BalkaNet Ref. No:IST-2000-29388

EUROPEAN COMMISSION
Directorate-General Information Society
Information Society Technologies: Content, Multimedia Tools and Markets,
Linguistic Applications

# Design and Development of a Multilingual Balkan Wordnet Balkanet, IST-2000-29388


## WP6: Improvement and Extension of the monolingual WordNets

## December 10, 2003

### Deliverable D6.1: Comparison and testing results

BalkaNet Ref. No:IST-2000-29388

# Methodology and Tools Adopted for the Evaluation and Correction of the Monolingual WordNets

## 1. Introduction

The main objective of this workpackage is to extend the individual core WordNets developed in the previous workpackage (WP5) after performing the evaluation of the accuracy of the implementation of the monolingual wordnets and validation of the interlingual linking. The two phases (monolingual evaluation and cross-lingual validation) require first, the detection of possible problems and, subsequently their solving.

The goal of building the monolingual wordnets in a concerted manner and with a high level of cross-lingual coverage raised several problems and challenges. We should mention that most of the work carried on within this project is based both on statistical techniques and on human introspection and subjectivity. As such, since none of these approaches is error-free, various kinds of errors (omissions, conflicts, processing errors, etc) percolated into the wordnets. Also, it is likely that some others will show up later on during exploitation in real applications. As the pioneering work at Princeton shows, a wordnet is a continuously changing and evolving resource; this is even more characteristic for a multilingual wordnet.

The consortium decided on a set of tests to be applied by each team to its own wordnet so that all the detected problems are solved before a cross-lingual evaluation was started.

During the subtask the results of which are reported in this document, the members of the consortium and user groups performed intensive evaluations and tests on their monolingual core wordnets and most of the problems were solved. Some specific errors couldn’t be solved and there were good reasons for the postponement of their resolution which the report explains (where the case).

Also, during this subtask a lot of effort was invested in preparing the cross-lingual validation based on parallel corpora. The partners prepared in the appropriate format (CES-ANA) the (the entire or partial) test monolingual corpus (the translations of Orwell's "1984"). Then, the monolingual corpora were sentence aligned (only 1-1 alignments were retained in order to ensure –via transitivity alignment– processability of any language pair). A set of words in the English original of the aligned parallel corpus was selected so that all their senses are represented by ILIs in the commonly agreed BCSs (1, 2, 3 or 4). An innovative word-aligner and sense disambiguation program (WSDtool) have been developed. During the next phase of this work-package, the results of the cross-lingual validation will be discussed and the necessary restructuring and extensions of the inter-linked wordnets fulfilled.

The extended and restructured WordNets will be the final monolingual WordNets to be incorporated into the BalkaNet multilingual lexical database.

# 2. The commonly agreed set of tests for the monolingual core wordnets and quantitative comparisons

One significant achievement of the consortium since the last report was moving from Princeton WordNet 1.7.1 to the most recent version WordNet 2.0. As the previous upgrade (from Princeton WordNet1.5 to Princeton WordNet1.7.1) this step assumed applying a set of mapping rules and in some cases, where the mapping was not deterministic, manual mapping.

Based on the consortium consultation we designed a set of formal general constraints which every wordnet was expected to observe. The constraints were implemented as a set of tests and each partner applied them and worked towards removing or correcting all the structural elements of their wordnets that did not observe the rules of well-formedness. A couple of other language specific restrictions have been proposed and implemented by some partners.

The first quantitative evaluation, namely the number of the synsets and their part-of-speech distribution as compared with the specifications in the Technical Annex, showed that the consortium achieved more than it was promised.

The quantitative comparisons among the well-formed wordnets were meant to give an overall evaluation of the cross-lingual coverage and to this end we computed intersections among the cross-linked synsets in all languages.

A better indication of the quality and compatibility will be given by comparing the consistency of the interlinked wordnets against a parallel corpus. The comparison of the WordNets will be based on the equivalence relations to the EuroWordNet ILI records and the translation equivalence relations as featured by the parallel corpus.

## *2.1 General tests for the well-formed wordnets*

1. XML well-formedness of the wordnets (compliant with the VISDIC format).
2. Literals and sense ids: this is probably one of the hardest issue so solve. The easy part is to ensure that all the literals in any synset are already assigned a sense identifier. Also is easy to check that no identical literals (irrespective of the sense labels)

belongs to the same synset. We do agree with the Belgrade team concerning the sense identifiers: we don't think that sense identifiers should be obligatory integers and also, we don't think that the senses implemented for a given word should be consecutive. At least for the small wordnets, under developments, as ours are. That is to say that these should not be regarded as errors. The single conceptual restriction is that the combination literal+sense identifier should be unique. Since our implemented wordnets were centered on a subset of senses in PWN it is unavoidable to have words in the target wordnets for which only some of the senses were considered.

3. IDs validation (the synsets should be labeled with valid unique IDs)
4. POS validation: the synsets should be tagged only with one of the 4 categories n, v, a, b)
5. Internal relations validation (no duplicates, relations belonging to the standard set of relations, no loops)
6. network density validation (no dangling synsets or relations);
    i. an existing synset which has no hyperonym should be mapped to an ILI that in PWN is a topmost synset (such as unique beginners for the noun hierarchy); otherwise is a dangling node;
    ii. an existing (binary) relation which misses either of the two synsets it is supposed to connect is considered a dangling relation iff the missing synset would correspond to an ILI in the commonly agreed set. Otherwise it is not and it should be deleted.
7. glosses validation (no empty definitions, spellchecking, definition in the own language)
8. senses validation (no literal with the same sense label should appear in more than one synset);

## *2.2 Quantitative cross-lingual comparisons among the wordnets*

9. Cross-lingual intersections of the synsets in BCS1, BCS2, BCS3 and BCS4 (optional)
10. The number of common relations for common ILI's.

    This test is meaningful especially for the approaches that assume the principle of hierarchy preservation (see section 4).

    Let $R_{REF}$ be the set of relations in PWN so that, any relation in $R_{REF}$ links synsets in BCS1+BCS2+BCS3 (+BCS4). Let $R_{XX}$ be the set of relations in XX-WN so that any relation in $R_{XX}$ links synsets in BCS1+BCS2+BCS3 (+BCS4). Than for each type of common relation $R^i$ (semantic relations) one could check the following:

    c1) compute $|R^i_{REF}|/|R^i_{XX}|$ (the ratio between the number of relations in the two sets);

    c2) If $R_{REF}$ is partitioned among the relations between noun synsets, verb synsets, adjective synsets and adverb synsets so that $R_{REF}=R_{REF}^N+R_{REF}^V+R_{REF}^A+R_{REF}^B$ and similarly $R_{XX}= R_{XX}^N+R_{XX}^V+R_{XX}^A+R_{XX}^B$

    Indicative figures are the ratios $|R_{REF}^N|/|R_{XX}^N|$, $|R_{REF}^V|/|R_{XX}^V|$, $|R_{REF}^A|/|R_{XX}^A|$, $|R_{REF}^B|/|R_{XX}^B|$;

In the subsequent sections of the third chapter are described the methodologies for Wordnet's validation, correction and/or extension adopted and followed for the last couple of months by each contractor. The last section of the chapter 3 summarizes the results of the tests and comparisons.

Chapter 4 presents the methodology that will be followed for cross-lingual validation based on a parallel corpus. The cross-lingual validation based on Orwell corpus is ready to start. In Brno, during the next consortium meeting (January 2004), we will demonstrate the tool and explain the functionality. The restructuring of wordnets (adding new synsets, adding new literals in the synsets already implemented, etc), will be supported by the WSDtool (and maybe some other tools developed at different sites). The restructuring and the final wordnets will be the topic of the D6.2 report, due in March 2004.

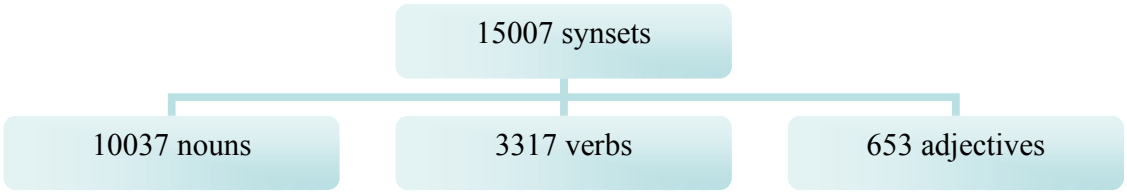
The last chapter provides a rough estimation of the workplan and an indicative timetable along with some general considerations for the forthcoming tasks.

# 3. Tests and results for the monolingual core wordnets

## *3.1 The Bulgarian wordnet*

**I. Current state of Bulgarian WordNet**

Currently the Bulgarian WordNet consists of 15 007 synsets, where 26 821 literals have been included (the ratio is 1,78). The distribution of synsets into parts of speech is as follows:




13 967 hyperonymy relations have been defined. The distribution of the remaining relations is as follows:

| | |
|---|---|
| Near_antonym – 1333 | Holo_member – 754 |
| Holo_part – 830 | Verb_group – 673 |
| Be_in_state – 383 | Derived – 234 |
| Also_see – 181 | Subevent – 142 |
| Cause – 102 | Holo_portion – 59 |
| Similar_to – 38 | Particle – 22 |

**II. Validating Bulgarian Wordnet**

Our approach to the validation of the Bulgarian Wordnet includes three separate steps of different degrees of complexity and significance which present different possibilities for automatic data correction – checking the syntax of the XML files in

which the data are organized, checking the completeness of the Bulgarian Wordnet, and checking for contradictions in the interpretation meanings of the synsets and the semantic relations between them.

The lowest level, which is also the easiest for processing and correction, is XML files syntax. In the following cases automatic checking as well as automatic data correction is possible:

- ✓ Some of the XML tags must always possess a value, and for others this is not compulsory: i.e. <USAGE>, <SNOTE>, <LNOTE>, <STAMP> and <ILR>. The empty tags of the second kind, being facultative, may be removed automatically.
- ✓ It is known that the literals in a given synset cannot be duplicated. Duplicated literals may also be removed automatically while keeping at least one of them.
- ✓ Currently, sense numbers are random in the Bulgarian Wordnet, they do not correspond to the arrangement of the meanings of polysemy words in an explanatory dictionary or to the frequency of usage of a certain meaning. For this purpose, another possible checking (and automatic reordering) is whether the <SENSE> tag is empty, whether it contains only numbers, as well as whether the sense numbers are consecutive and are not duplicated.

In other cases where automatic correction is possible, manual confirmation of replacements is necessary:

- ✓ <ID> tags may be checked whether they conform to the accepted standard for them - a certain number of digits and part of speech denotation, and if they do not conform the closest correspondences from English WordNet 2.0 are to be suggested. Due to the correspondence between meanings in the two languages, the automatic replacement should be avoided and decisions for connecting the synset with the correct <ID> must be taken manually.
- ✓ To empty <BCS> tags and to those, whose values differ from the corresponding English ones, respective English tag values are automatically entered. Manual confirmation of the replacement is again compulsory in these cases, because there are examples (rare as they might be) where the English and Bulgarian translation equivalents belong to different parts of speech.

The third possibility is an automatic checking and manual correction of missing or incorrect parts of an xml file:

✓ Only Cyrillic characters must be encountered in the text parts of xml files. Of course, mistakes are possible, where a Cyrillic “a” is replaced by a Latin “a”, or where parts that have not been translated from English have been kept. These errors must be checked and if necessary corrected. It is clear, however, that the Latin characters must be kept in some cases, because we could have chemical elements denotations, Latin names of plants, animals, etc. A continuation of this task is spelling checking of Bulgarian <GLOSSES>, <LITERALS>, <USAGE>, <SNOTE> and <LNOTE> tags.

✓ For <ID> tags a check must be performed whether there are empty <ID> tags or duplicated <ID> numbers. Correspondingly, the decision whether a correct <ID> tag shall be connected or which duplicated one shall be removed is to be taken by a human expert.

✓ Another important verification is whether there are duplicated relations between two synsets in a language. It is obvious that such a duplication of relations is impossible and a decision must be taken which relation is correct and which is to be removed.

✓ Each synset must contain at least one literal, possible mistakes are again subject to automatic checking and manual correction.

✓ When building the Bulgarian WordNet we conform to the requirement that an appropriate interpretation definition must be entered for each synset. Thus, another possible checking is for empty <DEF> tags, after which the missing definitions are to be formulated.

As a result of the application of the specified methodology for checking and correction of the Bulgarian WordNet, its current status is the following:

| |
|---|
| No empty tags |
| No <ID>, <POS>, <BCS>, <SENSE> and <DEF> tags without a value |
| At least one literal in a synset |
| No duplicated literals in a synset |
| The <SENSE> tags contain only numbers and are consistent |
| Full correspondence between English and Bulgarian BCS tags |

| |
|---|
| The <POS> tags contain only n, v, and a values |
| Only Cyrillic letters inside the glosses, literals, usage and notes |
| Unified graphical format of the glosses, literals, usage, notes |
| No spelling errors in glosses, literals, usage, notes |
| The <ID> tags are in unified design |
| No duplicated <ID> numbers |
| No duplicated relations between two synsets |

The next important step in the validation of the Bulgarian WordNet is checking its completeness. Under completeness we understand the presence of all members from the chosen up to now Base Concepts 1, 2 and 3 within the framework of the BalkaNet project, as well as the lack of any "dangling relations" (if a certain relation has been copied from the English WordNet then both members of the relation shall be present in the Bulgarian WordNet). “gaps” (if a certain synset is included in the Bulgarian WordNet then all of its hyperonyms shall be present up to the top of the tree) and “gap nodes” (every synset must be linked at list with one relation in the data base). As a result of the checks and corrections performed the Bulgarian WordNet currently contains all members of the sets Base Concepts 1 (1218 members), BC2 (3471 members) and BC3 (3789 members), all members of the tree of a given synsets up to the corresponding top-most synset, as well as the members of each relation entered.

The most difficult and important task is the verification of the consistency of the data – the semantic relations and the interpretation meanings with the synsets. When validating already defined relations the following tests may be used:

- ✓ All Bulgarian synsets with hyperonyms that differ from the English ones or that do not have a hyperonym were checked again. This check may be broadened to cover all relations, as well as all other languages currently developed in the BalkaNet project.
- ✓ The paths from the nodes that are roots or leaves for any relation should be checked again.
- ✓ The linked synsets that contain identical literals (with different senses) should be checked manually.

✓ There must be no hyperonym cycles, as well as any relation loops inside WordNet.

An important theoretical task is the formal defining of the dependencies between the relations (i.e. could we claim that near_antonyms should have a common hyperonym), their formal description in the WordNet Logic and the automatic verification of such dependencies. Criteria that are based on the correspondence between relations are as follows:

✓ hyponyms of two antonyms (nouns) should also be antonyms (*woman – man; female actor – actor*) – obligatory

✓ hyponym should have the same mero-parts (for concrete nouns} as its hypernym (*man – head, arm,… ; woman – head, arm, ..*) – obligatory;

✓ antonyms should have equivalent holo_parts: *woman - arm, head; man – arm, head.*

✓ collective nouns that are holo/mero_members should share the same hyperonym, not compulsory the immediate one *(football team is an organization, as well as football league*) - obligatory;

✓ nouns that are holo/mero_portions should share the same hyperonym, not compulsory the immediate one *(coffee – substance; caffeine - substance*) – obligatory.

When checking for glosses' consistency the following tests are used:

✓ Lines that contain literals with many senses are extracted and the defining of glosses are compared again;

✓ It may be automatically checked whether any literals in the Bulgarian WordNet coincide with their glosses. In such cases the glosses must be redefined.

✓ A check whether the glosses of different synsets are identical or almost identical can be performed. The interpretation definitions must be compared and differentiated in an appropriate manner.

✓ When building the Bulgarian WordNet we have adopted, for English synsets whose notions exist in the Bulgarian language consciousness but have not been lexicalized, to keep the node in the Bulgarian WordNet and to mark it with the phrase "no lexicalization". All entries, marked in this way, were checked again.

✓ The glosses check in corpora is a method that has been used for a long time. As is well known in this way interpretation meanings can be modified, differentiated, a sense that has not been described in uni-lingual dictionaries can be discovered, etc. We also use the traditional method of comparing the collocations of literals.

Up to now the following checks for contradictions in the relations and definitions have been performed on the Bulgarian WordNet:

| |
|---|
| All synsets whose hyperonyms differ with English were checked |
| All synsets without hyperonyms were checked |
| Some hyperonym tree paths were checked manually in both directions |
| No loops inside Bulgarian WordNet |
| All synsets without lexicalization were checked |
| Some glosses belonging to the literals with many senses were verified again. |
| There are no literals that coincide with their glosses |
| There are no equivalent glosses defined for different entries |

**III. The WordNet Validator**

The WordNet Validator (WNV) is a tool for validation (and correction) of WordNet completeness and consistency. The system is developed in the framework of the BalkaNet project and works with the adopted xml-file format. The WordNet Validator has the following main functions: automatic correction of xml syntax, validation of WordNet completeness and consistency, search for a given synset and visualization of semantic trees.

