concepts relevant for each language represented in the consortium; this assumed that each partner proposed a set of ILI concepts and those present in two or more proposed sets were also included into the common set.

The adverbs and non-descriptive adjectives were selected by each partner on own criteria.

Except for the restriction to have the common set of ILI records implemented in all the BalkaNet wordnets each team had the liberty to proceed to this goal (and beyond it) by using any available language resources and tools as well as developing new tools to facilitate reaching the common objective of the project. The methodologies adopted by each partner were different, mainly imposed by the language resources and personnel available.

The Language Resources and the Methodology for Building the Romanian Wordnet

The Romanian wordnet is developed by two teams of experienced computer scientists and linguists (one in Bucharest and the other one in Iaşi) that work in close collaboration. 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 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, The Dictionary of Antonyms as well as an in-house Romanian-English dictionary. We had some of these resources before-hand, some others had been turned into machine readable form for the purpose of this project.

The explanatory Dictionary of Romanian (EXPD) is a general dictionary of modern Romanian authored by the Linguistic Institute of the Romanian Academy and contains about 56.000 entries[3]. We further extended this dictionary so that our current version contains almost 70,000 entries. EXPD is XML heavily annotated

[1] The number is different from the one in EWN because they used as an interlingual index the 1.5 version of PWN, while in BalkaNet the interlingual index is build based on the version 2.0 of the PWN. Several synsets in PWN1.5 were split into finer grained synsets in the later versions (1.6, 1.7.1, 2.0)

[2] A dense sub-network of ILI concepts is represented by a set of ILI records so that for each concept in the set corresponding to a certain synset SYN_i in PWN, all the concepts corresponding the hyperonyms of SYN_i upwards the top level of the hierarchy are also included into the common set.

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

according to the encoding schema developed in the previous CONCEDE project. For the needs of the WNBuilder system, a much simpler encoding schema is necessary as exemplified below:

```
<ENTRY>
  <WORD>abandonat</WORD><POS>adjectiv</POS>
  <DEF>1. Care a fost părăsit.</DEF>
  <DEF>2. <USG>Despre copii nou – născuți</USG>
            Lepădat. </DEF>
 <ETYM>Vezi abandona </ETYM>
</ENTRY>
```

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), authored also by the Institute of Linguistics of the Romanian Academy was keyboarded, XML encoded and completed with more than 4000 new synonymy sets extracted from EXPD. The synonyms series in SYND contained both words used in modern language and archaisms and/or regionalisms (marked as such). 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 a comma.

The Romanian-English dictionary was automatically extracted from parallel corpora by our TREQ-ALL word aligner (Tufiş et al., 2003) and further on hand validated. It contains various statistical information, but in the simplified version as required by the wordnet development tool the bilingual lexicon should be a text file, containing on each line a translation equivalence pair (Source Target) and the common part of speech. If the source and target words have different part of speech, they are followed by the POSes of source and target words.

Besides these specific language resources we also used the XML format of the PWN.

All the above mentioned resources have been incorporated into a user-friendly system, called WnBuilder, which allows for cooperative work of a large number of lexicographers. Each lexicographer was responsible for implementing a distinct subset of the commonly agreed set of ILI-records. The lexicographer's subsets were computed so that the ILI-records in each subset corresponded to a dense sub-network in PWN. In the next section we will briefly discuss the basic functionality of the WNBuilder and the aggregation of the results of individual lexicographer's work.

The set of concepts we decided to implement in the Romanian Wordnet by the end of the project is a super-set of the commonly agreed 8515 ILI records. Following is a brief description of our approach aimed at ensuring an as wide coverage as possible for Romanian text processing. We started a series of quantitative analysis on a large corpus of journalistic texts, plus a few novels, collected from the web. The corpus (containing more than 100 million words) was automatically tagged, lemmatized and the content words of interest (nouns, verbs, adjectives and adverbs) were counted and sorted according to their frequency. We extracted this way, a list of more than 30,000 Romanian lemmas. Based on the frequency in the running texts, this list was divided into three parts, 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 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 net, the size of the data is not anymore a significant issue, but the relevance remains a systematic complain. The exact definition of what representative texts should be included into a corpus for quantitative data analysis is a long-standing debate and we won't get into this. Considering that our data consisted, almost exclusively, of journalistic texts, the relevance issue could certainly be raised. The Frequency Dictionary of Romanian–FDR [Julliard, 1965] published long time ago, based on a balanced corpus of 500,000 words of Romanian literature, legal texts, poetry and journalism contains a list of most frequent 5,000 lemmas. In spite of being quite contested, it is still used by many Romanian linguists as a reference. The comparison we made revealed that all of the 5000 words in FDR were also in our list, although not with the same frequency ranges. As frequency in running texts is a disputable criterion for deciding what would be words to be encoded into a core dictionary/thesaurus/ontology we considered that this criterion should be complemented with others, less controversial in the world of traditional lexicography.

