Methodological issues in building the Romanian Wordnet and consistency checks in Balkanet

Dan Tufiş *, Dan Cristea♥

* RACAI-Romanian Academy
13, "13 Septembrie", Bucharest 5, Romania
tufis@racai.ro
♥University A.I. Cuza
16, Berthelot, Iaşi 6600, Romania
dcristea@infoiasi.ro

Abstract

The requirements in building a multilingual ontology of the EuroWordNet kind are frequently conflicting and if not considered in the first stages of the project, later harmonizing might be extremely difficult if possible at all. To ensure as early as possible usability, the incrementally developed lexical stock of each individual wordnet, should cover the most frequent vocabulary of the language. On the other hand, given that this is a multilingual lexical resource, special care should be addressed to the compatibility problems. Specifically, there are two main compatibility issues to be considered: there should be a cross-language conceptual coverage, meaning that each monolingual lexicon should globally deal with the same conceptual areas or domains and the interpretation of the defined relations should be the same in any monolingual ontology considered by the multilingual harmonized ontology. This is why, drawing as much as possible from the EuroWordNet lessons, we decided to address these issues at the very beginning phase of the BalkaNet project.

1. Introduction

BalkaNet [1] (Stamou et al, 2002) is an EC funded project (IST-2000-29388) that aims to develop in accordance with EuroWordNet philosophy a core multilingual resource for the following Balkan languages: Greek, Turkish, Romanian, Bulgarian, Czech and Serbian. As in EuroWordNet, the monolingual lexical ontologies are projected onto an interlingual set of concepts (ILI), the correspondences being established by means of complex equivalence relations (eq-synonymy, eq-near-synonymy, eq-has-hyperonym, eq-has-hypernym etc).

The requirements in building a multilingual ontology of the EuroWordNet kind are frequently conflicting (Rodriguez et al, 1998) and if not considered in the first stages of the project, later harmonizing might be extremely difficult if possible at all. To ensure as early as possible usability, the incrementally developed lexical stock of each individual wordnet, should cover the most frequent vocabulary of the language. On the other hand, given that this is a multilingual lexical resource, special care should be addressed to the compatibility problems. Specifically, there are two main compatibility issues to be considered: there should be a cross-language conceptual coverage, meaning that each monolingual lexicon should globally deal with the same conceptual areas or domains and the interpretation of the defined relations should be the same in any monolingual ontology considered by the multilingual harmonized ontology. This is why, drawing as much as possible from the EuroWordNet lessons, we decided to address these issues at the very beginning phase of the BalkaNet project.

The first part of the paper will address the approach we took for the selection of the initial lexical stock to be included into the Romanian core wordnet so that to observe multilingual design criteria and cross-language compatibility issues. The synsets (in two or more languages) that are mapped onto the same ILI concept are implicitly semantically linked. The nature of these cross-lingual semantic links, which we call *translational links*, depends on the links between the ILI concept and the synsets in the monolingual wordnets. One way to check consistency of the ILI projection of the individual wordnets is comparing the translation links with the translation equivalents licensed by a parallel corpus. This issue will be discussed in the second part of the paper.

2. An overview of the language resources

The Romanian wordnet started, as in the case of other languages in this project, from scratch. However, in order to ease the work and make the process as reliable as possible we built on various valuable language resources and several tools we developed for their exploitation. In the following there is a brief account of these building blocks, each of them being largely described elsewhere.

2.1. Corpora

Within the Multext-East and TELRI European projects (Erjavec et al. 1997), (Dimitrova et al., 1998), (Tufiş, Bruda, 1997), (Tufiş et al. 1997, 1998, 1999) there were created one 7-language heavily annotated parallel corpus based on Orwell's famous novel "1984" and one 25-language heavily annotated parallel corpus based on Plato's "The Republic". The annotation initially used was TEI compliant, but it was later on converted into CES (Ide, 1998). These are two relatively small corpora (about 110,000 tokens in each language) but given the accuracy of tagging and interlingual sentence alignment (hand validated) they were extremely useful for various applications ranging from building language models for morpho-syntactic tagging (Tufiş, 1999) and document classification (Tufiş et al., 2000) to automatic sense

[1] Further information can be obtained from the project's web site http://dblab.upatras.gr

discrimination (Erjavec et al., 2001). Besides the multilingual corpora we constructed two other much larger monolingual corpora: a literary corpus based on various novels (containing about 1,500,000 tokens) and a journalistic corpus (containing more than 100,000,000 tokens). Both corpora were automatically tokenized, tagged and lemmatized.

