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

APRIL 25, 2004

Deliverable D6.2: The Final monolingual WordNets for each of the Balkan languages

Authors	{Dan Cristea, Georgiana Puscasu, Oana Postolache} **UAIC** {Harry Kornilakis, Maria Grigoriadou, Eleni Galiotou, Evangelos Papakitsos} **UOA** {Sofia Stamou} **DBLAB** {Cvetana Krstev, Gordana Pavlovic-Lazetic, Ivan Obradovic, Dusko Vitas} **MATF** {Ozlem Cetinoglu} **SABANCI** {Dan Tufis, Eduard Barbu, Verginica Mititelu, Luigi Bozianu, Radu Ion} **RACAI** {Karel Pala, Pavel Smrz, Anna Sinopalnikova } **FI MU** {Svetla Koeva} **DCMB** {George Totkov} **PU**
EC Project Officer	Erwin Valentini
Project Coordinator	Prof. Dimitris Christodoulakis ***Director, DBLAB*** Computer Engineering and Informatics Department Patras University GR-265 00 Greece Phone: +30 610 960385 Fax: +30 610 960438 e-mail: dxri@cti.gr
Keywords	Validation, final monolingual wordnet
Actual Distribution	Project Consortium, Project Officer, EC

Table of Contents

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 that 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 issues 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) belong 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

Enrichment of the Bulgarian WordNet

The last stage of the development of Bulgarian WordNet is directed both to its enrichment and validation. Our team has been already finished the work connected with the implementation of the common Base concepts – set one, two and three. That is why we concentrated our efforts to the covering as many senses as possible of the words in *1984*. We were following the procedure in which: all the nouns, verbs, and adjectives have been extracted from the English and Bulgarian versions of *1984*; all the English and Bulgarian synsets which contain these words were generated; then the English synsets that were not mapped onto Bulgarian synsets were identified; then the senses of the English synsets were compared with the real senses used in *1984* text; and finally we implemented most of the missing senses. The table below gives some statistical illustrations.

	Nouns	Verbs	Adj	Total
1984 literals	16142	14620	5486	36248
1984 unique words	2937	2361	1480	6778
BulNet literals	9412	4952 (6324)	2014	16378
BulNet unique words	2652	1492 (1932)	1179	5323
% of implemented unique words	0,9	0,63	0,8	0,79

Table 1. Covering of *1984* word senses in Bulgarian WordNet.

It is seen that we had already covered most of nouns and adjectives – the numbers for verbs in brackets show implemented literals that are not validated yet and that is why

not included in this report. Most of other words that are not included in Bulgarian WordNet are constructed by the author and do not really function in the language.

Quality control

The main positive characteristics of the BulNet are its completeness and consistency. Under completeness we understand the presence of all members from the Base Concepts chosen up to now within the framework of the BalkaNet project. These are Base Concepts subset 1 (1 218 synsets), Base Concepts subset 2 (3 471 synsets) and Base Concepts subset 3 (3 827 synsets). We measure the completeness of the BulNet not only by the number of the common synsets in all languages but also according to several additional criteria: lack of any "dangling relations" in the data base - that is both members of the defined relation have to be present in the WordNet; lack of any "gaps" - if a certain synset is included in the Bulgarian WordNet, then all of its hypernyms should be present up to the top of the tree; lack of any "free" nodes - a synset included in a WordNet should be in a relation at least with one different synset. Each synset must contain at least one literal, as well as at least one language-internal relation must be defined for each synset.