Among the criteria one could find pleas for, we opted for two that we could easily turned into operational selectors. The one is the number of senses a headword would have in a reference dictionary. The second one is the number of word definitions that use the headword in case.

Considering only the first two frequency ranges described above (the first most 20,000 words in the journalistic corpus) we extracted from our Explanatory dictionary a list of more than 8000 nouns and nominal compounds (accounting for more almost 35,000 senses) so that the definitional productiveness DP (the number of sense definitions a noun participates in) was at least 3. 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 and nominal compounds. The final common set of ILI (nominal) records represent a strict subset of the set we computed as above.

The conceptual density criterion, adopted for ensuring cross-lingual coverage, more often than not requires the implementation of only some of the senses of the literals represented in our wordnet. For instance for the word *acţiune,* although EXPD glosses over 13 senses, in the current version of the Romanian wordnet are implemented only seven (the concepts they stand for were also implemented in all the BalkaNet wordnets). Ensuring the implementation in a given wordnet of all senses described in a reference dictionary for the language concerned is what we call the *lexicographic density*. This property is obviously language dependent both by the different lexicalizations of the concepts represented in the interlingual index and by the explanatory dictionary taken as reference. The lexicographic density issue was outside the scope of the BalkaNet project and it would be dealt with by each partner at a later stage.

For the task of choosing the adjectives and adverbs we used a parallel English–Romanian corpus consisting of

George Orwell's *1984*, Romania's Constitution and one year issues of the daily newspaper "Evenimentul Zilei" (aprox. 900,000 tokens per language). Part of this corpus (the "1984" novel) was used for the validation of the monolingual wordnets alignment to the PWN. The corpus is sentence aligned, lemmatized and POS tagged. Selection of the adjectives and the adverbs was done in the following three steps:

1. We computed the frequency of 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 each distinct synset.
4. The ILI records corresponding to the highest scored synsets, (900 adjectival and 800 adverbial synsets) were selected for our wordnet.

For the process of the proper building of the 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 numbers according to the EXPD numbering 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 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 targeted concepts; the sense number of a Romanian literal which would fit a target concept was not listed in EXPD (although the lexicographers felt it should have been); some sense definitions in EXPD were too coarse grained and had to be refined, etc. Concerning the sense labeling one should note that one general criticism of PWN 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 labeling notation, we kept it in our wordnet with the same interpretation.

After implementing the 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 (Tufiş & Cristea, 2002) it is conjectured the *Hierarchy Preservation Principle* which is the basic motivation for automating the import of the hierarchical and part/whole structures from PWN into our wordnet. Most of other semantic relations from PWN were found to be applicable in the Romanian 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).

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:

- Synsets 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. If the source synsets $S_{1SOURCE}$ and $S_{2SOURCE}$ are linked by importable (sequence of) relations R^+ and if the $S_{1TARGET}$ and $S_{2TARGET}$ are the correspondingly aligned synsets in the target wordnet, than they will be linked by the relation R. If in the source wordnet there are intervening synsets between $S_{1TARGET}$ and $S_{2TARGET}$ then, R is imported between the corresponding target synsets only if it is declared as a transitive (non-limited number of compositions) or partially transitive (a user-specified maximum number of compositions) relation. The typical partial transitive relation is meronymy.
- 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 pairs of Source/Target languages (provided the required language resources are available) we will exemplify for the English/Romanian languages. The graphical interface of WNBuilder has four frames that we will denote with: UL (upper left) frame, UR (upper right) frame, LL (lower left) frame, LR (lower right) frame. In the UL frame it is loaded the list of ILI codes that are in the lexicographer's responsibility. Clicking any ILI code would show in the UR frame:

- the English synset together with its associated gloss which is mapped onto the respective ILI record.
- a list of translation equivalents for the words in the English synset. The list of translation equivalents are taken from the bilingual dictionary. The lexicographer has the possibility to add new translation equivalents. By selecting (clicking) one translation equivalent in this list, the interface will display the following information:
  1. the definitions of the selected translation (in the LL frame; they are extracted on the fly from EXPD).
  2. 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.