The user should define two WordNets for comparison and validation – the order of the languages is important, because the first language is compared against the second one. The languages can be set among the last versions of the English, Check, Bulgarian, Greek, Turkish and Serbian WordNets or can be browsed. The language is accepted if it corresponds to several conditions: an appropriate xml format, no empty ID tags and no duplicated ID's.

As we described above our approach to the validation of the WordNets includes three separate levels of different degrees of complexity and significance which present

different possibilities for automatic data correction – checking the syntax of the XML files, completeness checking of the WordNets, and checking for consistency in defining the semantic relations. The functions of the WordNet Validator correspond to those tree levels.

In the following cases the ***automatic correction*** function operates:

Facultative empty tags are removed; duplicated literals in a synset are removed while keeping one of them; sense numbers being random are reordered so that there are no empty tags, all tags contain only numbers, all sense numbers are contiguous and are not duplicated. Statistics of the automatic correction appears at the subdirectory *autocorrect* and a result file *autocorrect.xml* is constructed in which the above listed errors are fixed.

If the user selects ***validation*** function the list box appears in which one, several or all of the following operations could be selected:

(***Checking Wordnet completeness):*** **check BC1, check BC2, check BC3** – validating the presence of all members from the chosen up to now Base Concepts 1, 2 and 3 within the framework of the BalkaNet project; **check "dangling" relations** – checking whether both members of the defined relation are presented in the WordNet; **check "gaps"** – verifying whether all of the hyperonyms up to the top of the tree are present, if a certain synset is included in the WordNet; **check synsets without <DEF> tags; check synsets without any literals**.

(***Verifying consistency of the data***): **check ID format -** here we count also language specific synsets with specific IDs i.e. 19 for Bulgarian.**; check duplicated relations** – checking whether there are duplicated relations between two synsets in a language; **check differences in relations** – finding all synsets whose hyperonyms differ from those of the second selected language, this check may be broadened to cover all relations (the differences in the relations might not be errors, if they exist they show the differences between languages); **check loops -** verifying for lack of hypernym cycles, as well as any relation loops inside the WordNet.

The directory named *results* consists of subdirectories corresponding to the queries that were made and a log file. Each subdirectory contains an xml file with wrong synsets. The log file shows the statistics of errors for a given query.

The ***search*** function allows ID searching – the result is all the available information for the synset associated with the ID – literals, gloss, and all immediate relations in both directions.

The ***visualization*** function enables the tree visualization for a given synset – by putting in the check box the wanted relation (for example up to the topmost hyperonyms or down to the bottommost holo-parts) can be selected.

The WordNet Validator can be used for practical work during constructing monolingual WordNets of Balkan languages as well as for evaluation of the completeness and consistency of different WordNets.

The current versions of the five Balkan WordNets are compared in the following table.

| WordNets | Bulgarian | Czech | Greek | Romanian | Serbian | Turkish |
|---|---|---|---|---|---|---|
| **ID** | 19 | 0 | 817 | 0 | 8 | 0 |
| **DUPLICATED RELATIONS** | 0 | 64 | 0 | 54 | 22 | 45 |
| **NON-LEXICALIZED SYNSETS** | 0 | 3 | 1 | 608 | 0 | 680 |
| **DANGLING hypernym** | 0 | 0 | 0 | 0 | 0 | 0 |
| **DANGLING be_in_state** | 0 | 0 | 0 | 0 | 0 | 1 |
| **DANGLING also_see** | 0 | 0 | 0 | 0 | 0 | 1 |
| **DANGLING similar_to** | 0 | 0 | 0 | 0 | 0 | 23 |
| **DANGLING holo_part** | 0 | 0 | 0 | 0 | 0 | 1 |
| **DANGLING holo_member** | 0 | 0 | 0 | 0 | 0 | 3 |
| **DANGLING subevent** | 0 | 0 | 0 | 0 | 0 | 0 |
| **DANGLING causes** | 0 | 0 | 0 | 0 | 0 | 0 |
| **DANGLING derived** | 0 | 0 | 0 | 0 | 0 | 0 |
| **DANGLING particle** | 0 | 0 | 0 | 0 | 0 | 0 |

| | | | | | | |
|---|---|---|---|---|---|---|
| **DANGLING verb_group** | 0 | 0 | 0 | 0 | 0 | 0 |
| **DANGLING near_antonym** | 0 | 0 | 0 | 0 | 1 | 4 |
| **DANGLING holo_portion** | 0 | 0 | 0 | 0 | 0 | 0 |
| **MISSING BC1** | 0 | 0 | 0 | 0 | 7 | 0 |
| **MISSING BC2** | 0 | 0 | 8 | 0 | 526 | 0 |
| **MISSING BC3** | 0 | 321 | 2575 | 32 | 3445 | 904 |
| **ANY LOOPS** | 0 | 0 | 0 | 0 | 0 | 0 |
| **WITHOUT DEFINITIONS** | 0 | 25665 | 30 | 549* | 743 | 5843 |

* All these synsets are non-lexicalized synsets in Romanian

## 3.2 The Czech wordnet

### Automatic and Semi-automatic Validation

The quality control has been one of the priorities of the BalkaNet project. As our evaluation proves even the actual data from the second year of the project are more consistent that the results of previous wordnet-development projects. Part of the success story definitely lies in the implementation of strict quality control and data consistency policy.

Data consistency checks can be considered from various points of view. They can be fully automatic or need less or more manual effort. Even if supported by software tools, manual checks present tedious work that moreover need qualified experts. Another criterion for applicability of checks is whether they can be applicable all languages or they are language-specific (e.g. constraints on characters from a particular codepage). An important issue is also the need for additional resources and/or tools (e.g. annotated monolingual or parallel corpora, spell-checkers, explanatory or bilingual dictionaries, encyclopedias, lemmatizers, morphological analyzers).

Similarly to the scripts for quantitative characteristics we have developed a set of checks that validate wordnet data in the XML format. The following inconsistencies are regularly examined on all BalkaNet data:

- XML validation – empty ID, POS, SYNONYM, SENSE, ... ;
- XML tag data types for POS, SENSE, TYPE (of relation), characters from a defined character set in DEF and USAGE;
- duplicate IDs;
- duplicate triplets (POS, literal, sense);
- duplicate literals in one synset;
- not corresponding POS in the relevant tag and in the ID postfix;
- hypernym and holonym links (uplinks) to a synset with different POS;
- dangling links (dangling uplinks);
- cycles in uplinks (conflicting with PWN, e.g. *goalpost*:1 is a kind of *post*:4 is a

kind of *upright*:1; *vertical*:2 which is a part of *goalpost*:1);

- cycles in other relations;
- top-most synset not from the defined set (unique beginners) – missing hypernym or holonym of a synset (see BCS selecting procedure above);
- non-compatible links to the same synset;
- non-continuous numbering where declared (possibility of automatic renumbering).

The results of the checks are also regularly sent to the developers that are responsible for corrections. The current practice will be probably even further simplified when a new tool for consistency checking with a user-friendly graphical interface will be developed.

Semi-automatic checks that need additional language resources to be integrated are usually performed by each partner depending on the availability of the resources:

- spell-checking of literals, definitions, usage examples and notes;
- coverage of the most frequent words from monolingual corpora;
- coverage of translations (bilingual dictionaries, parallel corpora);
- incompatibility with relations extracted from corpora, dictionaries, or encyclopedias.

In addition to the above-mentioned checks, BalkaNet developers often work with outputs of various pre-defined queries retrieving “suspicious” synsets or cases that could indicate mistakes of lexicographers. For examples, these queries can list:

- nonlexicalized literals;
- literals with many senses;
- multi-parent relations;
- autohyponymy, automeronymy and other relations between synsets containing the same literal;
- longest paths in hyper-hyponymic graphs;
- similar definitions;
- incorrect occurrences of defined literals in definitions;

- presence of literals in usage examples;
- dependencies between relations (e.g. near antonyms differing in their hypernyms);
- structural difference from PWN and other wordnets.

Besides all the mentioned validation checks, quality of created resources is evaluated in their application. Several partners already used their data to annotate corpus text for WSD experiments. Such an experience usually shows missing senses or impossibility to choose between different senses. Another type of work that helps us to refine information in our wordnet was the comparison between the semantic classifications from the wordnet with the syntactic patterns based on computational grammar.

## *3.3 The Greek Wordnet*

### 3.3.1 Evaluation Approach

In the subsequent paragraphs the methodology adopted for the validation of Greek Wordnet's quality is summarized. In particular, we describe the main objectives of the validation task as these were defined shortly after the review meeting and provide the methodology followed for meeting these objectives.

Objectives of the validation task as defined on June 2003.

A) **Improvement of synset glosses**. Shortly after the review meeting members of both DBLAB and CTI teams started improving the quality of the Greek Wordnet. In specific all synsets of the Greek Wordnet that had English glosses and/or synsets that had no gloss appended were traced and corrected. Correction of synset glosses concerned supplying qualitative translations in both synset literals and glosses as well as the modification of existing glosses, which did not reflect Greek lexicalized concepts. By the reporting period all synsets of the Greek Wordnet falling within BCs subsets I and II have a literal name and gloss attached, which are correct in terms of quality and comply with the concepts being lexicalized by the other monolingual Wordnets. For performing this task various specialized terminological resources were used as well as document collections, which helped linguists define a correct gloss in cases explanatory dictionaries were not sufficient for providing such information.

B) **Validating the quality of the relations that hold among Greek Wordnet synsets**. For carrying out this task we have determined some lexicosyntactic patterns found within dictionary definitions that help linguists evaluate the correctness of existing links. E.g. the "*type of/kind of*" patterns indicate a hyponymic relation between two term senses, whereas the "*part of / consists of*" indicate a meronymic relation and so forth. To automatically apply these patterns in the available lexical resources the UOA team has developed all tools required. In particular, these tools given a pattern search within dictionary definitions and

extract both the lemma and its associate terms. Depending on the type of the pattern the lexical relation that holds between the terms in question is determined. To avoid encoding of the wrong relations among Greek Wordnet's synsets manual verification of the extracted links is performed and when needed corrections were manually performed.

C) **Tracing domain-specific usage of some terms** by consulting specialized glossaries (e.g. environmental, medical etc.).

D) **Checking glosses' completeness and vocabulary coverage**. For the performance of this task some preliminary efforts have been performed concerning the semantic annotation of available Greek corpora and/or document collections. Moreover, the 1984 corpus will shortly be semantically annotated with balkanet synsets in order to reassure vocabulary completeness and glosses coverage across Balkan languages. To that end the 1894 corpus have been processed in all languages involved in the project, alingned and morphosyntactically tagged.

In detail, each of the validation tasks was carried out as follows:

## Methodology Followed for Validating the Greek Wordnet.

### A. Improvement of synset glosses:

The main objective of the task was to make all necessary corrections to synsets that either missed a gloss or to the ones that still had an English gloss attached. Missing glosses were due to translational equivalencies problems and in particular when the English gloss indicated in the ILI records could not be confirmed against Greek explanatory dictionaries.

The first step of the task was to trace all synsets falling into any of the two categories and try to correct or improve their glosses. For the performance of the task additional resources were consulted such as the on line encyclopedia "Science and Life", which is hosted at the web site www.gnosinet.gr and the Valentine's floral creations web site www.valentine.gr are used, so as to gather the necessary information for the definition of the missing glosses. Following, the information obtained out of these

sources was compared against the ILI gloss in order to find out whether they refered to or described the same concept. For further confirmation, the information provided by the web sites was also compared with the hyperonym's gloss of the synset in question. This is due to the fact that some general information retrieved by the documents collected could be possibly located in the hyperonym's gloss. By following the abovementioned procedure the Greek gloss was defined. In cases where it was impossible to locate a correct gloss for the terms in question, a more general gloss was assigned as adopted by the term's hypernym gloss so as to define its general meaning and check it again in the future.

**B. Validating the quality of the lexical relations encoded in Greek Wordnet.**

Taking into account the necessity for verifying the correctness of the already existing lexicosemantic relations that hold among Greek Wordnet, the methodology adopted has been partly based on English Wordnet's bibliography.

Essentially, the two lexicosemantic relations that are of interest are the hyponymic and meronymic ones that hold among Greek Wordnet's nouns. And that is because in some cases there is a difficulty in their discrimination. For example whether a noun should be encoded as a hyponym or meronym to another noun. An indicative example of such cases is whether the synset Germany:1;Federal Republic of Germany:2;Deutschland:1;FRG:1; (07220009-n) should be encoded as a hyponym or as meronym of the synset (07164229-n)European country:1;European nation.

Therefore in order to deal with this difficulty the patterns that were used for the validation of the hyponymic and meronymic relations in English Wordnet's bibliography have been selected and translated in Greek. In specific, the hyponymic relation is indicated by the following lexicosyntactic patterns: *kind of, such as, including, especially, branch of*; and for the meronymic relation: *part of, member of, belongs to, substance of, respectively*.

As soon as the specific patterns have been translated, an attempt of their automatic location in the glosses of Triantafillidis and Patakis Greek explanatory dictionaries took place. Following on from this, the glosses in which a specific pattern was

located were stored in different documents, so as for each pattern a different file was created, where each file contains the literals and their definitions.

Continuously, each pattern was checked separately. Actually, the literal and its gloss were compared against the equivalent literals encoded within Greek Wordnet. During this procedure the context of each pattern (that was located in the Patakis and Triantafillidis glosses) was taken into consideration since it indicated hypernym- hyponym, holonym – meronym relations. Thereafter, based on the relations discovered by using the patterns, the hyponymic or meronymic relations holding among the equivalent synsets of Greek Wordnet were examined.

*For example:*

a. in the "branch of" pattern – file the literal cardiology is located with the definition: "branch of medicine....."which indicates that medicine has cardiology as a hyponym. Then, we verified whether this relation was denoted in the synset cardiology (05151565-n) of Greek Wordnet.

b. in "the part-of" pattern – file the literal face is located with the definition: "the front part of the human head..." which indicates that "head" has as mero-part the synset "face (04816017-n)". Then, we verified whether this relation was denoted in the synset "face" in Greek Wordnet, as well.

**C. Tracing domain-specific usage of terms**

Taking into consideration the difficulty of developing domain - specific synsets, the usage of specialized lexical resources such as medical and phytologic encyclopedias, technological dictionaries, maps and relevant sites arose.

In specific, for the domain-specific synsets, which belong to the category of phytology, the following resources have been used:

- The www.prasino.gr/greek-trees/index.htm web site, which contains a variety of plant categories and their description.
- The http://homepages.pathfinder.gr/agropolis/fyta.html web site, which contains a table of plants classification, including the plant's name, the plant's species, the family and plant's genus.

- The **"Systematic Botanic- Angiosperm"** module, volume 1, Maria Stefanaki – Nikiforaki, Ath. Stamoulis Edition, which contains more detailed information about angiosperm plants, their families and their description as well.

For the domain-specific synsets, which belong to the category of medicine the following resources have been used:

- The www.med.auth.gr web site that contains an English – Greek glossary of medical terms. This glossary has been mainly used for the location of the synset name.
- The "Human Body" and "Body Anatomy" encyclopedias, Domiki Editions. In specific the "Body Anatomy" encyclopedia has been used for the location of the synset name, and the "Human Body" encyclopedia for the description of the required synset respectively.

For the domain-specific synsets, which belong to the category of geography the following resources have been used:

- The **"Geographic Atlas, The continents"**, Ag. Siola – E. Alexiou, new edition, that has been used for the definition of the synset's name.
- Web sites such as www.geocities.com/world_greek_geografia, which contains general information about geography (countries, cities, rivers, lakes, mountains) have been used. The information retrieved is related to the history, economy or the culture of each country that is helpful for the definition of the synset's gloss.

Last but not least the www.gnosinet.gr/es/thematic.html web site, which contains information about various topics such as technology, politics, science, sports has been used for the definition of glosses.

However, during the development of domain-specific synsets some obstacles appeared. In specific, up to this point, the difficulties met focus on domains such as phytology. In these categories the synset name derived from taxonomies in which the Latin term is considered to be valid so it was not possible for an equivalent

synset name to be traced in Greek. For example, the results given by the forenamed resources have as following; the domain specific synset is located with its latin name only and its description in Greek e.g. Ananas, genus Ananas (10546618-n).

Therefore, the Latin term is maintained and a Greek definition extracted from the abovementioned resources is appended.

In addition there are cases where although the domain specific synset is located in our resources, its description is not helpful for the definition of the gloss so a more general gloss is adopted by its hypernym. e.g. *kauri* (09596268-n), in Greek Wordnet its gloss is adopted by its hypernym *wood* [*12772693-n*].