Finally, we consider the WordNet complete if the following tags have received a value: the synset ID tag which makes the relation to the corresponding synset in English WordNet2.0 explicit, the synset POS tag ensuring that each synset is specified for the part of speech it belongs to, the synset DEF - an appropriate interpretation definition must be entered for each synset, the SENSE tag - each literal has to receive unique sense number that distinguishes it from the homographic literals with different meaning, the synset BCS tag - each synset has to be defined as to whether or not it belongs to a particular Base Concept subset. On the other hand, there are some XML tags such as USAGE, SNOTE, LNOTE, STAMP which are not obligatory, so they may not possess a value and are removed automatically if empty. The completeness of the current state of the BulNet can be exemplified with the following Table 2:

NUMBER OF SYNSETS	18 716
NUMBER OF LITERALS	35 307
BASE CONCEPTS SUBSET 1	1 218

BASE CONCEPTS SUBSET 2	3 471
BASE CONCEPTS SUBSET 3	3 827
EMPTY TAGS	0
SYNSETS WITHOUT ID TAG VALUE	0
SYNSETS WITHOUT POS TAG VALUE	0
SYNSETS WITHOUT BCS TAG VALUE	0
SYNSETS WITHOUT DEFINITION	0
SYNSETS WITHOUT LITERALS	0
SYNSETS WITHOUT ILR	0
"FREE" SYNSETS	0
"DANGLING" RELATIONS	0
"GAPS"	0
LITERALS WITHOUT SENSE TAG VALUE	0

Table 2. The completeness of Bulgarian WordNet

The second important characteristic of the BulNet is its consistency. As a result of the application of the specified methodology for checking and correction of the Bulgarian WordNet, the current status of the XML syntax is the following (Table 3):

DUPLICATED LITERALS IN A SYNSET	0
DUPLICATED SENSE NUMBERS	0
INCONSEQUENT SENCE NUMBERS	0
MISSING SENSE NUMBERS	0
DEFECTED ID TAGS VALUES	0
DEFECTED POS TAGS VALUES	0
DEFECTED BCS TAGS VALUES	0
SPELLING ERRORS	0
WORDS IN LATIN CHARACTERS (correct)	961
EMPTY ID'S	0
DUPLICATED SYNSETS	0
DUPLICATED RELATIONS	0
LITERALS IN CONFLICT	0

Table 3. The consistency of the Bulgarian XML file

When validating semantic relations already defined for a given synset the following tests were used:

-- All Bulgarian synsets whose hypernym differs from the English ones and synsets without a hypernym were checked again by a lexicographer. This check was broadened to cover all relations. Thus every difference in relations between EWN2.0 and the Bulgarian WordNet is either language specific and linguistically substantiated or is due to the fact that one of the synsets is not yet presented in the Bulgarian WordNet.

-- There must be no hypernym cycles, as well as any relation loops inside WordNet. The cycle is defined easily in such (artificial) examples like following:

"Rose" has hypernym "flower".

"Flower has hypernym rose". (one step cycle)

It is not clear how to define the errors in cases of multiple hypernymy (or any transitive relation - e.g., *"eye"* may be a part both of *"face"* as well as a part of *"visual system", "face"* may be a part both of *"human"* as well as part of *"head", "head"* may be part of *"body"* and *"animal".*

In some cases there are wrongly connected nodes, but some cases may be instances of different “subrelations”. For example, the distinction between the following types of hyponymy is not included for the time being in the Bulgarian WordNet: *"kingdom"* is a kind of *"state",* while *"Bulgaria"* is an instance of *"state"; "actor"* is a role of *"person",* while *"man"* is a type of *"person".* If we allow such “subrelations”, we could avoid multiple transitive relations for a synset and thus we could successfully apply the consistency validation.

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

-- It was checked whether there were literals in the Bulgarian WordNet that coincide with their glosses. In such cases the glosses were redefined.

-- Another check was whether the glosses of different synsets were identical and if they were -- the interpretation definitions were compared and differentiated in an appropriate manner.

When building the Bulgarian WordNet, we have come across the problem of English synsets that denote concepts existing in the Bulgarian language consciousness but are not lexicalized in Bulgarian. In such cases we have adopted the strategy of

keeping the node in the Bulgarian WordNet and marking it with the phrase "no lexicalization". At the moment we have 99 language specific concepts defining relative relations such as {\it *"baldaza"* (the sister of one's wife) and some adjectives.

The next table illustrates the level of the consistency in the Bulgarian WordNet (differences in the relations does not involve inconsistency).