2.2. Lexicons and dictionaries

One delivery of the Multext-East project was a large wordform lexicon (more than 450,000 entries) containing triples <wordform, lemma, morpho-syntactic_code>. The encoding used in this lexicon is compliant with the Eagles recommendations for morpho-syntactic annotation and is largely documented in (Tufiş et al. 1997).

The reference dictionary we used for our analysis is The Explanatory Dictionary of Romanian (DEX,1996), work of the Romanian Academy Institute of Linguistics. This most authoritative lexicographic source for contemporary Romanian was partially digitized and converted into a lexical database (XML encoded) by RACAI under the European Project CONCEDE (Tufiş et al.1999). This core XML-dictionary has been extended to the full content of the printed dictionary by a follow-up project funded by Romanian Academy.

Another extremely useful lexical resource we relied on was the Romanian Dictionary of Synonyms-RDS (Seche, Seche 1997), which was transposed into electronic form by the NLP group at the University A.I. Cuza din Iaşi. The electronic form of RDS has been converted into an XML format so that the same query interface we developed for DEX works also with RDS.

From the multilingual parallel corpora mentioned before and using our translation equivalents extraction program (Tufiş, Barbu 2000, 2001a, 2001b) we constructed a bilingual Romanian English dictionary (also XML-encoded). This bilingual lexicon has been hand validated and extended with new entries from several public domain sources.

Finally, an extremely valuable resource was the ILI of the EuroWordNet, exported in XML format by means of the VisDic editor produced by the Masaryk University of Brno (Pavelek and Pala, 2002).

All these resources have been integrated by means of a series of tools developed for the purpose of the BALKANET project. They are user-friendly and allow for editing and mapping the Romanian synonymy series in RDS to the sense definitions in DEX and ILI records from EuroWordNet. The output of these tools is further subject to primary local consistency checks (such as detecting word sense appearing in more than one synset) and generated as an XML-encoded file appropriate for import in VisDic. We will provide a brief overview of these tools in Section 5.

3. Lexical stock selection

In order to ensure practical utility for the core wordnets to be delivered by the BALKANET project and to facilitate further extensions towards as large as possible coverage for the languages concerned, the project consortium decided to start the development process with a common set of concepts likely to be lexicalized in all the project languages. This special set of concepts, called *Base Concept* Set, was selected from the EuroWordNet interlingual index for reasons convincingly argued in (Vossen, 1998). The Base Concept Set contains 1310 concepts, each of them being attached a gloss and a Top Ontology Description (see Vossen, 1998). All project partners developed in a harmonized way the synsets in their languages corresponding to the Base Concepts. After this step, the monolingual wordnets will be further developed in a top-down approach starting with the synsets already mapped onto the Base Concepts.

Let us give a few definitions for some notions that will be used in the following.

When we place ourselves in a monolingual environment we speak about *senses, meanings* and *synsets.* A word has one or more *senses.* A sense refers to one *meaning.* In EuroWordNet the senses of a word are numbered according to their frequency and a sense of a lemma is denoted by appending the sense number to orthographic form of the lemma in case. A set of such numbered senses (eg. action2 activity1 activiteness1) referring to the same meaning is called a synset, which itself stands as a denotation of the common meaning of the senses in the synset. A meaning has a gloss that obviously applies for all senses in a corresponding synset.

When we want to abstract away from one language, we speak about the *concepts* referred to by the *word meanings*. So, we may speak about concepts with or without the reference to a specific language. Therefore, in trying to establish cross-lingual dependencies, via an interlingual index, it is convenient to refer to the entities used for this purpose as *concepts.* A concept is a language independent cognitive construct, which in EWN is always lexicalized at least in one language. A concept is further refined in terms of basic semantic distinctions (semantic features, sometimes referred to as semantic fields) so that one could speak about concept clustering along the basic semantic features.

According to these definitions we will use the term *Base Meaning* to refer to a basic (language specific) meaning in terms of which other word meanings can be defined and *which is directly mapped on a Base Concept*.

In EuroWordNet, and thus in BALKANET, ILI is defined as an unstructured collection of concepts represented by records of the form (<ILI-index> <ontological description> <gloss> {<domain>}). The initial ILI has been constructed from Wordnet1.5 and thus the gloss of each concept has been imported directly from the English synset referring to the meaning conceptualized in ILI.

According to the aims of the project regarding the interlingual coverage, language representativity, maximum usage of the core wordnet and scalability we started a series of quantitative analysis on a very large corpus made of several novels and a collection of journalistic texts, collected from the web. The corpus (containing more than 100 million words) was automatically tagged, lemmatized and the content words of interest (common 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).