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:

a) which are the best translations for the literals in the selected English synset;
b) which of the synonymic sets fits best 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;
c) which of the definitions (if different) of the translation and its synonyms fits best the English gloss;

d) 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 completing the local synset together with the definition and specifying the interligual relation the lexicographer will save his/her work. When saving the work one of the validation functions of WnBuilder comes into action, which checks the well-formedness of the synsets. The interface will signal the lexicographer all the errors he/she made during his/her assignment.

In the end, with most of the detected errors corrected, the lexicographer will export his/her work (name stamped) in VisDic[4] compatible format (Horak & Smrz, 2004). If errors are still present in the generated semantic subnetwork, they are recorded into a separate file for a subsequent correction.

The errors most difficult to correct are those arising due to different granularities among the PWN and the reference dictionary (in our case EXPD). With distributed work among different lexicographers the error chance due to the different resource granularity is further increased. For solving this very problem we developed another user-friendly interface called WnCorrect which allows the lexicographer to correct these problems in a focused way. Both WnBuilder and WnCorrect are implemented in Jscript and Perl runs under IE 6.0 or higher and are freely available on demand.

Current status of the Romanian Wordnet

The quantitative data pertaining to the Romanian wordnet[5] are summarized in the tables below.

Noun synsets	Verb synsets	Adjective synsets	Adverb synsets	Total
10725	2930	844	801	15300

Table 1: POS Distribution of the Synsets

hypernym	13681	category_domain	515
near_antonym	1776	also_see	333
holo_part	1005	subevent	139
similar_to	896	holo_portion	107
verb_group	888	causes	106
holo_member	779	derived	28
be_in_state	546		

Table 2: Internal relations

Table 1 shows the number of validated synsets for each part of speech. Other 1436 verbal synsets are implemented but they still contain conflicts and mapping errors and therefore were not included in these statistics. The Table 2 lists the internal relations used in our wordnet. Most relations also have a corresponding reverse relation, but these were not counted in the table 2.

The comparison shown in Table 3 reveals an average longer synset in Romanian wordnet as compared to PWN2.0 and also a higher ambiguity degree per literal. However, there is a simple explanation for these disparities: as one descends in the PWN hierarchies, the synsets get shorter and the literals less ambiguous.

Language	Synsets	Token lit.	Type lit.	Avg. synset length	Avg. senses/lit
Romanian	15300	27694	16887	1.81	1.64
English	115424	203147	145627	1.76	1.39

Table 3: Comparison between ROWN and PWN2.0

On the lowest hierarchical levels in PWN there are also specialized terms which are more often than not unambiguous. Most of the ILI concepts BalkaNet wordnets implemented (thus, Romanian too) correspond to upper levels PWN synsets. It is very likely that further extensions of our wordnet (downwards expansion of the hierarchies) will decrease the two figures discussed above.

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.

Acknowledgements

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

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

Coteanu, I., Seche, L., Seche, M. (Eds.). (1996) Dicționarul Explicativ al Limbii Române, Bucureşti: Univers Enciclopedic.

Horak, A., Smrz, P. (2004). Wordnet Browsing and Editing Tool. In Proceedings of the Global Wordnet Conference (pp. 136--141), Brno.

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

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

Seche L., Seche M. (1997). Dicționarul de sinonime al limbii române. Bucureşti: Univers Enciclopedic.

Tufiş, D., Barbu, A.M., Ion, R. (2003): A word-alignment system with limited language resources. In Proceedings of the NAACL 2003 Workshop on Building and Using Parallel Texts; (pp. 36--39), Edmonton.

Tufiş, D., Cristea, D. (2002). Methodological issues in building the Romanian Wordnet and consistency checks in Balkanet. In Proceedings of the Workshop on Wordnets, LREC2002 (pp. 35--41), Las Palmas.

Tufiş, D., Ion, R., Barbu, E., Barbu, V. (2004). Cross-Lingual Validation of Multilingual Wordnets. In Proceedings of the Global Wordnet Conference (pp. 332--340), Brno.

[4] VisDic is the standard visualization and maintenance system for the BalkaNet multilingual wordnets.

[5] At the time of this writing, March 1st, 2004.