DIFFERENCE IN ID's	0
EQUIVALENT GLOSSES	0
GLOSSES EQUAL WITH LITERALS	0
DIFFERENCE IN RELATIONS hypernym	0
DIFFERENCE IN RELATIONS be in state	16
DIFFERENCE IN RELATIONS also see	369
DIFFERENCE IN RELATIONS similar to	490
DIFFERENCE IN RELATIONS holo part	68
DIFFERENCE IN RELATIONS holo member	10
DIFFERENCE IN RELATIONS subevent	0
DIFFERENCE IN RELATIONS causes	0
DIFFERENCE IN RELATIONS derived	0
DIFFERENCE IN RELATIONS particle	0
DIFFERENCE IN RELATIONS verb group	27
DIFFERENCE IN RELATIONS near antonym	9
DIFFERENCE IN RELATIONS holo portion	9
ANY LOOPS	0

Table 4. The consistency of the encoded relations and definitions

Current status of the Bulgarian WordNet

Bulgarian WordNet contains 18 716 synonyms (synsets), distributed into four parts of speech. Every synset has one definition which encodes the meaning common for all the literals in the synset -- thus the number of the definitions has to be equal to the number of the synsets. The number of the literals included in the Bulgarian WordNet is 35 307 and the average number of literals per synset is 1.89. Some of the words included in the WordNet have more than one sense and the number of the graphic words is 27 088 -- this represents almost half of the standard Bulgarian orthographic dictionary. The average value of polysemy included in BulNet is 1.3 senses per graphic word. The

language-internal relations (semantic, morpho-semantic and extralinguistic) included in the Bulgarian WordNet are seventeen (following the Princeton WordNet), their occurrences are 32 213, the average number of relations per synset is 1.72. The figures representing the current state of the Bulgarian WordNet are exemplified in the Table 5.

	Nouns	Verbs	Adjectives	Adverbs	Total
Synsets	12 274	3 559	2 881	2	18 716
Literals	21 986	8 532	4 786	3	35 307
Literals/synsets	1,79	2,39	1,66	1,5	1,83
Graphic words	17 992	5 467	3 626	3	27 088
Literals/synsets	1,79	2,39	1,66	1,5	1,84
Graphic words	17 992	5 467	3 626	3	27 088
Literals/words	1,22	1,56	1,32	1	1.3
ILR	19 470	8 304	4 436	3	32 213
ILR per synset	1,58	2,33	1,54	1,5	1,73
Definitions	12 274	3 558	2 881	2	18 715

Table 5. Statistical data characterizing BulNet

Each synset included in the WordNets is part of a semantic tree which consists of chains of hyponymy and hypernymy relations. The tree structures of Bulgarian and English noun WordNets end with the same number of tops. It is obvious that the hierarchies for nouns are quite deep and the density of Bulgarian noun trees is much greater than the average for Bulgarian verbs. The difference in the number of verb tops is due to the different number of synsets encoded in the Bulgarian and the English WordNet (Table 6). The hierarchies for both nouns and verbs are quite deep. The average density for Bulgarian noun tree is 1 365.78 (in English wordNet2.0 it is 8 854.33), and the average density for Bulgarian verb trees is 9.16 (compared to 24.38 for English wordNet2.0).

WN	N nodes	Tops N	V nodes	Tops V
Eng2.0	79 689	9	13 508	554
BulNet	12 274	9	3 559	389

Table 6. Number of tops per Bulgarian nouns and verbs

The major part of the relations in BulNet are semantic relations: ALSO SEE, CAUSE, HOLO MEMBER, HOLO PART, HOLO PORTION, HYPERNYM, NEAR ANTONYM, SIMILAR TO, SUBEVENT, VERB GROUP. There are also some morpho-semantic relations: BE IN STATE, BG DERIVATIVE, some morphological (derivational) relations: DERIVED, PARTICLE, and some extralinguistic ones: REGION DOMAIN, USAGE DOMAIN, CATEGORY DOMAIN.