In deciding which is the most important subset of a lexical stock for a language, the frequency in running texts

is considered by many lexicographers to be a very subjective criterion. Among the strongest arguments they would come with is the volume and representativity of the texts included into the corpus subject to the quantitative analysis. With more and more texts available on the net, the size of the data is not anymore a significant issue, but the representativity 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 entirely, of journalistic texts, the representativity issue could certainly be raised. The Frequency Dictionary of Romanian Words– FDRW (Julliand et all., 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 most of the 5000 words in FDRW were also in our list, although not with the same frequency ranges.

As frequency in running texts is a disputable criterion for deciding what words should 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 turn 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. A third criterion, not considered yet, might be the number of derivatives of a given headword (this last criterion is preferred by most Romanian etymologists).

In this phase of the BALKANET project we concentrated our attention to the Romanian nouns and the experimental data reported below refers to nouns. Since the technical procedures do not depend on the specific part of speech, the same would apply for verbs, adjectives and adverbs.

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 more than 8000 entries for nouns and nominal compounds (accounting for almost 35,000 senses) so that the definitional productiveness DP (the number of sense definitions a noun participates in) was at least 3. The list was sorted according to the definitional productivity.

Noun	**Definitional productivity**	**Number of definitions**	**FRECV$_{range}$**
acţiune	2279	13	I
persoană	1979	9	I
parte	1882	94	I
formă	1286	21	I
obiect	1204	16	I
fapt	1044	11	I
apă	743	29	I
• • •	• • •	• • •	• • •
rasism	3	1	II

Table 1: scoring the headword candidates

For all these nouns we extracted EN translations from our translation equivalence dictionary. The procedures for automatic extraction of translation equivalents from parallel corpora as well as the sense discrimination procedure are largely described in (Tufiş&Barbu, 2001a,b), (Erjavec et al, 2001). As the translation equivalents found by our extractor are limited by the available parallel corpora we have, provisions were made for automatic updating of the Ro-En dictionary with web resources.

All pairs containing an English word (or a synonym of it) in the English synsets corresponding to the base concepts were also associated with the corresponding top-ontology description. Practically for all English words corresponding to the base concepts there were found translations in our translation lexicon and these translations appeared in the upper top of our 8000-noun list. Those few EN nouns not translated in our lexicon were given manual translations. Because our translation equivalence lexicon is based on sense equivalence in context, transferring the ontological description from one EN word to its equivalent translation was considered to be a legitimate option. Thus, at the end of this step we collected a list of Romanian nouns associated with one or more English translations out of which at least one was present in the base concept list. Each such an association was further enriched with additional information extracted from other resources:

a) the RO word was attached with all its definitions extracted from the Explanatory Dictionary of Romanian;

b) the EN word was attached with its entry in the WordNet1.5

The Romanian Dictionary of Synonyms (RDS), digitized and encoded as an ACCES database by University A.I. Cuza of Iaşi, was used to extract the synonymy series for the selected RO words. In RDS some members of the synonymy series are provided with usage information (old, regionalism, specific area of usage, domain, etc). Preliminary discussions lead to the idea to eliminate all the words marked as such (based on the assumption that we would like to construct a lexical stock for general use in contemporary Romanian). However, if later on this filtered out words (together with their usage information) would be necessary, their recovery was ensured. The synonymy series were taken as possible Romanian synsets and added to the RO-EN associations described above.

We have thus assembled the basic linguistic material that the lexicographer should use in making the decisions (linking) necessary for building the noun subset of the core Romanian wordnet. All this information is currently available in a java-based editor, showing in different frames, the following information (see figure 1):

- the list of the base concepts (upper-left frame), identified by the ILI record and an English word in the synset mapped on this concept (ex. *life_3_03941565-n*)
- the synset (life_3 living_1), its gloss and top-ontology description, possible translations and association boxes (right-upper frame)
- the numbered sense definitions from the Explanatory Dictionary of Romanian for the selected translation (left-lower frame);
- synonyms of the selected Romanian translation word (right-lower frame)

- pop-up menus for selecting the relevant sense numbers and the equivalence relation to the ILI concept.