**D. Checking glosses' completeness and vocabulary coverage.**

In order to check glosses' completeness we plan to semantically annotate the National Hellenic Corpus (HNC), delevoped by ILSP, Greece[1]. In particular, semantic annotation concerns assigning Greek Wordnet's synset glosses to corpus terms (mainly, nouns and verbs) so as to check possible glosses that might be missing from Wordnet and verify the completeness of the existing ones. Sense assignment will be performed manually by experienced lexicographres with the aid of a semantic annotation tools designed specifically for this purpose. The annotation tool will be completed shortly and the actual annotation of corpus terms will begin. Due to the fact that semantic annotation is a rather time-consuming task a selection of some synsets will take place and these will from the final set on which annotation will be perfomed. Selection criteria will mainly focus on polysemy issues. Specifically, semantic annotation is envisaged to take place as follows: for each selected synset the respective literal will be searched against HNC sentences. Then, a Wordnet sense will be manually assigned to the term in question. Where more than one sense is applicable then all possible senses will be assigned and disambiguation will be performed on the basis of collocations information provided by the corpus. We expect that semantic annotation will be helpful not only for validating the completeness of the synsets' glosses but also for future enrichment of Wordnet's.

[1] http://corpus.ilsp.gr

### 3.3.2 Statistical data of the Greek Wordnet (as of November 2003)

In the subsequent tables all statistics performed by DBLAB concerning the coverage and POS-distribution of Greek Wordnet are summarized. As illustrated by the figures what still needs to be accomplished in terms of quantitative data is the enrichment of nouns, which will follow naturally the development of BC 3 and the enrichment of adverbs. The latter impose some additional difficulties due to their dependence on the respective adjectives. However, by the end of the project it is estimated that an adequate number of adverbs will be encoded.

| **Current status of Greek Wordnet** |
|---|
| BC 1 = 1221, BC2 = 3455, BC 3 = 1098 / 3827 |
| Total number of synsets: 15560 |
| Nouns: 12320 (79,1%) |
| Verbs: 2882 (18,5%) |
| Adjectives: 344 (2,2%) |
| Adverbs: 14 (~1%) |

***Table 1****: Overall statistics of the POS distribution of Greek Wordnet's synsets*

| **Relations holding among synsets in Greek Wordnet** |
|---|
| Hypernyms: 9238 synsets |
| Hyponyms: 11692 synsets |
| Holo_parts: 1623 synsets |
| Mero_parts: 1742 synsets |
| Holo_member: 2095 synsets |
| Mero_member: 330 synsets |
| Holo_substance: 59 synsets |
| Mero_substance: 59 synsets |
| Antonyms: 27 |

| Near_antonyms: 168 |
|---|

***Table 2**: Overall statistics of the lexical relations encoded in Greek Wordnet*

### 3.3.3 Evaluation Results (as of November 2003)

Validation tasks are being performed by all teams during the last 3 months (August 2003-November 2003) of the project and even though quality control is a time-consuming and contious task, nevertheless some of the results obtained so far with respect to the Greek Wordnet validation are summarized below:

- All literals in any of the Greek Wordnet's synsets are assigned a sense identifier. Moreover, there are no identical literals witin the same synset. Each literal has a unique sense identifier.
- Each synsets is tagged with a unique POS tag.
- Each literal is appended at least one gloss and all glosses are checked in terms of spelling and quality. Quality control reassures that the correct concepts is lexicalized by a given gloss literal.
- All synsets are inter-linked with one or more of the pre-defined lexical relations and there are no loops.
- All dangling links and/or synsets have been eliminated.

### 3.3.4 Tasks currently in progress (to be finalized by the end of March 2004)

During the reporting period and for the subsequent four months further testing of each monolingual Wordnet will take place so as to reassure their completeness, vocabulary coverage and lexical links validity. A testing phase that is at about to begin concerns the semantic annotation of the multilingual 1984 (Orwell) corpus so as to verify that there is a significant qualitative and quantitative overlap across Wordnets. For the performance of semantic annotation several tools are currently under development by all teams. Such tools need to be evaluated in order to decide on the technical infrastructure that will be adopted towards the semantic annotation task.

Moreover, subset III of the BCs is currently under development and is expected to be finalized latest by the 5$^{th}$ progress meeting to be held in January 2004. Several difficulties that have been faced so far concerning the lexical relations and glosses to be encoded for BCs 3 will be discussed and resolved during the meeting. Briefly according to the T.A

and the project's internal timeplan it is expected that by the end of March 2004 the final versions of the monolingual wordnets will be delivered and their qualitative content will be guaranteed by each contractor. Quality control is a time-consuming but nevertheless essential task that needs to be performed in order for the project's application to deliver satisfactory results.

## *3.4 Romanian wordnet*

### A. Syntactic validation

The approach on the Romanian Wordnet is not a translation approach. Due to different granularity in sense definitions between Princeton Wordnet and the Explanatory Dictionary of Romanian (plus the Dictionary of Synonyms) and because the mapping over the interlingual index has been independently done by the members of the Romanian team, it was unavoidable to face difficulties in the correct mapping. As such, various literals with the same sense number appear in more than one synset. This is an error that can be easily traced by syntactic validation methods, but correcting it needs introspective analysis as the solution might not be obvious:

- one could simply modify some synsets, leaving the conflicting literal and sense number in only one synset (decide on which should remain and which should be deleted)
- one could assign different sense numbers to the conflicting literal (decide on which sense number will be preserves in which synset and which sense numbers will be modified in which synsets); this case raises the issue of defining new senses not previously recorded in our reference dictionary.

Besides this type of errors, there are several other purely syntactic errors that can be also easily traced and corrected.

In the first phase of the project, as decided by the consortium, we were concerned only with the cross-lingual ILI coverage and quantitative aspects of the synsets implementation and therefore the first 8.000 ILI's (BC1, BC2 and BC3) were implemented without checking their syntactic correctness. Once the quantitative aspect has been fulfilled, the consortium decided to bring in the forefront of our activities the quality insurance and let aside, for the moment, the task of adding new synsets, with more emphasis on the correction of the syntactic errors (monolingually) and on the cross-lingual validation.

We adopted a two-way strategy:

- for the synsets already done we have written a script which checks the syntactic correctness;
- for the synsets that were to be done we modified the interface so that it does not allow anymore building syntactically incorrect synsets.

The general structure of an entry for a synset in an XML file, which stores the Romanian WordNet, is:

```
<SYNSET>
      <ID>ENG171-00003135-n</ID>
      <POS>n</POS>
      <SYNONYM>
            <LITERAL>ființă<SENSE>1</SENSE></LITERAL>
            <LITERAL>viețuitoare<SENSE>1</SENSE></LITERAL>
            <LITERAL>vietate<SENSE>1</SENSE></LITERAL>
      </SYNONYM>
      <DEF>Tot ceea ce are viață</DEF>
      <STAMP>cineva</STAMP>
      <BCS>1</BCS>
      <ILR><TYPE>hypernym</TYPE>ENG171-00002956-n</ILR>
</SYNSET>
```

A1) The script we created verifies the following:

The general structure of the <SYNSET> tag is well-formed, i.e. it contains the tags <ID>, <POS>, <SYNONYM>, <DEF> and, optionally, the tags <STAMP>, <BCS>, <ILR>, <ELR>.

- According to a much disputed decision of the consortium, the synsets of the BALKANET wordnets are to be interlingually mapped to ILI only by the EQ-SYN external relation. As such because the ILI record is uniquely identified by the content of the ID tag, the <ELR> (external language relation) became redundant. However, since we do believe that various other external relations are extremely useful representation devices we retained it in the source format of the Romanian Wordnet. For compatibility with other Wordnets in the consortium based on a translation approach, the external relations different from EQ-SYN are automatically converted into an EQ-SYN by means of creation of an internal non-lexicalised synset. A non-lexicalised synset has similar structure to a usual synset but the sub-structure:

`<SYNONYM><LITERAL>…</LITERAL></SYNONYM>` becomes `<NL>yes</NL>`. For instance if the previous synset were not lexicalized in Romanian, then its encoding would have been:

```
<SYNSET>
   <ID>ENG171-00003135-n</ID>
   <POS>n</POS>
   <NL>yes</NL>
   <DEF>Tot ceea ce are viață</DEF>
   <STAMP>cineva</STAMP>
   <BCS>1</BCS>
   <ILR><TYPE>hypernym</TYPE>ENG171-00002956-n</ILR>
</SYNSET>
```

Some of the non-lexicalized synsets have been given a gloss representing the translation in Romanian of the English gloss attached to corresponding synset in PWN. Currently the Romanian wordnet contains 608 non-lexicalized synsets which are subject to further scrutiny. Besides leaving the non-lexicalized synsets as they are now, another possible solution would be to define multiword lexical items (as many English synsets do for our present non-lexicalized synsets). This will be solved the way the consortium will decide at the meeting in January 2004.

For the tags enumerated under A1) it checks:

- for **<ID>**: this has to contain a valid ILI identifier; no such error exists in our wordnet.
- for **<POS>**: this has to have the same value for **<POS>** as the corresponding ILI record; no such error exists in our wordnet.
- for **<SYNONYM>**: it has to contain only **<LITERAL>** tags; in its turn, this has to contain a string in the UTF-8 format followed by the tag **<SENSE>**: generally, the value of the **<SENSE>** tag is an integer; however it may be an alphanumeric string; the BNF description of the value of a sense identifier is the following:

```
<sense-identifier>::=<integer>|                    (a)
                     <integer1>.<integer2>|        (b)
                     <integer>.<letter>|           (c)
```

```
<integer1>.c<integer2>    (d)

<letter>                  (e)

<letter>.c<integer>       (f)
```

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 labeling used in the Explanatory Dictionary of Romanian and we kept it as it represents information that we don't want to loose. It stands for the <integer2>$^{th}$ sub-sense of the <integer1>$^{th}$ sense of the current literal. 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 the (b) notation, we kept it in our wordnet with the same interpretation;

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. 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 coarse grained sense which must be split into sub-senses if not a sense-assignment error made during the wordnet construction. After introspective analysis, the notation of this type should be, in general, turned into a notation of type (c). In this case, the glosses might need particularization so that to make distinction between the finer grained senses.

A sense identifier of **type (e)** represents a sense which is not listed in the Explanatory Dictionary of Romanian but we felt as a legitimate distinct one. In this case, the gloss represents simply the translation of the corresponding sense in PWN. Instead of a letter we could have used one integer larger than the one of the last definition listed in the reference dictionary. However, with more than a single missing sense for a given headword, currently 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.

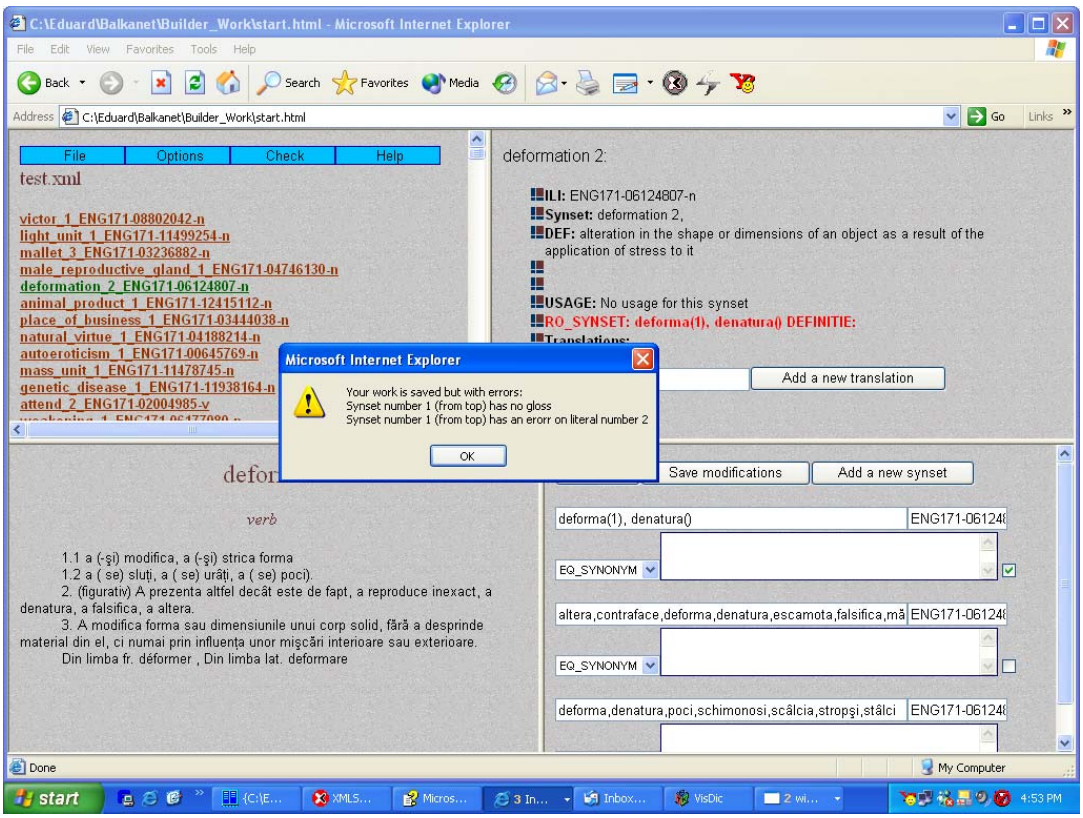
Finally, a sense-identifier of **type (f)** represents sub-senses of unlisted senses of the current literal. This notation is analogous to a (b) notation.

We should mention that the last four types of sense-identifiers could be automatically turned into a notation of the type (a) or (b) unless the sense-numbering sequence is not used or is not relevant. However, in the Explanatory Dictionary of Romanian the numbering order of senses is assumed to be meaningful.

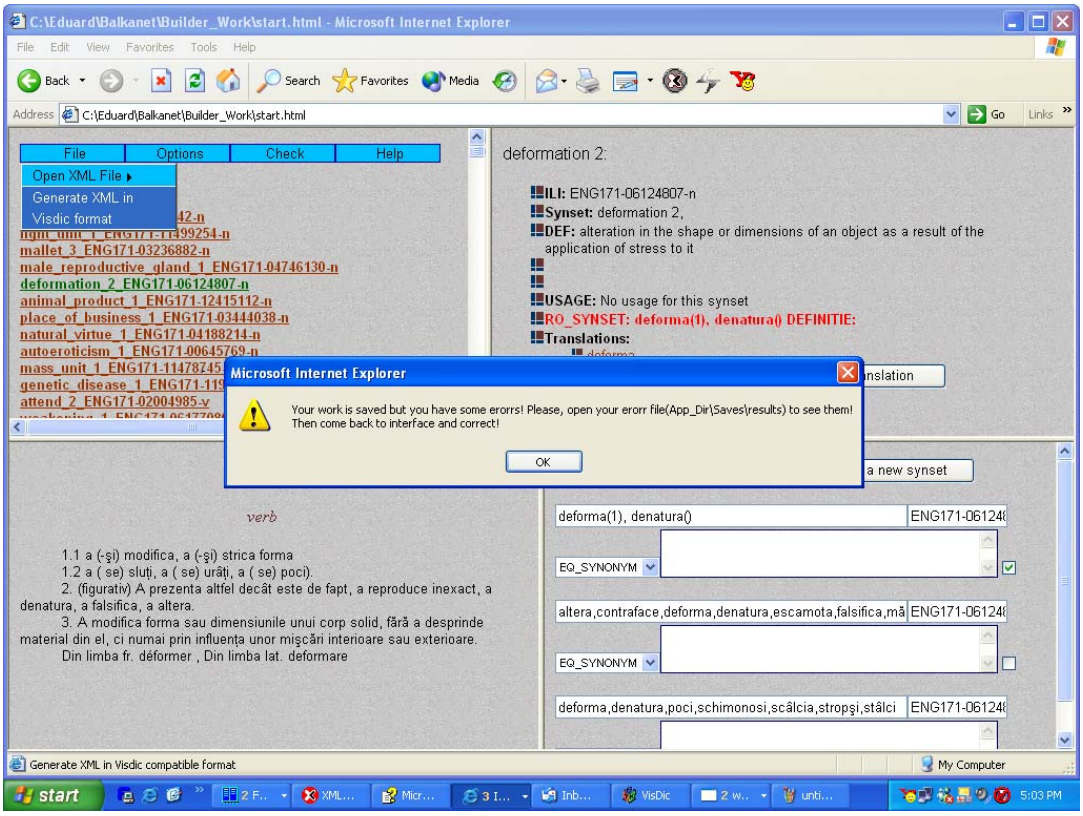
- for **<DEF>**: it should be a piece of text in the language for which the wordnet is built; in our case, the vast majority of glosses are automatically extracted from the Explanatory Dictionary of Romanian; when the definitions were not available, they were translated from the corresponding glosses of PWN; no synset in our wordnet misses its gloss, except for the (majority) non-lexicalized synsets. We plan to translate all the glosses for the non-lexicalized synsets in the immediate future;
- for **<STAMP>**: it contains the name of the person who last modified the synset; this is not verified;
- for **<BCS>**: it checks if its value is the same as the value of the **<BCS>** in the corresponding ILI record; no such error exists in our wordnet.