The hypernym-hyponym relation rates highest in terms of number of occurrences - 15 838 in 18 716 synsets - approximately 84 percent of the total number of synsets are assigned hypernyms. The distribution of the BulNet relations in comparison with the English WordNet2.0 is shown in Table 7.

ILR	POS/POS	EW2.0	BulNet
ALSO SEE	A/A V/V	3 240	895
BE IN STATE	A/N	1 296	591
BG DERIVATIVE	N/V	36 630	6 469
CATEGORY DOMAIN	N/N V/N A/N B/N	6 166	638
CATEGORY MEMBER	N/N V/N A/N B/N	6 166	638
CAUSES	V/V	439	104
DERIVED	A/N	6 809	1 071
HOLO MEMBER	N/N	12 205	841
HOLO PART	N/N	8 636	1 241
HOLO PORTION	N/N	787	107
HYPERONYM	N/N V/V	94 844	15 838
HYPONYM	N/N V/V	94 844	15 838
IS CAUSED BY	V/V	439	104
IS DERIVED FROM	N/A	6 809	1 071
IS STATE OF	N/A	1 296	591
IS SUBEVENT OF	V/V	409	162
MERO MEMBER	N/N	12 205	841
MERO PART	N/N	8 636	1 241
MERO PORTION	N/N	787	107
NEAR ANTONYM	N/N A/A V/V	7 642	1 847
PARTICLE	A/V	401	56
REGION DOMAIN	N/N V/N A/N B/N	1 280	4

REGION MEMBER	N/N V/N A/N B/N	1 280	4
SIMILAR TO	A/A V/V	22 196	1 479
SUBEVENT	V/V	409	162
VERB GROUP	V/V	1 748	848
USAGE DOMAIN	N/N V/N A/N B/N	983	22
USAGE MEMBER	N/N V/N A/N B/N	983	22
ID		115 424	18 715
L NOTE		0	1 520
LITERAL		203 147	35 306
POS		115 424	18 715
SENSE		203 147	35 306
S NOTE		0	125
USAGE		48 231	8 816

Table 7. Distribution of the BulNet relations

Cross-lingual validation

Our team developed a Web-based system (WordNet Validator) for validation (and correction) of the WordNets completeness and consistency (http://dcmb.ibl.bas.bg). In the WordNet Validator the predefined queries are used. The system works with the adopted xml-file format. The WordNet Validator has the following main functions:

a) Automatic correction of xml syntax;

b) Validation of WordNet completeness and consistency;

c) Search for a given synset;

d) 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 latest versions of English, Czech, Bulgarian, Greek, Turkish and Serbian WordNets or can be browsed (Figure 1). The browsed language is accepted if it corresponds to several conditions: an appropriate xml format, no empty ID tags and no duplicated ID's.