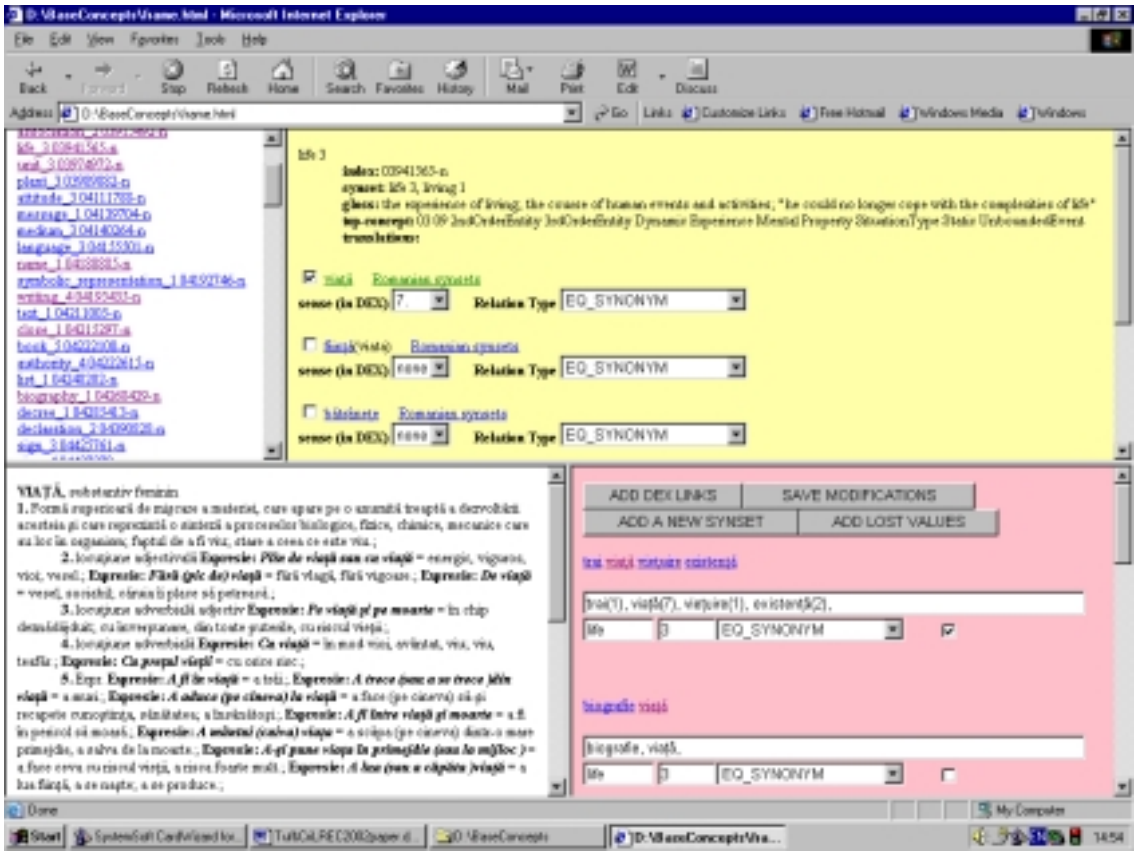

Figure 1: The editor for building synsets for the base meanings

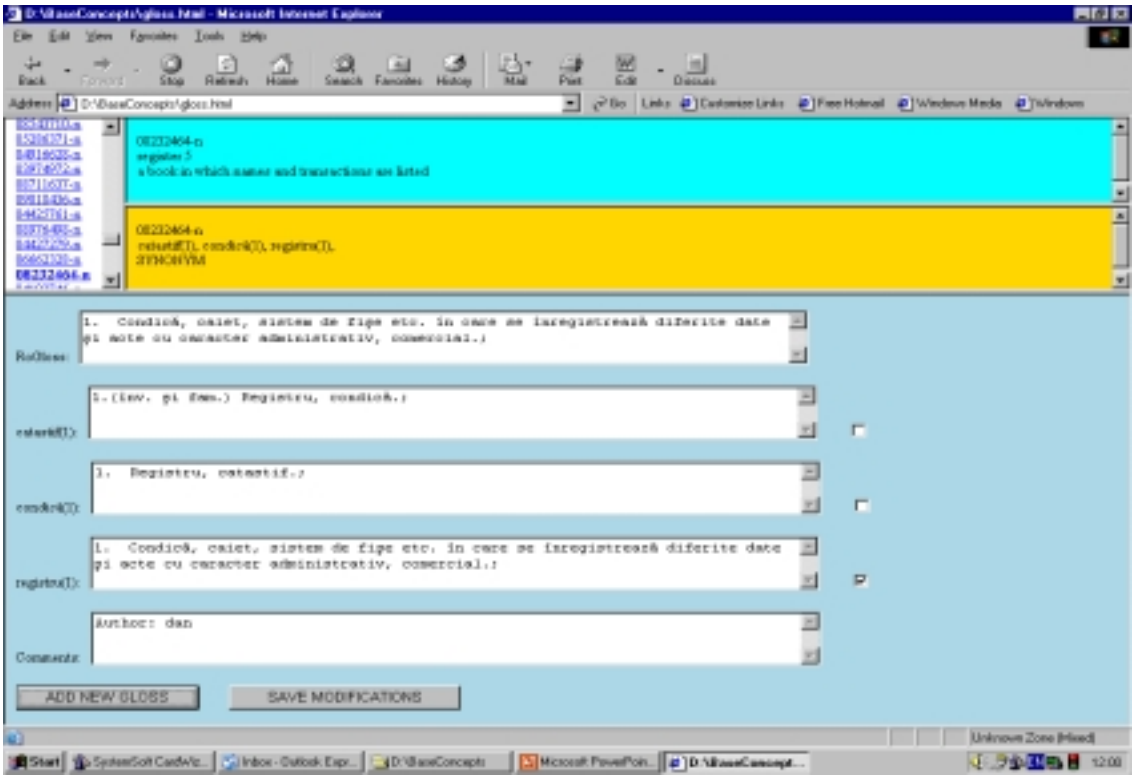

Figure 2: The editor for gloss assignment

The editor has been instantiated into 10 differently populated copies, each containing a different set of base concepts. Each incarnation of the editor has been given to a different expert who was in charge of building his/her set of Romanian synsets and map them onto the appropriate base concepts. When this building phase was finished we performed a few simple error-checking such as:

- all literals appearing in a synset should have attached a sense number
- no sense (literal and sense number) should appear in two or more synsets
- each synset should have an equivalence relation to a unique base concept.

Once the synsets were constructed and mapped onto base concepts, the second phase was to add a Romanian gloss to each Romanian synset. In the vast majority of cases, the definitions extracted from DEX corresponding to the senses in a synset were different in wording so, the lexicographers had to chose the best definition, closest to the definition of the corresponding base concept. The Figure 2 shows that the base concept 08232464-n corresponding to the 5th sense of the English word *register* (a book in which names and transactions are listed) corresponds in Romanian to the synset (catastif_1 condică_1 registru_1). The selected senses for the three Romanian words have in DEX different definitions. By checking the box to the right of the third definition (lower frame in Figure 2) the lexicographer decided that the definition given to *registru_1* is the one to be attached to the synset.

It is worth mentioning that during the gloss assignment phase it became apparent that several synsets were not correct, requiring modifications. In some cases, the Romanian Explanatory Dictionary includes under the same definition two senses that are differentiated in ILI as two distinct concepts. In such cases, the general strategy was to split the Romanian definition and attach the relevant part as a gloss.

4. A proposal for cross-lingual validation of the ILI mapping