- for **<ILR>**: it has to contain both the tag **<TYPE>** whose value has to be a relation from the agreed set of relations, and an ILI record which has to be in the set of ILI records for which we assigned synsets; no such error exists in our wordnet.

A2) After checking the approximately 8.000 synsets in BCS1,2,3 using the above mentioned script, we modified the WNBuilder interface so that it does not allow the human user to make syntactic mistakes when implementing new synsets. When the user wants to save the implemented synsets, the interface checks its well-formedness according to the criteria mentioned before and, if the case, a message appears on the screen, warning him about the syntactic mistakes he did:




The user may either ignore the message and postpone the correction or correct it. If he chooses to postpone the correction, when the interface exports the work of the user in an XML file compatible with VisDic format, the interface will warn him again about the mistakes as in the following snapshot:




A3) Other syntactic tests, not included in the WNBuilder interface, but available as command line scripts are described below.

**Dangling relations**: according to the definition given in the previous section, this test checks whether or not there are dangling relations in a given wordnet; no such error exists in our wordnet.

**Dangling nodes**: according to the definition given in the previous section, this test checks whether or not there are dangling relations in a given wordnet; no such error exists in our wordnet.

**The same literal occurring more than once in a synset**: this problem does not exist in Romanian wordnet any longer; by means of the function implemented in VisDic, we identified the synsets which had this problem; we found very few such situations which we manually corrected.

Most of the detected errors were corrected manually assisted by VISDIC or WNBuilder. However, as these tools were not developed especially for error corrections but mainly

for synset implementation and error detection, we built a specialized tool, WN-Correct, meant to allow a more friendly and effective control over the corrections in sense assignment errors. WN-Correct has two variants, one oriented on literals and the other one oriented on synsets. Although they are functionally equivalent, some members of the development team prefer one version while the others prefer the other version.

Working with WN-Correct1 assumes the following steps:

- Identify the literals with senses in conflict, i.e. the same literal appears in two or more synsets with the same sense;
- Collect all synsets containing those literals;

Each member of the validation team has a set of literals that were used with the same sense number in two or more synsets; their list is displayed in the upper left panel in figure 1 below; when clicking such a literal, in the upper right panel appears the list of synsets containing the offending 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 delete the literal from the synset if it does not belong to that synset. The figure 1 suggests how the conflict for "osie [1]" has been solved: the sense number in the first synset (ENG20-02669073-n) was changed from 1 to 3. Since in the Explanatory Dictionary the headword "osie" has only two senses, a new gloss has been created. The newly defined sense is automatically added to our Explanatory Dictionary (shown in different color in the bottom panel).

The advantage of this procedure is that at the end of the validation task, there will not be any 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. These have to be implemented again using the WN-Builder interface.

Another main problem is that this procedure does not allow the lexicographers to modify the synsets except for the conflicting literal.

Moreover, the same synsets are checked by several lexicographers, which is a too much time consuming procedure.

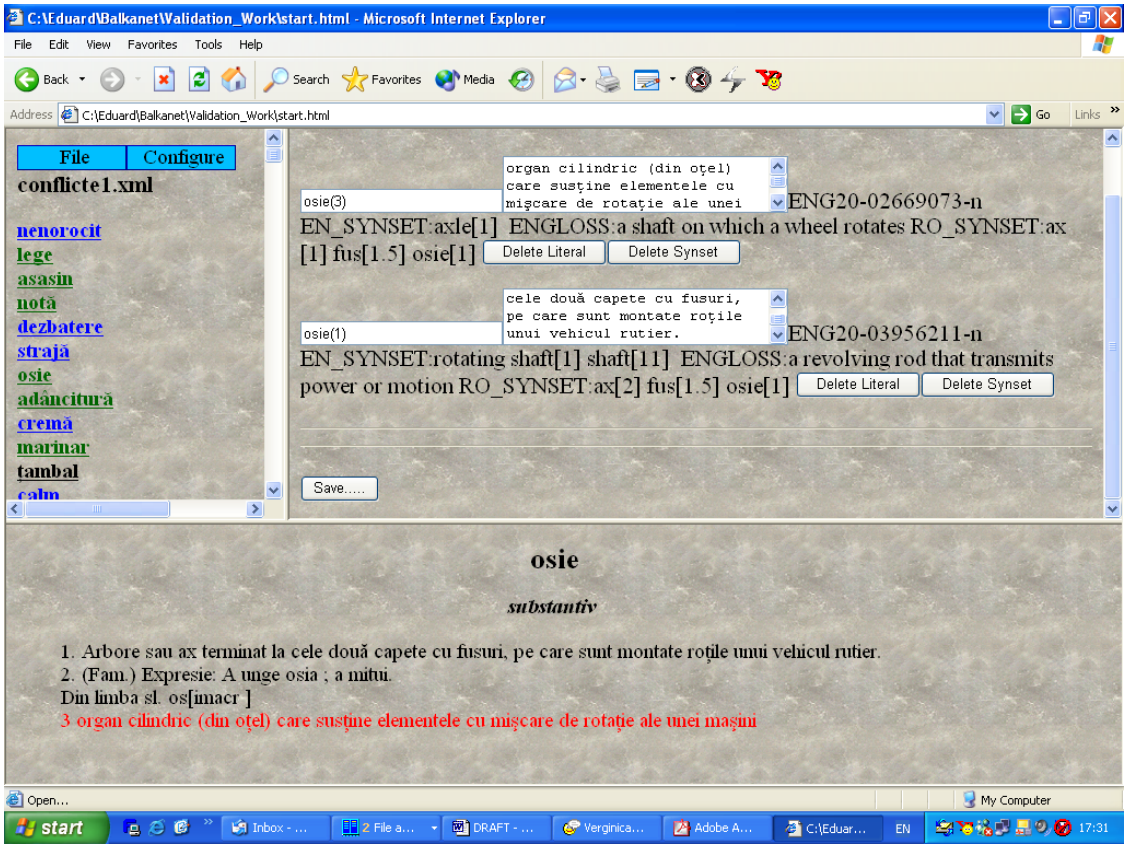



Figure 1

Using the the second variant, WN-Correct-2, assumes the following steps:

- Identify the synsets with literals in conflict;
- Different lexicographers will be given disjoint sets of synsets;
- As the lexicographer is now responsible for the correctness of the whole synset, 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.
- WN-Correct-2 has a function which checks on the fly the work of the lexicographer for new conflicts. If there are any, they will be solved by the same lexicographer.
- The corrected synsets replace the initial ones in the WordNet database and the procedure is repeated from the first step until there are no more conflicts left.

In the figure 2 you can see a snapshot from Wn-Correct-2 session. The Add links button (top of the upper right panel) will add links to our explanatory dictionary.

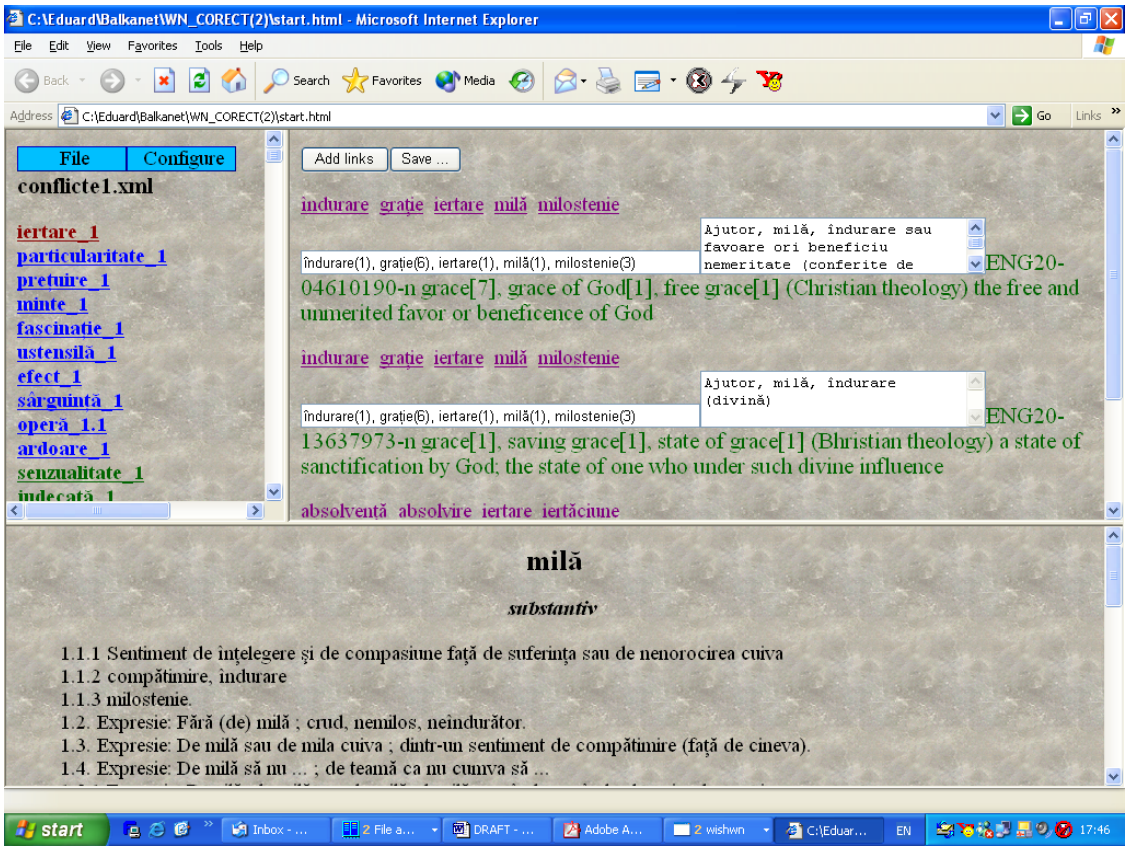



Figure 2

One problem that we dealt with during the improvement of our wordnet was marking up the reflexive pronouns that either co-occur obligatory or optionally with some verbs: the reflexive pronouns that obligatory accompany some verbs in verbalizing a specific meaning are put inside square brackets. The omission of an obligatory reflexive pronoun for a verb is either ungrammatical or radically changes the meaning of that verb. The reflexive pronouns which are not mandatory, are surrounded by vertical bars | |. Their omissions usually produce a slight meaning shift of the verb anyway.

Ex.: *[se] uita(7)* is the Romanian equivalent of the English *look(1)*;

*|se| spăla(2)* is the Romanian equivalent of the English *wash(2)*.

## *3.5 The Serbian wordnet*

Section one gives a brief outline of the state of the art of the Serbian wordnet. Section two describes specific validation tasks already performed by the Serbian team. Section three describes the validation tasks that are still underway or are being planned.

**I State of the art of the Serbian wordnet**

The Serbian wordnet has been developed under conditions which differ from wordnets for other languages within the Balkanet project. The Serbian team has entered the project as a subcontractor of DBLAB at a later stage of the negotiations with limited man month and budget allocation. Due to this fact Annex I of the Consortium Agreement envisaged only a limited, approximately 1500 synset large Serbian wordnet.

In spite of its somewhat specific position, the Serbian team is making every effort to keep the pace with other Balkanet wordnets and the Serbian wordnet to date includes 4818 synsets. The wordnet is constantly being developed with the goal to attain lexical coverage as close as possible to the one targeted by other languages.

The distribution of developed synsets within the BC sets is summarized in the following table:

| | No of synsets | Planned | Realized (%) |
|---|---|---|---|
| BC1 | 1212 | 1219 | 99.4% |
| BC2 | 2982 | 3508 | 85.0% |
| BC3 | 387 | 3788 | 10.2% |
| other | 237 | | |
| total | 4818 | | |

The next table shows the PoS related distribution of synsets and literals, the literal/synset (l/s) ratio, the number of duplicate literal+sense (l+sen) pairs that have not yet been resolved. The last column in the table shows the literals that have the greatest number of senses in certain PoS categories.

| | synsets | | literals | ratio l/s | duplicate l+sen | max. senses per lit. |
|---|---|---|---|---|---|---|
| nouns | 3161 | 65.6% | 5115 | 1.99 | 52 | "mesto", 11 senses |
| verbs | 1490 | 30.9% | 2968 | 1.62 | 98 | "drzxati", 13 senses |
| adjectives | 156 | 3.2% | 216 | 1.38 | 3 | "velik", 8 senses |
| adverbs | 7 | 0.1% | 7 | 1.00 | 0 | |
| total | 4818 | 100.0% | 8306 | 1.73 | 153 | |

The relations established between synsets in Serbian wordnet are summarized in the following table:

| | |
|---|---|
| Hypernym | 4391 |
| near_antonym | 362 |
| holo_part | 220 |
| verb_group | 132 |
| holo_member | 67 |
| be_in_state | 67 |
| Subevent | 55 |
| Causes | 44 |

Out of 4818 synsets, 4079 (84.7%) now have glosses, and for 739 synsets glosses remain to be added.

**II Performed validation and enhancement tasks**

1. The Serbian wordnet is being developed in accordance with the six volume standard explanatory dictionary of Serbian (Rečnik Matice srpske). The validity of all literals has been initially checked against this dictionary. The Serbian team has decided not to assign independently sense numbers to literals but rather use appropriate numbers from this dictionary whenever possible. However, for various reasons this has not always been possible and in those cases we have used non-numeric (x, y, z...) and mixed (1a, 1b, 1c...) sense annotation. For the same reasons, sense numbers do not necessarily follow a sequence but can have “gaps”. Presently, we do not envisage this specific feature as a shortcoming which could in any way affect other wordnets within the project. If however, it turns out that this assumption is wrong we will consider all possible measures to overcome this potential problem.

Additionally, the validity of literals has been checked using the morphological electronic dictionaries in Intex format for Serbian developed by the Serbian team. The system of morphological e-dictionaries of simple words in Intex format consists primarily of three parts: dictionary of lemmas (DELAS - around 60.000) , dictionary of word forms (DELAF - around 900.000) and regular expressions implemented by finite transducers that describe the inflectional properties of entries in DELAS. These dictionaries were used to include morphological and syntactic information related to synset literals using the LNOTE tag. Lack of this information in a wordnet is considered as an essential shortcoming in the case of Serbian language. Without this information the validation of the wordnet on a corpus, which is essential for determining the quality of a wordnet, is greatly impeded. The number of literals with morphosyntactic information in the LNOTE tag is presently 5549 (66.8%), while this information needs to be added to another 2757 literals (33.2%).

2. For further validation of the literals we have used both the Serbian monolingual corpus and parallel Serbian/French and Serbian/English corpora. The Serbian monolingual corpus has now more than 50MW and is constantly being enlarged. It consists of texts from various sources: newspaper, agency news, literature, and textbooks. A part of this corpus (22MW) is now available on-line at http://korpus.matf.bg.ac.yu/korpus (for authorized users). The size of both multilingual corpora is now close to 1MW. Texts in parallel corpora are aligned on the sentence level using different alignment programs.

   For corpora pre-processing the Intex system, based on appropriate e-dictionaries and finite state transducers, has been used. The standard distribution of this system incorporates morphological e-dictionaries for French and English. In addition to that, Serbian morphological e-dictionaries described in the previous section have been used.

   A brief description of the validation process follows. The validation process starts with the search for the occurrences of literal strings from Serbian synsets in the

Serbian monolingual corpus and the Serbian parts of multilingual corpora. For all occurrences it is checked whether they conform to the synsets to which the literal strings belong. This process can confirm the inclusion of a literal string into a synset or lead to its exclusion and possible move to some other synset. For instance, the verb boraviti has been originally placed in the synset (stanovati:1b, zxiveti:4, boraviti:1, prebivati:1) that corresponds to the synset (dwell:2, inhabit:1, live:6, make one's home:1, people:6, populate:1, reside:2, shake:3) from PWN. However, concordances produced by Intex showed that this verb has the exclusive meaning of a temporary stay and that it was misplaced in this synset, as shown in the following table:

| |
|---|
| atski predstavnici koji borave u Skoplju diskretno sugerisali |
| Zvornik, sxto, kako je, boravecxi danas u Loznici, objasnio |
| i princeza Katarina, boravicxe sutra u Novom Sadu, saopsx |
| avgustu Avramovicx je boravio u Sxvajcarskoj, pa posle u Am |
| im cxe, pored Beograda, boraviti i na Kosmetu i u Crnoj Gori. |

Bilingual corpora can be used for synset validation in a more fruitful way, especially having in mind the request that all synsets from a wordnet for languages other that English have to be associated, if possible, to a corresponding English synset via ILI. Thus between synsets in English (or French) wordnet and Serbian wordnet a one-to-one correspondence is established on basis of the EQ-SYNONYMS relation. For instance, a 1-1 correspondence exists between the following synsets:

(glava:1) <---> (head:8)
(glava:5, odgovorno lice:1)<--->(chief:2, head:19,top dog:1)
(glava:2,um:1a)<--->(brain:2,head:9,mind:1,nous:1,psyche:1,chief:1)

Between the literal strings from the English wordnet (or French wordnet) and the Serbian wordnet, however, a many-to-many correspondence exists. The purpose of the validation process is to investigate the nature of this many-to-many correspondence and confirm or reject its appropriateness.