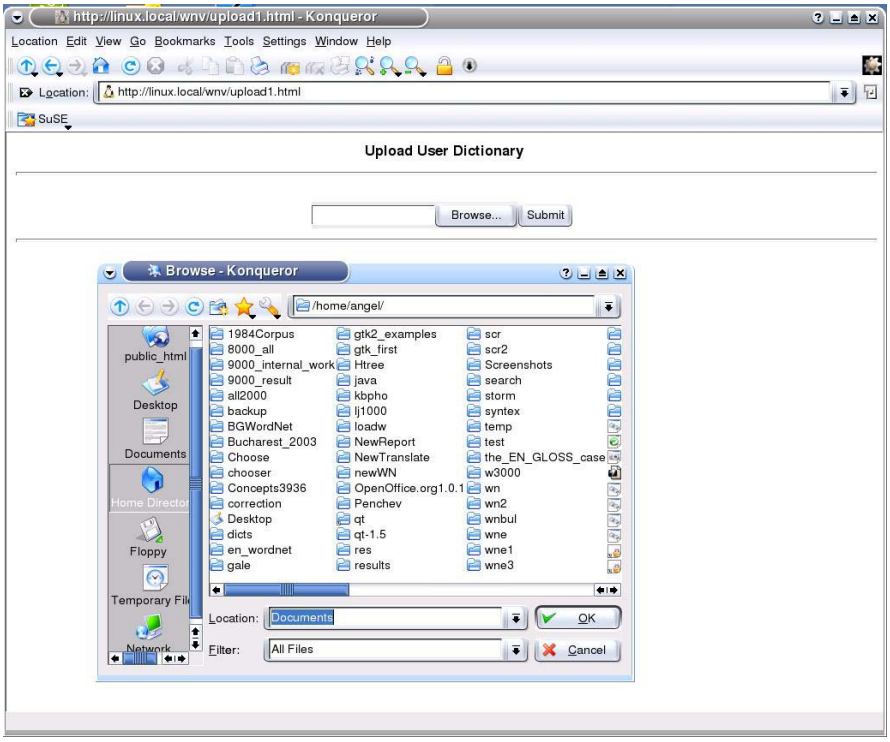


Figure 1. WordNet Validator Browse Function

In the following cases the automatic correction function of the WordNet Validator operates: facultative empty tags are removed; duplicated literals in a synset are removed while keeping only one of them; the SENSE tags are assigned values so that there are no empty tags, all tags contain only numbers, and are reordered to ensure that all sense numbers are contiguous and are not duplicated. Statistics of the automatic correction appears in the next window and a result file is constructed in which the above listed errors are fixed - the user can download it following the link on the file name.

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 Base Concepts (subsets one, two, and three), check "dangling" relations; and check "gaps".

- Verifying the consistency of the data: check ID format; check synsets without DEF tags; check synsets without literals; check duplicated relations; check differences in relations, verifying for lack of any loops inside the WordNet.

The search function allows ID searching - the result is all the available information pertaining to 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 - the wanted relation (for example, hypernyms up to the top or holo parts down to the leaves) can be selected in the check box (Figure 2).

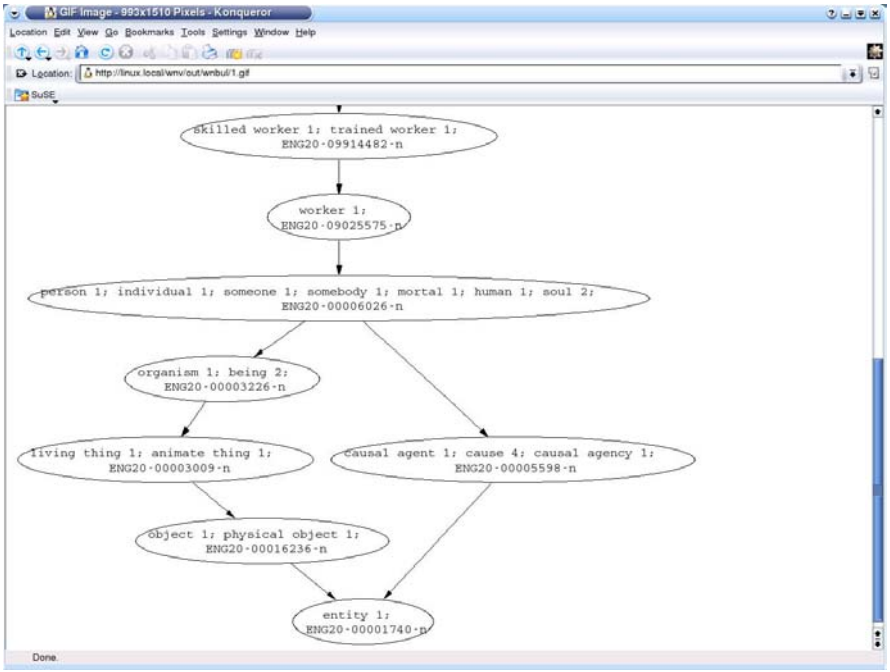


Figure 2. Visualization function of the WordNet Validator

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

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

Wordnet Improvement

During the reporting period a fair amount of synsets has been developed and incorporated within the Greek Wordnet. These mainly concern BCs3 which had not been finalized previously. To facilitate the development of the remaining BCs3, we automatically imported the Princeton WordNet structure and we manually corrected mistakes and/or mismatches. Currently the entire set of BCs3 (i.e. 1862 BCs3) is represented within the Greek Wordnet.