As we said before, one of the main objectives of the BALKANET project (which adopted a merge model approach) is to ensure as much as possible overlap between the concepts lexicalized in the concerned languages. A significant overlap may be hampered either by conceptually different lexical stocks for the different languages or by inconsistent projection of the monolingual concepts onto the ILI concepts. In order to ensure conceptual similarity for the lexical stocks across various languages, the development of the monolingual ontologies started in two different, but convergent ways: the minimalist one was to provide direct translations of the EuroWordNet Base Concept Set; the second way (language-centric) was to produce a ranked list of most important (according to prescribed lexical criteria) words in each language and to include in the monolingual wordnets at least those words, the meanings of which would cover the Base Concept Set. Irrespective of the approach taken towards ensuring lexical stock similarity across languages, we had to consider means for automatic check of the correctness of the mapping of the monolingual synsets over the ILI concepts. To this end we will describe in some details a proposal for an automatic consistency checking.

Our approached is based on the notion of translation equivalence over bitexts, on bilingual lexicons automatically extracted from parallel corpora (Tufiş, Barbu, 2001 a,b) and on sense disambiguation (Erjavec et al., 2001).

The parallel corpus we used in our experiments is the "1984", based on Orwell's famous novel, developed in the MULTEXT-EAST project, further cleaned up in the TELRI and CONCEDE projects. The corpus contains professional translations of the original novel in 6 languages (Bulgarian, Czech, Estonian, Hungarian, Romanian and Slovene), all aligned at the sentence level to the English original. Each monolingual part of this 7-language parallel corpus is segmented, tagged and lemmatized and also carefully hand validated.

From the 6 (integral) bitexts (CEE language texts aligned to the EN original) there were extracted bilingual lexicons (XX-EN, with XX one of the six CEE languages) and furthermore a 7-languages lexicon with EN as a hub. By removing all the non 1-1 alignments in the bitexts and using the EN sentence Ids as anchors, a partial (about 92% of the whole text) 7-lingual 1-1 alignment (EN-BG-CZ-EE-HU-RO-SI) was computed. The 7-language aligned corpus allows for extracting any of the 21 possible (partial) bitexts. A number of 104 nouns appearing in the English part of the multilingual corpus (altogether 3316 instances were hand annotated and used as a gold standard for our sense clustering algorithm (Erjavec, 2001).