The validation process proceeds in two steps:

- One literal string from Serbian wordnet is searched for in the Serbian part of the bilingual corpus and the matching English/French terms are identified in the English (or French) part of the corpus.

- All literal strings in the English (or French) wordnet that are in correspondence with the chosen Serbian literal string are searched for in the English (or French) part of the corpus and matching Serbian terms are identified in the Serbian part of corpus.

The nature of the correspondence is then analyzed on basis of the matched pairs of terms. This analysis can either lead to a removal of some links from the initial correspondence or to the addition of new Serbian literal strings and new links. An excerpt from the concordances of aligned corpus is shown in the following table:

| |
|---|
| easy.<Oshs.1.2.20.6> Trebalo je samo da prenese na papir onaj neprekidni i nesmireni monolog koji mu se doslovno godinama odvijao u glavi. .EOS <Oen.1.1.19.6> All he had to do was to transfer to paper the interminable restless monologue that had been running inside his head, literally for years. <Oshs.1.2.20.7> Medxutim, u tom trenutku je <Oshs.1.2.23.3> No cyudno je bilo to sxto mu se, dok je pisao, u glavi osvetlila jedna sasvim razlicyita uspomena, i to do te mere da se osetio sposobnim da je prenese na papir. <Oen.1.1.22.3> But the curious thing was that while he was doing so a totally different memory had clarified itself in his mind, to the point where he almost felt equal to writing it down. |

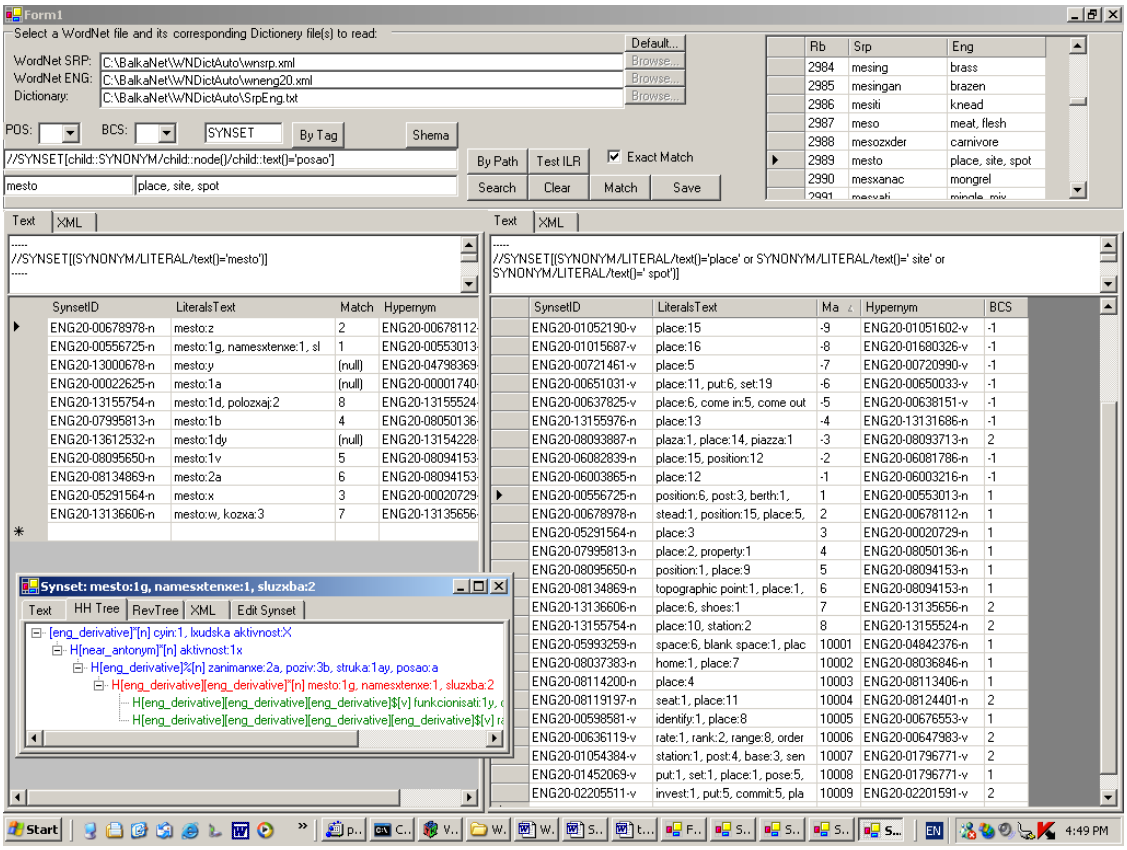
The results obtained by validating a representative group of synsets fully approve the usability of corpora approach to the validation of wordnet synsets. Besides the reestablishment of synsets themselves, this approach enables the establishment of relations between various derivatives, either by including them in the same synset, if they have the same PoS, or by setting up a cross-PoS relation. In this respect the corpora approach is particularly useful in detecting the derived forms in connection to the senses. The other useful issue here is the detection of phrases and their translation equivalents.

Another important use of Serbian corpora for validation purposes is the extraction of examples of literal usage from the corpora and their inclusion in the synsets under the USAGE tag. Presently, 286 synsets have been checked against corpora, and as a result 348 USAGE tags have been added to the Serbian Wordnet.

3. A tool for the integration of various lexical resources such as the Wordnet, e-dictionaries, and bilingual word lists is being developed by the Serbian team. A part of this integrated tool is already implemented and will be used for wordnet development and refinement. On basis of existing wordnet and bilingual word lists

the tool helps the user generate new synsets and validate the existing ones, including the addition of new literals. The tool uses XML files compatible with the VisDic standard.

The tool is illustrated by a figure showing the matching of synsets containing the Serbian literal “mesto” and its English counterparts from the bilingual word list (place, site, spot).




## III Further developments

In the period remaining for the fulfilment of WP6 the following tasks will be performed.

1. Synsets from BC1 and BC2 will be fully covered.
2. Synsets from BC3 will be included that fill in possible gaps in BC1 and BC2 as well as those related to BC1 and BC2 synsets with one of the following relations:

near_antonymy, mero_part/holo_part, mero_portion/holo_portion, mero_member/ holo_member, derived, causes, particle.

3. Missing glosses (DEF tags) will be filled and existing ones rechecked.
4. The contents of the USAGE tag will be filled in wherever possible.
5. The missing LNOTE tags will be added.
6. The tagged and disambiguated version of 1984 will be completed. The full Serbian version of 1984 exists in electronic form, with a markup structure to the sentence level and many phrase element tags (names, abbreviations, foreign words, etc.). Also, there exists a full version aligned to English on the sentence level. So far, 16 chapters have been tagged and disambiguated (14 already posted at the IS) and the remainder will be completed in the near future.

**IV Further plans**

The Serbian team also plans further validation of synsets based on their lexical frequency. The validation results will be used for removing existing or adding new literals to the synset. The information on synset validation will be stored in the LNOTE and NOTE tags. The NOTE tag will contain information whether a synset has been validated, and the type of corpus used (mono/multilingual). The LNOTE tag will contain, besides morphological and syntactic information discussed in the previous paragraph, one of more indices indicating the relevance of the appropriate literal within the synset in terms of its lexical frequency. The Serbian team has developed a set of these indices and will present them at GWN 2004. It should be noted that the envisaged validation task is a rather ambitious and time consuming one and that it is realistic to estimate that it can be fulfilled.

## The Turkish wordnet

Validation tasks gain more importance as our monolingual wordnets expand. The sections below will explain the validation tasks we have adopted so far, as well as the procedures we are planning to undertake in the near future.

### 3.6.1. Syntactic Quality

We first ensured the syntactic quality of Turkish Wordnet in XML format. Each opening tag has a closing tag. All synsets have one and only one <SYNSET> tag, one and only one <ID> tag, one and only one <POS> tag. Unless the synset is one of the nonlexicalized concepts, it has at least one <LITERAL> tag- with its subtag <SENSE>. Otherwise, it has the special <NL>yes</NL> tag. Since we have not finished assigning glosses to our synsets, some of our synsets do not contain a <DEF> tag. The <ILR>, <BCS>, <STAMP> tags are optional. Among those optional tags, only the <ILR> tag can have more than one occurrence within a synset. We also validated the values of the tags in terms of integrity, where possible. All the values in the <ID> tag are well-formatted valid ILI numbers, i.e., are in the form ENG20-XXXXXXXX-X. This criterion holds for the ILI numbers in <ILR> tags. All the <POS> tags have one of the following values: n, v, or a (b is also a valid value for the <POS> tag, but we do not have it currently, since we have not implemented adverbs in our wordnet yet). If a <BCS> tag exists, it can only contain one of the following values: 1, 2, or 3. This criterion has been verified for the latest TWN XML file. There are no empty tags such as <LITERAL><SENSE></SENSE></LITERAL> or <DEF></DEF>, etc.

### 3.6.2. Structural Quality

#### 3.6.2.1. Gaps

We obeyed the rule that the wordnets should not contain any "gaps". We managed not to have gaps by running a Perl script to find its gap hyperonmys in PWN 2.0 whenever we

add a bulk of synsets to our wordnet, and then, by translating all the gaps and adding them to the existing wordnet.

#### 3.6.2.2. Close-world Assumption and Dangling Relations

Due to the close-world assumption we have adopted, if a relation is defined between ILI 1 and ILI 2, in which ILI 1 is contained in the wordnet then ILI 2 should also be contained in the wordnet. Relations where ILI2 is not contained in the wordnet are called "dangling relations". All such relations have been identified with the help of a small Perl script and are being translated into Turkish. The current version of our wordnet may have some dangling relations which means the translation process is in progress.

#### 3.6.2.3. BCS1, BCS2, and BCS3

In line with the common decisions taken by all participants, each wordnet (except Serbian) should have synsets labeled BCS1, BCS2, and BCS3. We, as the Turkish team have finished BCS1 (1218 synsets) and BCS2 (3471 synsets) and tagged them with a <BCS> tag within the XML format. 2877 out of 3872 BCS3 synsets have been completed; the rest will be added to TWN until the end of the year.

#### 3.6.2.4. VisDic Tests

We applied VisDic's duplicate tests on our wordnet. We identified some duplicate ID's and manually deleted the one containing less information, i.e. synsets with fewer literals or with no definitions. The current version of our wordnet does not have any duplicate ID's.

The file also passed the duplicate synset literal test with very few mistakes. Another test VisDic offers allows us to identify duplicate links. This test prevents the user from linking two synsets via more than one relation. For instance, a synset cannot be both the "hyperonym" and "antonym" of another synset. But the new PWN 2.0 relations "verb_group", "similar_to", and "also_see" are exceptions to this case. In our recent wordnet we had 45 duplicate links, all of which are instances of these three relations.

The last VisDic test checks if the same literal with the same sense number occurs in more than one synset. As is the case with some other partners, we used our monolingual dictionary while assigning glosses and sense numbers to our synsets. Sometimes, PWN

2.0 is more fine-grained than our TDK Turkish Dictionary and therefore we assigned the same sense number to some literals which occur in different synsets. One solution may be to deliver the final product by assigning auto-incremented sense numbers. But, for the time being, duplicate literals and sense numbers will be kept.

### 3.6.3. Content Quality

#### 3.6.3.1. Spelling Correction

The linguistic content of our wordnet (synset members, glosses and usage examples, if any) should be passed through a spelling corrector. Although this may sound too simple and straightforward, it is a basic step to be achieved. All synset members have been examined and validated, but glosses have not been spellchecked yet.

#### 3.6.3.2. Validation of Relations

We believe that, all relations imported from Princeton WordNet (or any other wordnet) should be manually, semiautomatically or automatically validated. For the synsets we have translated, we have imported all relations contained in PWN 2.0. We then erased the English-specific relations "eng_derivative" and "region_domain". There are some other relations such as "also_see", "verb_group" and "similar_to" which explodes the number of related synsets and therefore affects the close world assumption. In other words, to satisfy the close-world rule, a vast number of synsets have to be added to the existing wordnet. For those cases, we kept the relation if we already had the related synset, but removed dangling relations.

### 3.6.4. Statistical Data Regarding Turkish WordNet (as of December 9th, 2003)

| FUNDEMENTALS | NUMBER OF OCCURRENCES |
|---|---|
| Synsets | 10350 |
| Literals | 14480 |
| Nonlexicalized Literals | 680 |
| Definitions | 4514 |
| Pos | 10350 |
| Ratio of literal/synset | 1,46 |

| SYNSET TYPE | NUMBER OF OCCURRENCES |
|---|---|
| BCS1 | 1218(100%) |
| BCS2 | 3471(100%) |
| BCS3 | 2877 (74,3%) |
| Noun | 7710(74,5%) |
| Verb | 2306(22,2%) |
| Adjective | 334(3,3%) |

| RELATION TYPE | NUMBER OF OCCURRENCES |
|---|---|
| Hypernym | 10034 |
| Holo_member | 208 |
| Holo_part | 1260 |
| Holo_portion | 162 |
| Causes | 96 |
| Be_in_state | 499 |
| Near_antonym | 1158 |
| Subevent | 119 |
| Also_see | 226 |
| Verb_group | 540 |
| Similar_to | 65 |
| Category_domain | 349 |
| Usage_domain | 5 |
| TOTAL | 14721 |

### 3.6.5. Ongoing Tasks

- During the progress meeting in Bucharest, the Consortium decided to use George Orwell's novel 1984 for wordnet validation purposes. An electronic version of the novel was not available, so, we scanned and OCR'ed the Turkish translation and aligned all sentences at the sentence level, using the alignment tool of the

TRADOS translation memory software. All Turkish sentences will be passed through our Turkish morphological analyzer and the resulting file will be sent to the Romanian partners with ambiguous POS tags, in line with the agreement reached in Bucharest.

- During the first two passes, our lexicographers could not find appropriate translations for 25.7% of BCS3 synsets. A third pass will be finished until 1 January 2004 and the number of untranslated synsets will be reduced as much as possible.

- Until today, our lexicographers decided that 680 synsets do not have an appropriate Turkish counterpart. Since we approach the final stages of the project, the validity of these decisions has to be checked. All Turkish synsets marked with the tag “non-lexicalized” will be reviewed and an attempt will be made to minimize their number until the end of the year.

- More than 5000 Turkish synsets lack a Turkish gloss. The assignment of glosses is a very time-consuming task and the task has therefore been spread over several months of the project.

# 4. Preparing the semantic cross-lingual validation of the monolingual wordnets

Semantic cross-lingual validation of the monolingual wordnets such as the ones in BalkaNet is defined as the checking of the inter-lingual alignments of the synsets in two or more wordnets. This type of validation assumes that the experts performing the task have very good command of the considered languages, and in order the validation be affected as least as possible by subjective judgment, we decided to use as additional source of knowledge the linguistic evidence as provided by a multilingual parallel corpus containing texts translated by professional translators. In principle, validation could be carried on for any pair of BalkaNet's languages or for any number of these languages, but we decided to consider the simplest case, namely the validation of pairs of wordnets, one for the native language of the experts and the other one for English. The parallel corpus is based on Orwell's novel "*Nineteen Eighty-Four*", containing 9 languages out of which 6 of these are in languages of interest for the Balkanet. For our experiments we selected, for the present moment, the English original plus translations in Bulgarian, Czech, Greek and Romanian. Currently, from the Serbian translation only half of it is available in the required format (tagged, lemmatized and sentence aligned to the English hub) and as such it provides insufficient data for statistical language processing. Unless the full version of Orwell's translation will be available in the appropriate format, the tests for Serbian will be carried under the reserve of less accurate results due to insufficient data.

The cross/lingual semantic validation is expected to pinpoint synsets alignment errors and incomplete synsets. An additional benefit from such a validation would be a word sense disambiguation (in terms of ILI labels) of the multilingual corpus for all the occurrences of the target evaluation words.

## *4.1 Interlingual Validation Based on Parallel Corpus Evidence*

If we take the position according to which word senses (language specific) represent language independent meanings, abstracted by ILI records, then the evaluation procedure

of wordnets interlingual alignment becomes straightforward: in a parallel text, words which are used to translate each other should have among their senses at least one pointing to the same ILI or to closely related ILIs. However, both in EuroWordNet and BalkaNet the ILI records are not structured, so we need to clarify what "closely related ILI" means. In the context of this research, we assume that the *hierarchy preservation* principle [4] holds true. This principle may be stated as follows:

*if in the language L1 two synsets* $M_1^{L1}$ *and* $M_2^{L1}$ *are linked by a (transitive) hierarchical relation H, that is* $M_1^{L1}$ $H^n$ $M_2^{L1}$ *and if* $M_1^{L1}$ *is aligned to the synset* $N_1^{L2}$ *and* $M_2^{L1}$ *is aligned to* $N_2^{L2}$ *of the language L2 then* $N_1^{L2}$ $H^m$ $N_2^{L2}$ *even if* $n \neq m$ *(chains of the H relation in the two languages could be of different lengths).* The difference in lengths could be induced by the existence of meanings in the chain of language L1 which are not lexicalized in language L2.