Besides developing new synsets, a significant amount of effort has been devoted in checking the quality of the already existing synsets and correcting mistakes. These tasks which are still in progress and are expected to be finalized soon concern mainly the (i) manual mapping of 1.7 and 1.5.1 to Princenton ILI's (completed by the time of this report), and (ii) the mapping of GRE-synsets to their Princenton equivalents (to be finalized by May 15, 2004).

Moreover, during the reporting period the lexical relations holding between Greek Wordnet synsets have been significantly enriched especially the ones holding between verbs and adjectives.

Validation Tasks

The current version of the Greek Wordnet is in valid XML format, reassuring that there are no emply tags and that the non-lexicalized synsets are denoted by the <NL> tag. In particular, the following checking has been performed concerning the evaluation of Greek Wordnet's quality.

- 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.

Spell checking

An improrttant task regarding the reassurance of the qualitative content delivered by Greek Wordnet concerned the spell checking of the synsets currently encoded. Specifically all BC1, BC2 and BC3 have been semi-automatically checked and mistakes encountered have been manually corrected. The most frequently occuring mistakes that have been traced as well as the remedial actions takes are listed below:

- Correction of misspellings
- Due to information retrieval reasons in the synset name of the adjectives only the male gender is kept
- Abbreviations met in the glosses are being replaced with their fully written
  - type
  - i.e. sb --→somebody
    - sth---→ something
- Enforcing a uniform format among synsets' glosses by:
  - allowing only commas among the words and not slashes e.t.c.
  - erasing any full stops at the end of the glosses

Currently, all remaining Greek Wordnet syssets (i.e. those that are not encoced as BCs) are being spell checked and their correction is expected to be finalized by the end of May 2004.

1984 Corpus Processing

The Multilingual *1984* Corpus

In the framework of BalkaNet validation the Greek text of George Orwell´s *Nineteen Eighty-Four* has been annotated, lemmatized, aligned and incorporated in a multilingual parallel corpus. This parallel corpus of the *Nineteen Eighty-Four* text has already been developed for all the participating languages in BalkaNet, except Greek and Turkish, during the Multext-East project (Erjavec et al., 1998).

For the annotation of the text, we used the standardized specification for the description of the morpho-lexical information of words that was proposed (Tufis et al., 1998) in the framework of the Multext-East project. The morpho-lexical information is provided as a string, using a linear, term-like encoding. In this notation, the position in a string of characters corresponds to an attribute, and specific characters in each position indicate the value for the corresponding attribute. That is, the positions in a string of characters are numbered 0, 1, 2, etc., and are used in the following way:

The character at position 0 encodes part-of-speech;

Each character at position 1, 2, n, encodes the value of one attribute (person, gender, number, etc.), using a one-character code.

If an attribute does not apply, the corresponding position in the string contains the special marker ‘-'.

For example, the string *"Ncns"* stands for:

Part-of-speech: **N**oun, Type: **c**ommon, Gender: **n**euter, Number: **s**ingular

Each sentence in the multilingual corpus is assigned a sentence number, which uniquely identifies it. Sentences with the same number are common for all languages. An example of such a sentence appears in Figure 1. The sentence with number 3751 appears in English, Romanian and Czech. The annotation on the text is done with XML and for each word its dictionary citation form ("lemma" attribute) and its morpho-lexical information ("ana" attribute) is given. As it can be seen in the figure the English word *"crash"* is