As BG, CZ and RO are languages of the BalkaNet project from the present data, our methodology could be used for checking the ILI-mapping consistency for any of the RO-EN, RO-CZ, RO-BG, EN-CZ, CZ-BG and BG-EN pairs of wordnets. In the current phase of the project we are able to consider only the interlingual mapping of the base concepts. Let us generically denote the language pairs subject to checking as XX-YY. The basic methodology is as follows:

1) From the XX-YY bitexts we extracted the XX-YY lexicon (http://www.racai.ro/~tufis/BilingualLexicons/AutomaticallyExtractedBilingualLexicons.html). The bilingual lexicon contains not only the translation pairs but also, for each entry the aligned sentences that licensed the translation equivalence relation. This lexicon is purged so that it contains only words that have (in the respective monolingual wordnets) at least one sense mapped on a base concept set. Put it otherwise, any pair (W_{XX} *translated as* W_{YY}) of the purged lexicon has the property that W_{XX} or W_{YY} or both have at least one sense in the language-specific base meaning set.

2) Let it be (W_{XX} W_{YY}) a translation equivalent. Let us denote with S_{WXX} the synsets in language XX containing the W_{XX} word (actually one sense of it) and S_{WYY} the synsets in language YY containing the W_{YY} word (actually one sense of it). Starting in the XX monolingual wordnet from the synsets in S_{WXX}, via ILI, one ends in the YY monolingual wordnet with the XX-synsets having translation links to YY-synsets. Let us call this set as S'_{WYY}. S_{WYY} and S'_{WYY} should have at least one synset in common. Please note that if the intersection of the two sets of synsets is non-empty, the described procedure ensures semantic tagging of the (W_{XX} W_{YY}) pair with one or more ILI-concept tags. If the intersection contains exactly one synset, its corresponding ILI record-number

could be used to semantically tag both W_{XX} and W_{YY}. With intersection containing more synsets, we still are able to reduce the semantic ambiguity of the considered words. In case the intersection is empty, we might have one of the following possible explanations:

2.1) (W_{XX} W_{YY}) is not a valid translation pair; by checking the sentences that licensed the extraction of this translation pair one could confirm or refute this possibility; please note that an error here might be due to the extraction algorithm or to a problematic human translation (for instance it is not uncommon that even professional translators would sometimes translate one word by a non-eq-synonym for various reasons like contextual semantic gaps or stylistic preferences)

2.2) (W_{XX} W_{YY}) is a valid translation pair and the two words share a meaning assigned to a concept which is not in the base concept set.

2.3) the interlingual mapping of the W_{XX} and W_{YY} is "wrong"; being "wrong" might be a real mapping error in the XX or YY language (or in both) or it might be motivated by a lexical gap in one of the languages concerned (or both); the lexicographer might have overcome the lexical gap by using a complex equivalence relation (not the eq-synonym); in the second case, one might get insights on possible concept clustering at the ILI level (creating so-called *soft-concepts*).

We claim that this procedure allows us to estimate both the cross-lingual coverage and the correctness of the interlingual mapping of the two considered monolingual wordnets. The procedure allows not only for estimation but also for pinpointing the incomplete or missing synsets as well as inconsistencies in mapping the synsets onto ILI concepts and gives hints on soft-concept clustering.

4.1. Condiments, spices, sauces and other ingredients

Let us consider the fragments of the Ro-Wordnet and WN1.5 shown in the Figure 3. The arrows represent hyponymy relations in the two wordnets. The gray heavy lines represent translational links between the synsets in the two languages, meaning that the respective synsets are mapped onto the same ILI concept. The heavy dashed line represents a translational link that is reported as wrong during the cross-validation of the two wordnets. The reason for this comes from the violation of what we called the *hierarchy preservation principle*. The inconsistency is signaled because in language RO the hierarchical relations (hyponym) between $^{M}mirodenie_{RO}$ H $^{M}condiment_{RO}$ as well as $^{M}ketchup_{RO}$ H $^{M}sos_{RO}$ are not verified in language EN by the equivalent pair meanings ($^{M}spice_{EN}$ $^{M}condiment_{EN}$) and ($^{M}ketchup_{EN}$ - $^{M}sauce_{EN}$)(in EN they are sisters). If the structuring in WN1.5 is taken to be the Truth, this example shows that *the hierarchy preservation principle* is not true. On the other hand, if it would be reasonable to consider that WN1.5 is amendable (for instance making $^{M}mustard_{EN}$ and $^{M}ketchup_{EN}$ direct hyponyms of $^{M}sauce_{EN}$) then the *hierarchy preservation principle* might be a very powerful consistency check.