Under this assumption, we define the *relatedness* of two ILI records $R_1$ and $R_2$ as the *semantic similarity* between the synsets $Syn_1$ and $Syn_2$ of PWN that correspond to $R_1$ and $R_2$. A semantic similarity function SYM($Syn_1$, $Syn_2$) could be defined in many ways. We used a very simple and effective one: $SYM(Syn_1, Syn_2) = \frac{1}{1+N}$ where N is the number of oriented links traversed from one synset to the other or from the two synsets up to the closest common ancestor. One should note that every synset is linked (EQ-SYN) to exactly one ILI and that no two different synsets have the same ILI assigned to them. Furthermore, two ILI records $R_1$ and $R_2$ will be considered closely related if *semantic-similarity* ($Syn_1$, $Syn_2$) $\geq$ k, where k is an empirical threshold, depending on the monolingual wordnets and on the measure used for evaluating semantic distance.

Having a parallel corpus, containing texts in k+1 languages (T, $L_1$, $L_2$…$L_k$) and having monolingual wordnets for all of them, interlinked via an ILI-like structure, let us call T the target language and $L_1$, $L_2$…$L_k$ as source languages. The parallel corpus is encoded as a sequence of *translation units* (TU). A translation unit contains aligned sentences from each language, with tokens tagged and lemmatized as exemplified below (for details on encoding see http://nl.ijs.si/ME/V2/msd/html/):

**Table 1.** A partial translation unit from the parallel corpus

```
<tu id="Ozz.113">
 <seg lang="en">
   <s id="Oen.1.1.24.2"><w lemma="Winston" ana="Np">Winston</w>
     <w lemma="be" ana="Vais3s">was</w>          ...   </s>
 </seg>
 <seg lang="ro">
   <s id="Oro.1.2.23.2"><w lemma="Winston" ana="Np">Winston</w>
     <w lemma="fi" ana="Vmii3s">era</w>   ...   </s>
 </seg>
 <seg lang="cs">
   <s id="Ocs.1.1.24.2"><w lemma="Winston" ana="Np">Winston</w>
     <w lemma="se" ana="Px---d--ypn--n">si</w>  ...   </s>
  </seg>
  . . .
</tu>
```

We will refer to the wordnet for the target language as T-wordnet and to the one for the language $L_i$ as the i-wordnet. We use the following notations:

T_word = a target word, say $w_{TL}$;

$T\_word_j$ = the j-th occurrence of the target word;

$eq_{ij}$ = the translation equivalent (TE) for $T\_word_i$ in the source language $L_j$, say $w_{SLj}$; a pair ($w_{TL}$, $w_{SL}$) so that in a given context (a translation unit) $w_{TL}$ and $w_{SL}$ are reciprocal translations is called a translation pair (for the languages considered);

EQ = the matrix containing translations of the T_word (n occurrences, k languages):

**Table 2.** The translation equivalents matrix (EQ matrix)

| | $L_1$ | $L_2$ | … | $L_k$ |
|---|---|---|---|---|
| Occ #1 | $eq_{11}$ | $eq_{12}$ | … | $eq_{1k}$ |
| Occ #2 | $eq_{21}$ | $eq_{22}$ | … | $eq_{2k}$ |
| … | … | … | … | … |
| Occ #n | $eq_{n1}$ | $eq_{n2}$ | … | $eq_{nk}$ |

$TU_j$ = the translation unit containing $T\_word_j$;

$EQ_i$ = a vector, containing the TEs of T_word in language $L_i$: ($eq_{1i}$ $eq_{2i}$ …$eq_{ni}$)

More often than not the translation equivalents found for different occurrences of the target word are identical and thus identical words could appear in the $EQ_i$ vector. If $T\_word_j$ is not translated in the language $L_i$, then $eq_{ij}$ is represented by the null string. Every non-null element $eq_{ij}$ of the EQ matrix is subsequently replaced with the set of all

ILI identifiers that correspond to the senses of the word $eq_{ij}$ as described in the wordnet of the j-language. If this set is named $IS_{ij}$, we obtain the matrix EQ_ILI which is the same as EQ matrix except that it has an ILI set for every cell (Table 3).

**Table 3.** The matrix containing the senses for all translation equivalents (EQ_ILI matrix)

| | $L_1$ | $L_2$ | … | $L_k$ |
|---|---|---|---|---|
| Occ #1 | $IS_{11}$ ={$ILI_p$ \| $ILI_p$ identifies a synset of $eq_{11}$ } | $IS_{12}$ ={$ILI_p$ \| $ILI_p$ identifies a synset of $eq_{12}$ } | … | $IS_{1k}$ = {$ILI_p$ \| $ILI_p$ identifies a synset of $eq_{1k}$ } |
| Occ #2 | $IS_{21}$ ={$ILI_p$ \| $ILI_p$ identifies a synset of $eq_{21}$ } | $IS_{22}$ {$ILI_p$ \| $ILI_p$ identifies a synset of $eq_{22}$ } | … | $IS_{2k}$ = {$ILI_p$ \| $ILI_p$ identifies a synset of $eq_{2k}$ } |
| … | … | … | … | … |
| Occ #n | $IS_{n1}$ ={$ILI_p$ \| $ILI_p$ identifies a synset of $eq_{n1}$ } | $IS_{n2}$ {$ILI_p$ \| $ILI_p$ identifies a synset of $eq_{n2}$ } | | $IS_{nk}$ = {$ILI_p$ \| $ILI_p$ identifies a synset of $eq_{nk}$ } |

If some cells in EQ contain empty strings, then the corresponding cells in EQ_ILI will obviously contain empty sets. Similarly, we have for the T_word the list T_ILI = ($ILI_{T1}$ $ILI_{T2}$ … $ILI_{Tq}$).

The next step is to define our target data structure. Let us consider a new matrix, called VSA (Validation and Sense Assignment):

**Table 4.** The VSA matrix

| | $L_1$ | $L_2$ | … | $L_k$ |
|---|---|---|---|---|
| Occ #1 | $VSA_{11}$ | $VSA_{12}$ | … | $VSA_{1k}$ |
| Occ #2 | $VSA_{21}$ | $VSA_{22}$ | | $VSA_{2k}$ |
| … | … | … | … | … |
| Occ #n | $VSA_{n1}$ | $VSA_{n2}$ | … | $VSA_{nk}$ |

with $VSA_{ij}$ = T_ILI $\cap$ $IS_{ij}$ , if $IS_{ij}$ is non-empty and $\perp$ (undefined) otherwise.

The $i^{th}$ column of the VSA matrix provides valuable corpus-based information for the evaluation of the interlingual linking of the the i-wordnet and T-wordnet.

Ideally, computing for each line j the set $SA_j$ (sense assignment) as the intersection $ILI_{j1} \cap ILI_{j2} \ldots \cap ILI_{jk}$ one should get at a single ILI identifier: $SA_j=(ILI_{T\alpha})$, that is the $j^{th}$ occurrence of the target word was used in all source languages with the same meaning, represented interlingually by $ILI_{T\alpha}$. If this happened for any T_word, then the WSD problem (at least with the parallel corpora) would not exist. But this does not happen, and there are various reasons for it: the wordnets are partial and (even the PWN) are not perfect, the human translators are not perfect, there are lexical gaps between different languages, automatic extraction of translation equivalents is far from being perfect, etc.

Yet, for cross-lingual validation of interlinked wordnets the analysis of VSAs may offer wordnet developers extremely useful hints on senses and/or synsets missing in their wordnets, wrong ILI mappings of synsets, wrong human translation in the parallel corpus and mistakes in word alignment. Once the wordnets have been validated and corrected accordingly, the WSD (in parallel corpora) should be very simple. There are two ways of exploiting VSAs for validation:

**Vertical validation (VV)**: the development team of i-wordnet (native speakers of the language $L_i$ with very good command of the target language) will validate their own i-wordnet with respect to the T-wordnet, that is from all VSA matrixes (one for each target word) they would pay attention only to the $i^{th}$ column (the $VSA(L_i)$ vector).

**Horizontal validation (HV)**: for each VSA all SAs will be computed. Empty SAs could be an indication of ILI mapping errors still surviving in one or more wordnets (or could be explained by lexical gaps, wrong translations etc) and as such, the suspicious wordnet(s) might be re-validated in a focused way. The case of an SA containing more than a single ILI identifier could be explained by the possibility of having in all i-languages words with similar ambiguity.

Our system called WSDtool implements the methodology described above and offers an easy-to-use interface for the task of semantic validation. It incorporates the translation equivalents extraction system (TREQ&TREQ-AL, described in [Tufiş et al., 2003] as well as a graphic visualization of the two wordnets used in the validation process. We exemplify a horizontal WSDtool validation session by considering the En-Ro language pairs. The intersection between ILI sets of $w_{en}$ and $w_{ro}$ is presented in a table for every

occurrence of $w_{en}$ in the parallel corpus. The cell at line *i* (labeled with the translation unit identifier of the sentence containing the i[th] occurrence of $w_{en}$) and column labeled with the target language name (ro) contains the intersection of ILI sets of literals $w_{en}$ and $w_{ro}^{i}$ where $w_{ro}^{i}$ represents the Romanian translation for the *i*-th occurrence of $w_{en}$. The cell's content ranges over the next three cases:

1. the cell contains an ILI set; this means that each of the literals $w_{en}$ and $w_{ro}^{i}$ are found in synsets which are mapped onto the same ILIs. The user is required to choose the ILI which points to the correct sense in both languages (see figure 2). If such an ILI cannot be found, the user is offered another choice: to indicate the missing sense in the Romanian wordnet for the $w_{ro}^{i}$ literal. Finally, if all the senses of $w_{ro}^{i}$ are implemented, the user is asked to remap one of $w_{ro}^{i}$ synsets to satisfy the translation equivalence pair;

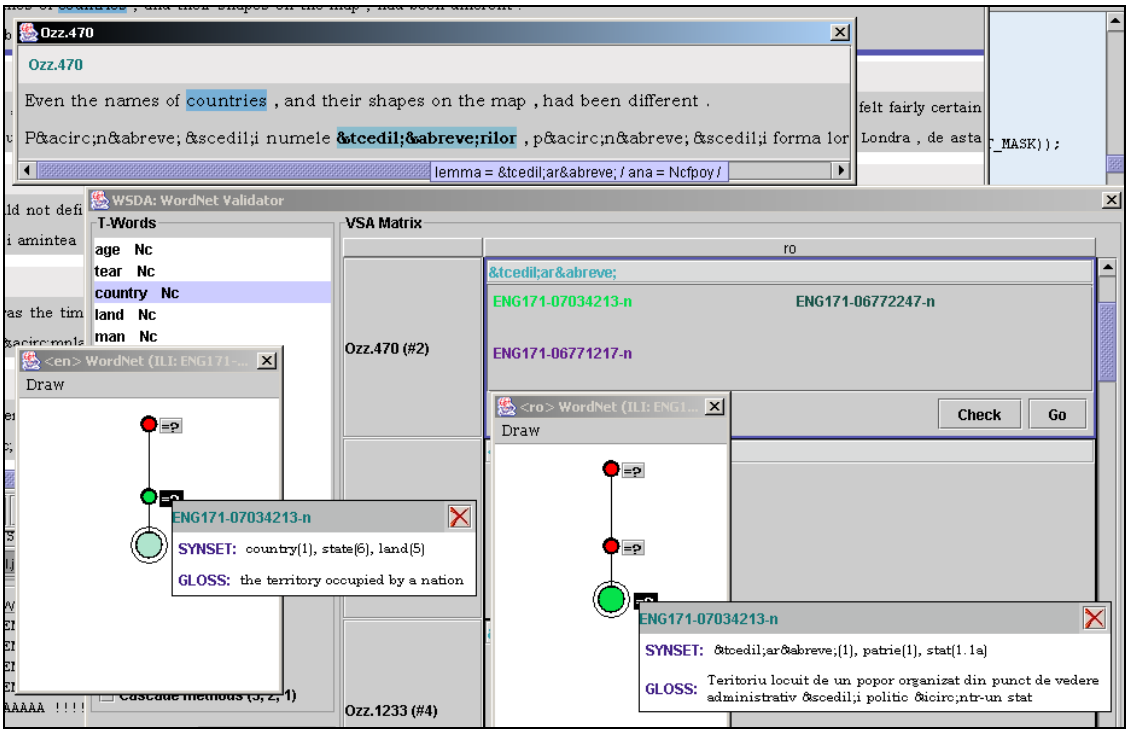



**Figure 2**

*The translation unit Ozz.470 contains the second occurrence of* $w_{en}$ ***'country'****. This occurrence is translated in Romanian by* $w_{ro}^{2}$ ***'țară'*** *( SGML entities encoding:* ***'ţară'****) and we can see that the selected table cell contains the ILI set of the intersection. In this case, ILI171-07034213-n is the identifier for the correct sense in both Romanian and English*

2. the cell contains pairs of ILIs; each pair ends with a real number denoting a similarity measure between the members of the pair; the similarity measure was calculated as $\delta_N = \frac{1}{1+N}$ where $N$ is the number of links between the pair members in the PWN hierarchy (it is easily seen that when $N = 0$, $\delta_0 = 1$ which means that the two ILIs are identical; for $N = 1$, $\delta_1 = 0.5$ which shows an HH relationship or a coordination between pair members); all pairs in the interval $[\delta_2, \delta_0]$ were retained. The user is now required to choose the pair which reflects the best HH relation between pair members ('the best' means that the pair member corresponding to $w_{en}$ should reflect the sense used – see figure 3). If such a pair does not exist, the preceding actions (from 1.) are to be followed;

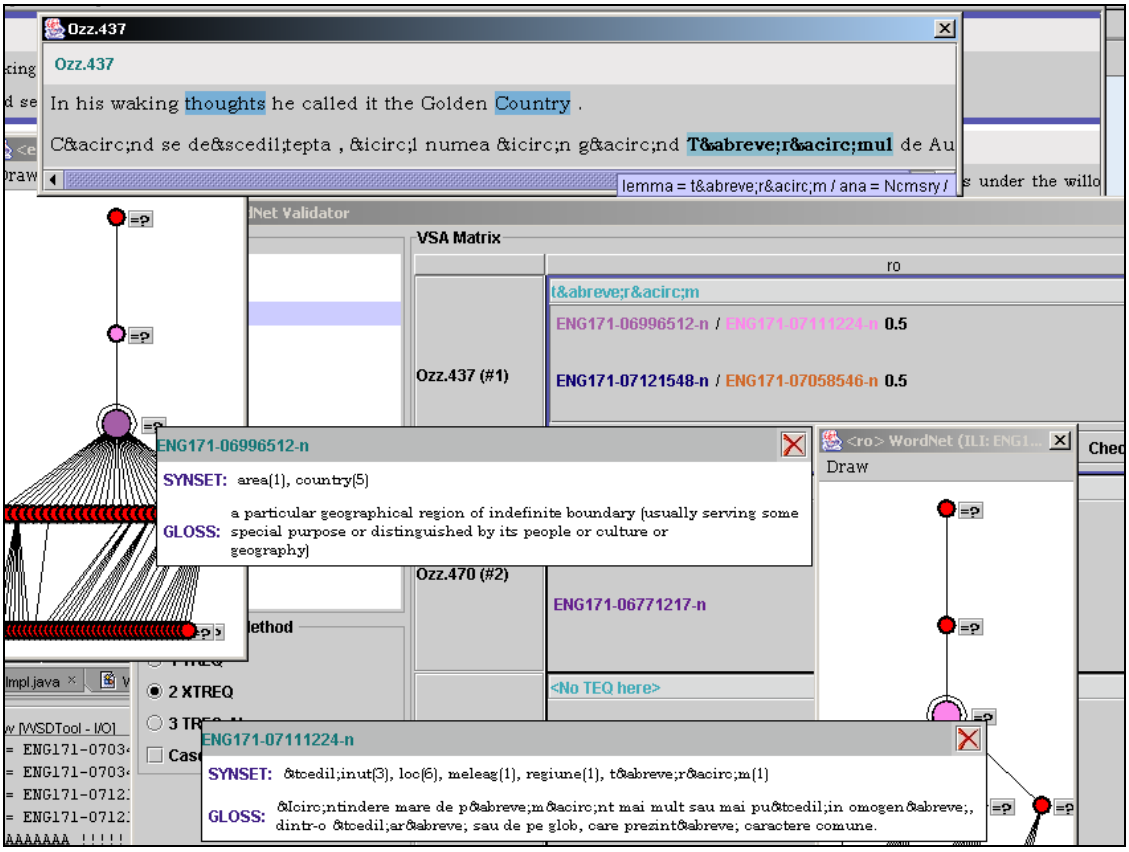



**Figure 3**

*The selected cell ( Ozz.437(#1), ro ) reflects the ILI intersection between **'country'** and **'tărâm'** (SGML entities notation: 'tărâm'). As none of the corresponding ILIs are the same, the cell presents two pairs of ILIs between which $\delta_N$ is maximal (0.5, with $N = 1$). In this case the first pair is correct.*

3. the cell is empty; this is a potential alignment error in the Romanian wordnet or an incomplete Romanian synset (see figure 4). If $(w_{en}, w_{ro}^{i})$ is a correct translation pair, then one of the following must hold: the relevant $w_{ro}^{i}$ synset is wrongly mapped, the sense of the $i^{th}$ occurrence of $w_{en}$ is not yet implemented for the corresponding translation equivalent literal $w_{ro}^{i}$ (see figure 5) or the literal $w_{ro}^{i}$ does not belong to the relevant Romanian synset. If the latter case holds, the user is asked to add the literal (with the appropriate sense number) to the correct synset (this way, synset expanding can be achieved in a focused way: context study).

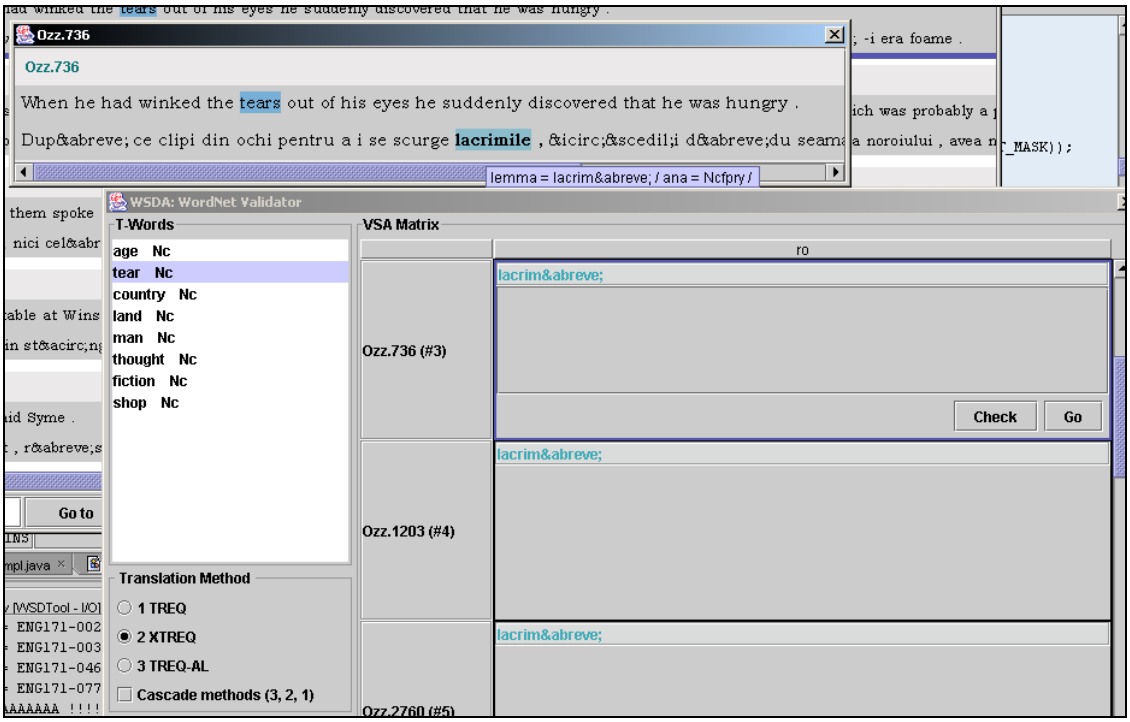



**Figure 4**

*The cell at ( Ozz.736(#3), ro ) is empty. The third occurrence of **'tear'** was translated by **'lacrimă'** (SGML entities notation: 'lacrimă') and this is a correct translation pair.*

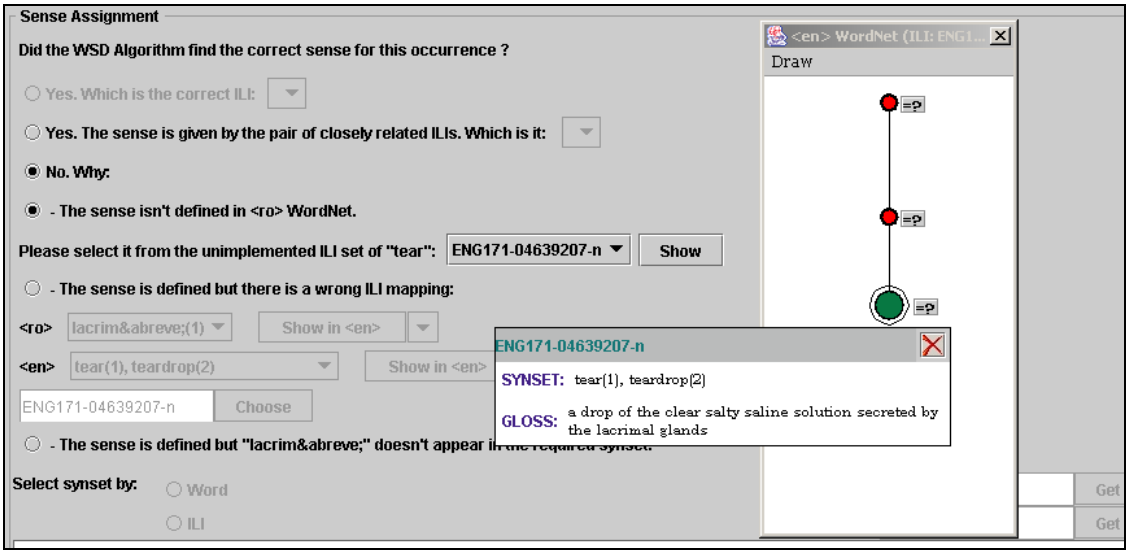



**Figure 5**

*The reason for the void intersection above is that **'tear'** was used in a sense that is not implemented in Romanian wordnet. The figure shows a portion of the check window where the user specifies that this sense of **'tear'** is not implemented in the current version of the Romanian wordnet*

## *4.2 The next step for cross-lingual validation of the BalkaNet wordnets*

Since the BalkaNet wordnets are partial (with number of synsets ranging between 4500 to 25,000) it is obvious that in the parallel corpus there might be words for which some or even all senses are missing from each monolingual wordnet. Therefore, in order to get meaningful results for the vertical evaluations of different pairs of wordnets (EN-XX), one has to select a bag of English target words with the property that all their senses are labeled with ILI numbers in the set of commonly agreed set of concepts. This approach is feasible among the time-span of the project and does not assume creating too many new synsets besides the already implemented. The disadvantage is that the wordnets will be semantically validated only partially (for the senses used in the corpus of the selected bag of words) and consequently only the target words and their translation equivalents in the other languages of the project will be sense disambiguated. Another approach would be to extract the ILI numbers pertaining to all content words in the English part of the parallel corpus and all the missing concepts be implemented by all partners. This approach assumes a lot of work on each partner in order to extend their wordnets so that

to cover the integral text in the parallel corpus. Although this is not feasible within the remaining time and budget of the current project this goal could be a goal for future developments of our wordnets, either in a concerted way (in a follow-up of this project) or on an individual basis, for some of the monolingual wordnets.

The procedure for identifying the bag of English words to be used for vertical semantic evaluations is the following:

- extract all lemmas for the English verbs and nouns occurring in "*1984"*such as all their senses are labeled as BCS 1 or BCS 2 or BCS 3 (these concepts are supposed to be implemented by all wordnets except for the Serbian one which was subject to implement the BCS1 but implemented also BCS2; in this case there will be considered only a subset of the bag of words, namely those that were used in the corpus with senses in BCS1 and BCS2- this information is supposed to be clarified when all the other language wordnets were validated and the translation equivalents of the target words in the respective monolingual texts of the parallel corpus were sense disambiguated);

The bag of target words thus selected contains 530 English words which every partner may use for the vertical semantic validation against the PWN. The bag of words with all their senses in BCS1, 2 or 3 is given in the APPENDIX 1.

To identify the concepts that might be used in the entire corpus, but are not implemented in a monolingual wordnet, the procedure can be summarized as follows:

- extract all lemmas for the English verbs and nouns occurring in "*1984"*;
- collect the ILI numbers of all these words as the full ILI-validation_set;
- eliminate from the full ILI-validation_set all the ILIs in a monolingual WN and thus obtain the set of *would-be-implemented* ILIs.

For the Romanian wordnet our *would-be-implemented* ILIs contains 2312 ILIs out of which we already implemented 1000 synsets.

# 4. Conclusions

The quantitative evaluation of the cross-lingual coverage of the monolingual wordnets uploaded on the BalkaNet information server is described in the following tables, considering different clusters of languages:

**Intersection of ILI's (two languages)**

| Language | Romanian® | Bulgarian(B) | Greek (G) | Turkish (T) | Serbian (S) | Czech (C) |
|---|---|---|---|---|---|---|
| Romanian | - | 11489 | 7336 | 8171 | 4646 | 12391 |
| Bulgarian | | - | 7250 | 8143 | 4659 | 12682 |
| Greek | | | - | 6459 | 4363 | 8871 |
| Turkish | | | | - | 4590 | 8755 |
| Serbian | | | | | - | 4649 |
| Czech | | | | | | - |

**Intersection of ILI's (three languages) :**

| Language | BG | BT | BS | BC | GT | GS | GC | TS | TC | SC |
|---|---|---|---|---|---|---|---|---|---|---|
| Romanian | 6865 | 7991 | 4632 | 10688 | 5965 | 4352 | 7031 | 4580 | 8046 | 4620 |
| Bulgarian | | | | | 5917 | 4352 | 6980 | 4580 | 8060 | 4623 |
| Greek | | | | | | | | 4329 | 6041 | 4347 |
| Turkish | | | | | | | | | | 4582 |

**Intersection of ILI's (four languages):**

| Language | BGT | BGS | BGC | BTS | BTC | BSC | GTS | GTC | GSC | TSC |
|---|---|---|---|---|---|---|---|---|---|---|
| Romanian | 5896 | 4350 | 6712 | 4572 | 7934 | 4610 | 4329 | 5909 | 4344 | 4573 |
| Bulgarian | | | | | | | 4329 | 5890 | 4345 | 4574 |
| Greek | | | | | | | | | | 4329 |

**Intersection of ILI's (five languages):**

| Language | BGTS | BGTC | BGSC | BTSC | GTSC |
|---|---|---|---|---|---|
| Romanian | 4329 | 5871 | 4343 | 4567 | 4329 |
| Bulgarian | | | | | 4329 |

**All language intersection:**
**RBGST =4329.**

**BCS statistics:**

| Language | BCS 1 | BCS 2 | BCS 3 | BCS final | |
|---|---|---|---|---|---|
| ILI database | 1218 | 3471 | 3827 | 8516 | |
| Romanian | 1218 | 3471 | 3795 | 8484* | |
| Bulgarian | 1218 | 3471 | 3827 | 8516 | |
| Greek | 1218 | 3463 | 1252 | 5933 | |
| Turkish | 1218 | 3471 | 2923 | 7611 | |
| Serbian | 1211 | 2945 | 382 | 4538 | |
| Czech | 1218 | 3471 | 3506 | 8195 | |

*We have a number of 608 nonlexicalized concepts

# POS statistics

| Language | Nouns | Verbs | Adjectives | Adverbs |
|---|---|---|---|---|
| Romanian | 10.716 (~72%) | 2927 (~20%) | 844(~6%) | 200(~1%) |
| Bulgarian | 11037 (~73%) | 3317(~22%) | 653 (~4%) | 0 |
| Greek | 12494(~79%) | 2921 (~18%) | 352 (~2%) | 14 (~0.1%) |
| Turkish | 7710(~74%) | 2306(~22%) | 334(3%) | 0 |
| Serbian | 3139(~65%) | 1471(~30%) | 154(~3%) | 7 (~0.1%) |
| Czech | 19286(~72%) | 4950(~18%) | 2128(~8%) | 164(~0.6%) |

**Other statistics:**

| | **Duplicate ILI** | **Not Well-formed synsets** | **Relations that should not be imported from PWN** | **Dangling Nodes*** | **Dangling Relations** | **Literals in Conflict** |
|---|---|---|---|---|---|---|
| **Romanian** | 0 | 0 | no | 58 | 0 | 0 |
| **Turkish** | 0 | 5182 | maybe** | 71 | 53 | 1523 |
| **Serbian** | 3 | 761 | maybe** | 82 | 2 | 151 |
| **Bulgarian** | 0 | 0 | maybe** | 19 | 0 | 48 |
| **Greek** | 0 | 30 | no | 2465 | 0 | 1191 |
| **Czech** | 0 | 25665 | no | 2561 | 0 | 68 |

* Adverbial synsets are not included in this statistics since they do not have a relational structure in BalkaNet.
** The relations *region-domain, usage-domain, particle* and *eng-derivative* should be manually checked to see if they pertain for the languages in case; if this is the case, they should be renamed as <lg>-region-domain, <lg>-usage-domain <lg>-participle and <lg>-derivative (as was done in the Bulgarian wordnet)

**Duplicate ILI** ------ number of the ILI's labeling more than one synset; in the error log file these ILIs are listed one per line
**ill-formed synsets** -----the number of synsets the structure of which is not conformant with the prescribed format. The error log lists for each ill-formed synsets the errors encountered in the respective synset. For example the following line shows a synset in a wordnet which has no ILI number, no pos value and no gloss.

```
no ID\no pos\no sense\noGloss\
<SYNSET><ID></ID>
   <SYNONYM>
     <LITERAL><SENSE></SENSE>
            <LNOTE>nema</LNOTE>
     </LITERAL>
   </SYNONYM>
   <POS></POS>
   <STAMP></STAMP>
</SYNSET>
```

**relations that should not be imported from PWN** ------these are relations that were introduces in WordNet2.0 that are language specific in PWN and should not be subject to automatic import.

1. *eng_derivative*. The semantics of the relation is that it links nouns and verbs that are related morphologically (in English of course).
   This is a language specific and it was accordingly prefixed (as in bg_derivative)
2. *region_domain*. It is related with the area where a specific word with a particular sense is used (language depended). When used in a specific wordnet (other than PWN) is should designate areas where the literals in the respective synsets are used.
3. usage_domain (language dependent)

examples:

*potted 3* **region domain** *is United Kingdom, UK, Great Britain, GB, Britain, United Kingdom of Great Britain and Northern Ireland*

The *particle* relation existed before in the VISDIC representation of the PWN1.7.1 but actually this should be named as in the original **participle.** It is also language dependent. For example adsorbing is **participle** of the verb adsorb.

**dangling nodes** --- the number of **dangling nodes** (nodes that have no link with other nodes); in the error log file they are listed one per line.

**dangling relations** --- the number of **dangling relations** (see the definition above) ; in the error log file they are listed one per line.

Example:

*Dangling:ENG20-00165384-v(hypernym)ENG20-00198579-v*

According to the definition we gave before, if an outgoing link is specified for a synset, the incoming synset of that relation should be also implemented. This example shows
In the error log file, each line always signals a missing incoming synset of a given relation outgoing from a specific synset.
In the example above, hypernym relation starting from *ENG20-00165384-v* is dangling because its arrival synset (*ENG20-00198579-v)* is missing.

**Literals in conflict** ---- number of literals appearing in multiple synsets with the same sense identifier. In the error log file, for every pair <literal sense> that appears in more that one synset, the list of the ILIs assigned to the respective synsets is generated:
*Example:*
*potreba@@3 ENG20-13629894-n ENG20-13630974-n*

The line says that the word potreba with the sense 3 is present in ENG20-13629894-n and ENG20-13630974-n.