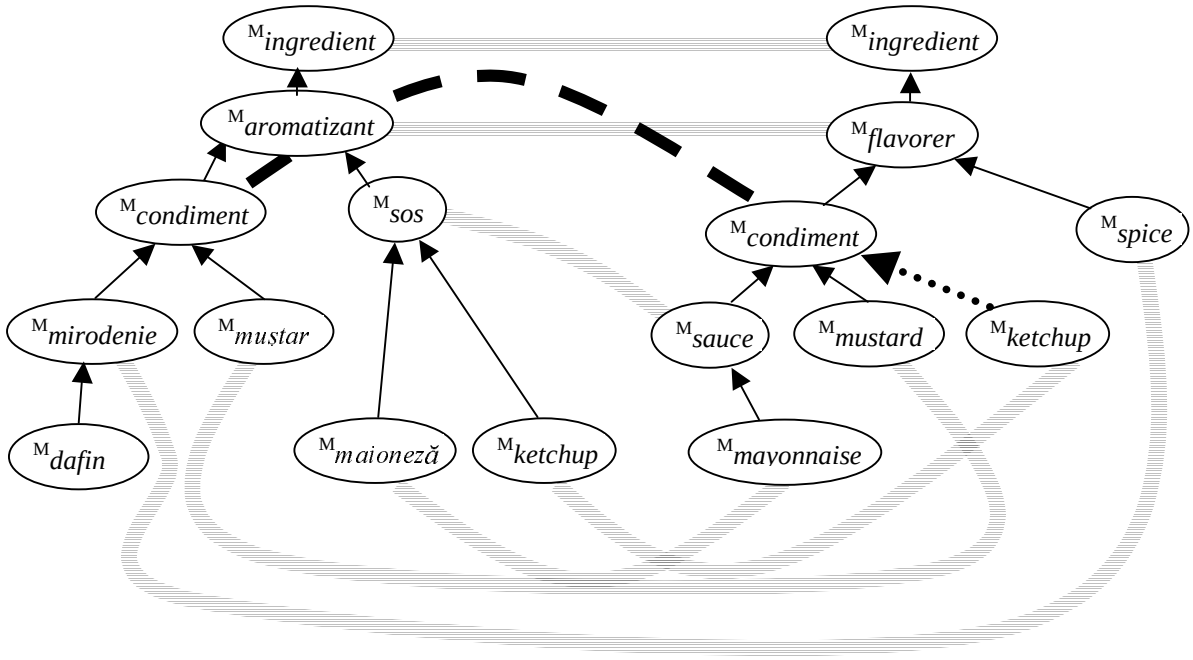

Figure 3: Translational links and consistency checks

5. Conclusions and further work

The approach on consistency checking based on translation equivalents in multilingual parallel corpora has some methodological similarity with (Resnik et al., 1999) on the multilingual corpus built up from many translations of the Bible. Speaking about useful sense distinctions (for machine translation for instance) Resnik (personal communication) identifies *strong sense distinctions* of one word in a source language as those that are lexicalized as

different words in the target languages. When some senses carried by a source word are found in a target word the distinction between them is called a *light sense* distinction. In the area of machine translation trying to disambiguate among light distinctions is not a very productive enterprise and therefore being able to identify, for a given pair of languages, which are the strong/light sense distinction might be extremely useful for machine translation. Our approach could be used to enhance the strong/light dichotomy with a third dimension: *fuzzy sense* distinction. This term is strongly related to that of *soft concept* used in EuroWordNet for clustering different ILI concepts that are lexicalized in two or more languages by words considered to be legitimate translations of one another.

In the next phase of the project, in order to extend the monolingual Romanian wordnet up to the level of the promised size, our strategy will be language-centric meaning that the new entries will be the top ranked words selected from our noun/verb/adjective/adverb lists sorted as described in the section 3.

6. References

Bloksma L., Diez-Orzas and Vossen P. (1996) The User Requirements and Functional Specification of the EuroWordNet-project *EWN-deliverable D.001*, LE-4003

DEX (1996). Coteanu, I., Seche, L., Seche, M. (coord.). Dicţionarul Explicativ al Limbii Române, Ediţia a II-a, *Univers Enciclopedic*, Bucureşti, 1996

Erjavec T., Ide N., Tufiş D.(1997) Encoding and Parallel Alignment of Linguistic Corpora in Six Central and Eastern European Languages" in Michael Levison (ed) *Proceedings of the Joint ACH/ALL Conference* Queen's University, Kingston, Ontario, June 1997 (also on http://www.qucis.queensu.ca/achallc97)

Erjavec T., Ide N., Tufiş, D.(2001) *Automatic Sense Tagging Using Parallel Corpora*, in Proceedings of the 6th Natural Language Processing Pacific Rim Symposium, Tokyo, Japan, 27-29 November, pp. 212-219, 2001

Ide, N. (1998) *Corpus Encoding Standard: SGML Guidelines for Encoding Linguistic Corpora* First International Language Resources and Evaluation Conference, Granada, Spain. See also http://www.cs.vassar.edu/CES/.