**Statistics of the relations used by each monolingual wordnet**

Bulgarian

| | |
|---|---|
| hypernym | 14300 |
| bg_derivative | 6379 |
| near_antonym | 1392 |
| holo_part | 998 |
| verb_group | 851 |
| holo_member | 771 |
| category_domain | 617 |
| be_in_state | 541 |
| also_see | 269 |
| derived | 256 |
| subevent | 150 |
| causes | 104 |
| holo_portion | 102 |
| similar_to | 40 |
| particle | 22 |
| usage_domain | 22 |
| region_domain | 1 |

Czech

| | |
|---|---|
| hypernym | 22262 |
| holo_part | 1742 |
| near_antonym | 1720 |
| similar_to | 1138 |
| category_domain | 1029 |
| verb_group | 916 |
| also_see | 762 |
| be_in_state | 602 |
| holo_portion | 357 |
| holo_member | 250 |
| subevent | 217 |
| causes | 117 |

Greek

| | |
|---|---|
| hypernym | 12308 |
| holo_part | 1763 |
| holo_member | 334 |
| near_antonym | 287 |
| holo_substance | 59 |
| antonym | 44 |

Romanian

| | |
|---|---|
| hypernym | 13669 |
| near_antonym | 1476 |
| holo_part | 1007 |
| similar_to | 896 |
| verb_group | 888 |
| holo_member | 778 |
| be_in_state | 546 |
| category_domain | 508 |
| also_see | 333 |
| subevent | 139 |
| holo_portion | 107 |
| causes | 106 |
| derived | 28 |

Turkish

| | |
|---|---|
| hypernym | 10034 |
| holo_part | 1260 |
| near_antonym | 1158 |
| verb_group | 540 |
| be_in_state | 499 |
| category_domain | 349 |
| also_see | 226 |
| holo_member | 208 |
| holo_portion | 162 |
| subevent | 119 |
| causes | 96 |
| similar_to | 65 |
| usage_domain | 5 |
| derived | 1 |

Serbian

| | |
|---|---|
| hypernym | 4399 |
| srb_derivative | 1881 |
| near_antonym | 364 |
| holo_part | 249 |
| category_domain | 167 |
| verb_group | 137 |
| also_see | 99 |
| be_in_state | 90 |
| holo_member | 69 |
| derived | 66 |
| subevent | 58 |
| causes | 44 |
| holo_portion | 21 |
| similar_to | 10 |
| particle | 9 |
| usage_domain | 1 |

APPENDIX 1:
The list words, occurring in the parallel corpus, all senses of which belong to BCSs 1, 2 or 3

| WORD | POS | # OF SENSES |
|---|---|---|
| course | n | 8 |
| lie | v | 7 |
| wish | v | 7 |
| portion | n | 6 |
| unit | n | 6 |
| country | n | 5 |
| part | v | 5 |
| happen | v | 5 |
| search | n | 5 |
| structure | n | 5 |
| party | n | 5 |
| concern | n | 5 |
| beginning | n | 5 |
| commit | v | 5 |
| device | n | 5 |
| like | v | 5 |
| increase | n | 5 |
| effort | n | 4 |
| measure | v | 4 |
| paint | v | 4 |
| balance | v | 4 |
| transmit | v | 4 |
| disc | n | 4 |
| require | v | 4 |
| win | v | 4 |
| shout | v | 4 |
| amount | n | 4 |
| intend | v | 4 |
| include | v | 4 |
| people | n | 4 |
| station | n | 4 |
| store | n | 4 |
| behaviour | n | 4 |
| market | n | 4 |
| danger | n | 4 |
| promise | v | 4 |
| year | n | 4 |
| demonstrate | v | 4 |
| leadership | n | 4 |
| relationship | n | 4 |
| describe | v | 4 |
| perform | v | 4 |
| path | n | 4 |
| forget | v | 4 |
| competition | n | 4 |
| replace | v | 4 |
| destruction | n | 3 |
| flatten | v | 3 |
| improvement | n | 3 |
| need | v | 3 |
| ache | v | 3 |
| heap | n | 3 |
| choice | n | 3 |
| money | n | 3 |
| affair | n | 3 |
| prize | n | 3 |
| universe | n | 3 |

| WORD | POS | # SENSES |
|---|---|---|
| hardship | n | 3 |
| disagreement | n | 3 |
| supply | n | 3 |
| chance | v | 3 |
| struggle | n | 3 |
| chest | n | 3 |
| polish | v | 3 |
| hurry | v | 3 |
| slide | v | 3 |
| experience | n | 3 |
| intellect | n | 3 |
| tin | n | 3 |
| fate | n | 3 |
| town | n | 3 |
| shut | v | 3 |
| educate | v | 3 |
| satisfy | v | 3 |
| comprehend | v | 3 |
| scratch | v | 3 |
| harm | n | 3 |
| encourage | v | 3 |
| week | n | 3 |
| rinse | v | 3 |
| crumble | v | 3 |
| battle | n | 3 |
| rub | v | 3 |
| smell | v | 3 |
| boundary | n | 3 |
| disorder | n | 3 |
| luck | n | 3 |
| marry | v | 2 |
| persuade | v | 2 |
| hostel | n | 2 |
| saloon | n | 2 |
| shudder | v | 2 |
| effect | v | 2 |
| goodness | n | 2 |
| neighbourhood | n | 2 |
| team | n | 2 |
| mutter | v | 2 |
| judge | n | 2 |
| remark | v | 2 |
| being | n | 2 |
| soldier | n | 2 |
| mine | n | 2 |
| atom | n | 2 |
| slaughter | v | 2 |
| grasp | v | 2 |
| message | n | 2 |
| weapon | n | 2 |
| swarm | v | 2 |
| accumulate | v | 2 |
| route | n | 2 |
| robe | n | 2 |
| murmur | v | 2 |
| childhood | n | 2 |

APPENDIX 1:
The list words, occurring in the parallel corpus, all senses of which belong to BCSs 1, 2 or 3

| WORD | POS | # OF SENSES | WORD | POS | # OF SENSES |
|---|---|---|---|---|---|
| task | n | 2 | munition | n | 2 |
| conduct | n | 2 | ointment | n | 2 |
| dry | v | 2 | lamp | n | 2 |
| refrain | v | 2 | succeed | v | 2 |
| soothe | v | 2 | whole | n | 2 |
| increase | v | 2 | forest | n | 2 |
| consciousness | n | 2 | apple | n | 2 |
| crisis | n | 2 | profit | v | 2 |
| regain | v | 2 | risk | v | 2 |
| improve | v | 2 | discussion | n | 2 |
| mentality | n | 2 | conviction | n | 2 |
| prison | n | 2 | instance | n | 2 |
| extent | n | 2 | cause | v | 2 |
| weary | v | 2 | cost | v | 2 |
| exist | v | 2 | swarm | n | 2 |
| bathroom | n | 2 | approve | v | 2 |
| confer | v | 2 | residue | n | 2 |
| prevent | v | 2 | carelessness | n | 2 |
| discrimination | n | 2 | ruler | n | 2 |
| accomplish | v | 2 | forbid | v | 2 |
| passageway | n | 2 | symbol | n | 2 |
| estimate | v | 2 | religion | n | 2 |
| imagine | v | 2 | certainty | n | 2 |
| hat | n | 2 | fluid | n | 2 |
| chief | n | 2 | expend | v | 2 |
| month | n | 2 | wound | v | 2 |
| bottle | n | 2 | bore | v | 2 |
| accident | n | 2 | comfort | v | 2 |
| last | v | 2 | swim | v | 2 |
| emphasize | v | 2 | din | n | 2 |
| attempt | n | 2 | bread | n | 2 |
| characterize | v | 2 | uncover | v | 2 |
| existence | n | 2 | army | n | 2 |
| happiness | n | 2 | musician | n | 2 |
| uncertainty | n | 2 | mouse | n | 2 |
| hammer | v | 2 | adapt | v | 2 |
| metal | n | 2 | ability | n | 2 |
| pronounce | v | 2 | morality | n | 2 |
| zip | n | 2 | disconcert | v | 2 |
| rebelliousness | n | 2 | human | n | 2 |
| mend | v | 2 | entrust | v | 2 |
| pause | n | 2 | aeroplane | n | 1 |
| urinate | v | 2 | pub | n | 1 |
| owner | n | 2 | fanaticism | n | 1 |
| island | n | 2 | roam | v | 1 |
| committee | n | 2 | unpack | v | 1 |
| proliferate | v | 2 | dirty | v | 1 |
| stupidity | n | 2 | kind | n | 1 |
| crowd | n | 2 | fireplace | n | 1 |
| emblem | n | 2 | trousers | n | 1 |
| drip | v | 2 | ignorance | n | 1 |
| cease | v | 2 | delude | v | 1 |
| accord | v | 2 | underclothes | n | 1 |
| meaning | n | 2 | chunk | n | 1 |
| railway | n | 2 | fidget | v | 1 |
| individual | n | 2 | trumpet | n | 1 |
| status | n | 2 | murder | n | 1 |

APPENDIX 1:
The list words, occurring in the parallel corpus, all senses of which belong to BCSs 1, 2 or 3

| WORD | POS | # OF SENSES | WORD | POS | # OF SENSES |
|---|---|---|---|---|---|
| journey | n | 1 | machine gun | n | 1 |
| urinal | n | 1 | cooking | n | 1 |
| postpone | v | 1 | citizen | n | 1 |
| animal | n | 1 | hatred | n | 1 |
| scientist | n | 1 | artist | n | 1 |
| weather | n | 1 | dwelling | n | 1 |
| squeak | v | 1 | dwelling house | n | 1 |
| detect | v | 1 | own | v | 1 |
| rodent | n | 1 | leather | n | 1 |
| long | v | 1 | astonishment | n | 1 |
| projectile | n | 1 | recollect | v | 1 |
| liken | v | 1 | shirt | n | 1 |
| disprove | v | 1 | cliff | n | 1 |
| corpse | n | 1 | rivalry | n | 1 |
| rival | n | 1 | nostalgia | n | 1 |
| select | v | 1 | sunlight | n | 1 |
| loathe | v | 1 | wade | v | 1 |
| briefcase | n | 1 | airfield | n | 1 |
| saucepan | n | 1 | slope | v | 1 |
| pantry | n | 1 | expert | n | 1 |
| explosive | n | 1 | wriggle | v | 1 |
| squirm | v | 1 | bakery | n | 1 |
| archipelago | n | 1 | staircase | n | 1 |
| grandfather | n | 1 | ancestor | n | 1 |
| porch | n | 1 | inflict | v | 1 |
| water closet | n | 1 | drug | n | 1 |
| attendance | n | 1 | thank | v | 1 |
| nakedness | n | 1 | convince | v | 1 |
| tennis | n | 1 | awake | v | 1 |
| buttock | n | 1 | grovel | v | 1 |
| coin | n | 1 | compete | v | 1 |
| purchase | v | 1 | nonexistence | n | 1 |
| lifetime | n | 1 | dustbin | n | 1 |
| questioning | n | 1 | hallway | n | 1 |
| emotion | n | 1 | disgrace | n | 1 |
| persevere | v | 1 | cosmetics | n | 1 |
| opportunity | n | 1 | proprietor | n | 1 |
| laugh | v | 1 | matter | v | 1 |
| armchair | n | 1 | mineral | n | 1 |
| military | n | 1 | commodity | n | 1 |
| actuality | n | 1 | doorway | n | 1 |
| mattress | n | 1 | rely | v | 1 |
| sanity | n | 1 | sailing ship | n | 1 |
| sky | n | 1 | orifice | n | 1 |
| frock | n | 1 | revolt | n | 1 |
| entertainment | n | 1 | hate | n | 1 |
| exploit | n | 1 | garment | n | 1 |
| motion | v | 1 | roughen | v | 1 |
| unconsciousness | n | 1 | table tennis | n | 1 |
| footpath | n | 1 | summer | n | 1 |
| chew | v | 1 | dice | n | 1 |
| offensive | n | 1 | whisper | v | 1 |
| incredulity | n | 1 | flee | v | 1 |
| spyhole | n | 1 | tribunal | n | 1 |
| praise | v | 1 | tinkle | v | 1 |
| misdemeanour | n | 1 | disseminate | v | 1 |
| produce | n | 1 | police | n | 1 |

APPENDIX 1:
The list words, occurring in the parallel corpus, all senses of which belong to BCSs 1, 2 or 3

| WORD | POS | # OF SENSES |
|---|---|---|
| achieve | v | 1 |
| despair | v | 1 |
| whimper | v | 1 |
| parachute | n | 1 |
| disguise | v | 1 |
| humiliate | v | 1 |
| furniture | n | 1 |
| clock | n | 1 |
| calamity | n | 1 |
| poem | n | 1 |
| parent | n | 1 |
| winter | n | 1 |
| refrigerator | n | 1 |
| swine | n | 1 |
| poverty | n | 1 |
| bicycle | n | 1 |
| stair | n | 1 |
| hiding place | n | 1 |
| shoelace | n | 1 |
| disgust | n | 1 |
| hate | v | 1 |
| trickle | v | 1 |
| resemble | v | 1 |
| wife | n | 1 |
| discard | v | 1 |
| knowledge | n | 1 |
| love affair | n | 1 |
| mankind | n | 1 |
| persecution | n | 1 |
| notice board | n | 1 |
| truncheon | n | 1 |
| razor | n | 1 |
| cloth | n | 1 |
| factory | n | 1 |
| saw | v | 1 |
| adherent | n | 1 |
| recurrence | n | 1 |
| syringe | n | 1 |
| cigarette | n | 1 |
| anodyne | n | 1 |
| prisoner | n | 1 |
| shrub | n | 1 |
| insanity | n | 1 |
| supersede | v | 1 |
| yap | v | 1 |
| obey | v | 1 |
| disobey | v | 1 |
| desk | n | 1 |
| punish | v | 1 |
| lighthouse | n | 1 |
| retaliation | n | 1 |
| effigy | n | 1 |
| gaze | v | 1 |
| corridor | n | 1 |
| ship | n | 1 |
| ascribe | v | 1 |
| selfishness | n | 1 |

| WORD | POS | # OF SENSES |
|---|---|---|
| fortress | n | 1 |
| convict | v | 1 |
| sticking plaster | n | 1 |
| feed | n | 1 |
| prostitution | n | 1 |
| conversation | n | 1 |
| muse | v | 1 |
| pillow | n | 1 |
| grandmother | n | 1 |
| fright | n | 1 |
| mayor | n | 1 |
| victory | n | 1 |
| enroll | v | 1 |
| daughter | n | 1 |
| protector | n | 1 |
| method | n | 1 |
| slap | v | 1 |
| friendship | n | 1 |
| funeral | n | 1 |
| furnace | n | 1 |
| inhabitant | n | 1 |
| amputate | v | 1 |
| crinkle | n | 1 |
| demeanour | n | 1 |
| breathing | n | 1 |
| periodical | n | 1 |
| concrete | n | 1 |
| helicopter | n | 1 |
| ankle | n | 1 |
| haunt | n | 1 |
| syllable | n | 1 |
| pistol | n | 1 |
| salary | n | 1 |
| embezzlement | n | 1 |
| infant | n | 1 |
| gramme | n | 1 |
| denture | n | 1 |
| doctrine | n | 1 |
| wipe | v | 1 |
| lettering | n | 1 |
| pendulum | n | 1 |
| flower | v | 1 |
| clothing | n | 1 |
| ugliness | n | 1 |
| brooch | n | 1 |
| insurrection | n | 1 |
| stitch | v | 1 |
| intellectual | n | 1 |
| ladle | n | 1 |
| kitchen | n | 1 |
| paraphernalia | n | 1 |
| gabble | v | 1 |
| sandwich | n | 1 |
| hint | v | 1 |
| utterance | n | 1 |
| district | n | 1 |
| annihilate | v | 1 |

APPENDIX 1:
The list words, occurring in the parallel corpus, all senses of which belong to BCSs 1, 2 or 3

| WORD | POS | # OF SENSES |
|---|---|---|
| wrist | n | 1 |
| perish | v | 1 |
| lingua | n | 1 |
| bookcase | n | 1 |
| disbelieve | v | 1 |
| reflex | n | 1 |
| achievement | n | 1 |
| bulge | n | 1 |
| rove | v | 1 |
| gymnastics | n | 1 |
| happening | n | 1 |
| stroll | v | 1 |
| gratitude | n | 1 |
| trolley | n | 1 |
| photograph | n | 1 |
| blowlamp | n | 1 |
| therapy | n | 1 |
| dislike | v | 1 |
| uselessness | n | 1 |
| lack | n | 1 |
| affection | n | 1 |
| directive | n | 1 |
| reptile | n | 1 |
| bookshelf | n | 1 |
| weep | v | 1 |
| writhe | v | 1 |
| gambling | n | 1 |
| battlefield | n | 1 |
| surname | n | 1 |
| waste pipe | n | 1 |
| ant | n | 1 |
| chisel | n | 1 |
| equipment | n | 1 |
| ampoule | n | 1 |
| enrol | v | 1 |
| lawyer | n | 1 |
| amplifier | n | 1 |
| credulity | n | 1 |
| toil | v | 1 |
| familiarize | v | 1 |
| partisanship | n | 1 |
| poet | n | 1 |
| household | n | 1 |
| cattle | n | 1 |
| vomit | v | 1 |
| uniform | n | 1 |
| guardian | n | 1 |
| statue | n | 1 |
| overhear | v | 1 |
| repeat | n | 1 |
| firearm | n | 1 |
| jew | n | 1 |
| popularity | n | 1 |
| handle | n | 1 |
| lack | v | 1 |
| singlet | n | 1 |
| stimulus | n | 1 |
| museum | n | 1 |
| ridicule | v | 1 |
| fighting | n | 1 |
| insult | v | 1 |
| disease | n | 1 |
| civilian | n | 1 |
| pigeon | n | 1 |
| gesticulate | v | 1 |
| tremble | v | 1 |
| feat | n | 1 |
| creak | v | 1 |
| punishment | n | 1 |
| husband | n | 1 |
| relevance | n | 1 |
| scuttle | v | 1 |
| sheaf | n | 1 |
| concept | n | 1 |
| morals | n | 1 |