Julliand, A., Edwards P.M.H, Julliand I. (1965). The Frequency Dictionary of Rumanian Words. *Mouton & CO.*, London-The Hague-Paris, 1965

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

Pavelek T., Pala K. (2002) *VisDic : A new Tool for WordNet Editing* in Proceedings of the 1st International Wordnet Conference, Mysore, January 21-25, 2002

Resnik, P. (1999) Disambiguating Noun Groupings with Respect to WordNet Senses, in S. Armstrong, K. Church, P. Isabelle, S. Manzi, E. Tzoukermann and D. Yarowsky (eds.), *Natural Language Processing Using Very Large Corpora.* Kluwer Academic Publishers, 1999, pp. 77-98.

Resnik, P., Broman Olsen M., Diab M.(1999) The Bible as a Parallel Corpus: Annotating the `Book of 2000 Tongues', *Computers and the Humanities*, 33(1-2), pp. 129-153, 1999.

Resnik P., Yarowsky D. (2000) Distinguishing Systems and Distinguishing Senses: New Evaluation Methods for Word Sense Disambiguation, *Natural Language Engineering* 5(2), pp. 113-133.

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 Piek Vossen (ed.) *EuroWordNet: A Multilingual database with lexical semantic networks*, Computers and Humanities, Vol. 32, Nos 2-3, 1998

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

Stamou S., Oflazer K., Pala K., Christoudoulakis D., Cristea D., Tufiş D., Koeva S., Totkov G., Dutoit D., Grigoriadou M..(1997) BALKANET A Multilingual Semantic Network for the Balkan Languages, in *Proceedings of the International Wordnet Conference*, Mysore, India, 21-25 January 2002

Tufiş D., Şt. Bruda (1997) Structure Markup in CES and Preliminary Statistics on Romanian Translation of Plato's "Republica", *Proceedings of International Seminar on Encoding*, Ljubliana, February, 1997, also in *TELRI News*, nr. 5, May, 1997.

Tufiş, D. Tiered Tagging and Combined Classifiers In F. Jelinek, E. Nöth (eds) *Text, Speech and Dialogue*, Lecture Notes in Artificial Intelligence 1692, Springer, 1999

Tufiş D., Barbu A.M., Pătraşcu V., Rotariu G., Popescu C. (1997). Corpora and Corpus-Based Morpho-Lexical Processing, in Tufiş D., P. Andersen (eds.) *Recent Advances in Romanian Language Technology*, Editura Academiei, 1997.

Tufiş, D., Rotariu, G., Barbu, A.M. (1999) TEI-Encoding of a Core Explanatory Dictionary of Romanian. In Kiefer, F. and Pajzs J. (eds.) *Papers in Computational Lexicography*, Hungarian Academy of Sciences, 1999, pp. 219-228

Tufiş D., Popescu C., Roşu R.: Automatic classification of documents by random sampling in *Proceeding of the Romanian Academy*, Series A, vol 1, no. 2, p. 18-28, 2000

Tufiş, D. (2000). Blurring the distinction between machine readable dictionaries and lexical databases. R*esearch Report, RACAI-RR56*, 1999

Tufiş, D. (2001) Romanian wordnet of BALKANET: selecting the lexical stock. R*esearch Report, RACAI-RR68*, October 2001

Tufiş, D, Cristea D. (2001) *Methodological issues in selecting the candidate concepts to be included into the Romanian Wordnet.* Progress Report on BALKANET project. November 2001

Tufiş D., Barbu A.M.(2001a) *Computational Bilingual Lexicography: Automatic Extraction of Translation Dictionaries*, in International Journal on Science and Technology of Information, Romanian Academy, ISSN 1453-8245, Vol.4, No.3-4, 2001, pp.325-352

Tufiş D., Barbu A.M.(2001b) *Extracting multilingual lexicons from parallel corpora*, in Proceedings of the ACH-ALLC conference, New York, 12-17 June, 2001.

Vossen P. (ed.) (1998) "A Multilingual Database with Lexical Networks", Kluwer Academic Publishers, Dordrecht