assigned the grammatical information *"Ncns"* which, as mentioned, means that it is a common neuter, singular noun.

```
<tu id="Ozz.3751">
<seg lang="en"><s id="Oen.2.10.33.8"> <w lemma="there" ana="Pt3">There</w>
<w lemma="be" ana="Vmis3s">was</w> <w lemma="another" ana="Dg--s">another</w>
<w lemma="crash" ana="Ncns">crash</w> <c>.</c></s></seg>

<seg lang="ro"><s id="Oro.2.10.70.6"> <w lemma="sine" ana="Px3--a--------w">Se</w>
<w lemma="auzi" ana="Vmis3s">auzi</w> <w lemma="un" ana="Tifsr">o</w>
<w lemma="nou" ana="Afpfsrn">nou&abreve;</w>
<w lemma="bufnitur&abreve;" ana="Ncfsrn">bufnitur&abreve;</w> <c>.</c></s></seg>

<seg lang="cs"><s id="Ocs.2.10.33.8">
<w lemma="zazn&iacute;t"  ana="Vmps-sfan----n">Zazn&ecaron;la</w>
<w lemma="dal&scaron;&iacute;"  ana="Afpfsn---c">dal&scaron;&iacute;</w>
<w lemma="r&aacute;na" ana="Ncfsn">r&aacute;na</w><c>.</c></s></seg>
</tu>
```

Figure 1: An annotated, aligned and lemmatized sentence for English, Romanian and Czech taken from the Multext-East project.

Building the Greek *1984* Corpus

Making the Greek text of *Nineteen Eighty-Four* appropriate for incorporation in the multilingual corpus and therefore for BalkaNet's validation, initially involved the scanning of the hardcopy version of the book and the use of an Optical Character Recognition (OCR) program in order to obtain the text in machine readable form. Afterwards it was necessary to align the text to the rest of the texts in the multilingual corpus. The final step is to annotate with morpho-lexical information and find the citation form (lemma) of each word in the corpus.

Sentence Alignment

The purpose of the sentence alignment process is to take each sentence in the Greek text and find which is the corresponding sentence in the English text. By aligning to the English text, we are simultaneously aligning to all the other languages, since English in Multext-East was used as a hub language. The alignment task is not trivial, since it is often the case that one of the following problems exists:

An English sentence has been translated into two Greek sentences

Two or more English sentences have been translated into one Greek sentence.

An English sentence has been left out of the Greek translation.

A sentence of the original text is not present in the aligned corpus of the Multext-East project. This case is very common since the multilingual corpus is the set of sentences that are common for all languages. Therefore if a sentence was not present in even one of the languages it will not appear in the final multilingual corpus. Specifically, of the 6737 sentences in the original English text only 5466 sentences were present in the aligned multilingual version we were working with, meaning that almost 18% of the original text was missing.

Certain characteristics of the 1984 text made some of these existing methods for sentence alignment, which are based mainly on machine learning, hard to use. For example we had no previously annotated parallel corpus for training and in the English text there were no paragraph or section markers or anything else except line breaks that could be used as a delimiter. Additionally, as we mentioned before a very large part of the English text was missing making manual post-processing of the text necessary to a large extend. Due to all these problems we finally opted for a more simplistic approach, which, nevertheless, would be much faster to implement.

Our approach was based on a tool we have developed and that works semi-automatically. It performs an initial alignment of the text and then it offers an interface to the human editor who will correct the alignment. The initial alignment works by scanning the text for punctuation marks such as:”.”,”;” and “!”, and considers these as sentence separators. Some heuristics are used in order to find the cases when these symbols don’t correspond to the end of sentence. For example, when the symbol “.” appears after the symbols “κ” or “κα” (“mr” or “mrs”) or after a single capital letter, the program assumes that this symbol is used to show abbreviation and it is not a sentence final full stop.

After the first step an initial alignment of the text is achieved, but it still requires human editing. The interface offered for this editing appears in Figure 2. The number of the sentence, the sentence in English and the sentence in Greek appear side by side. It is possible for the user to delete a sentence, to split a sentence into two sentences or to join two sentences together. Once any of those actions has been performed the numbering of the sentences is refreshed so as to reflect the new alignment between the two texts.

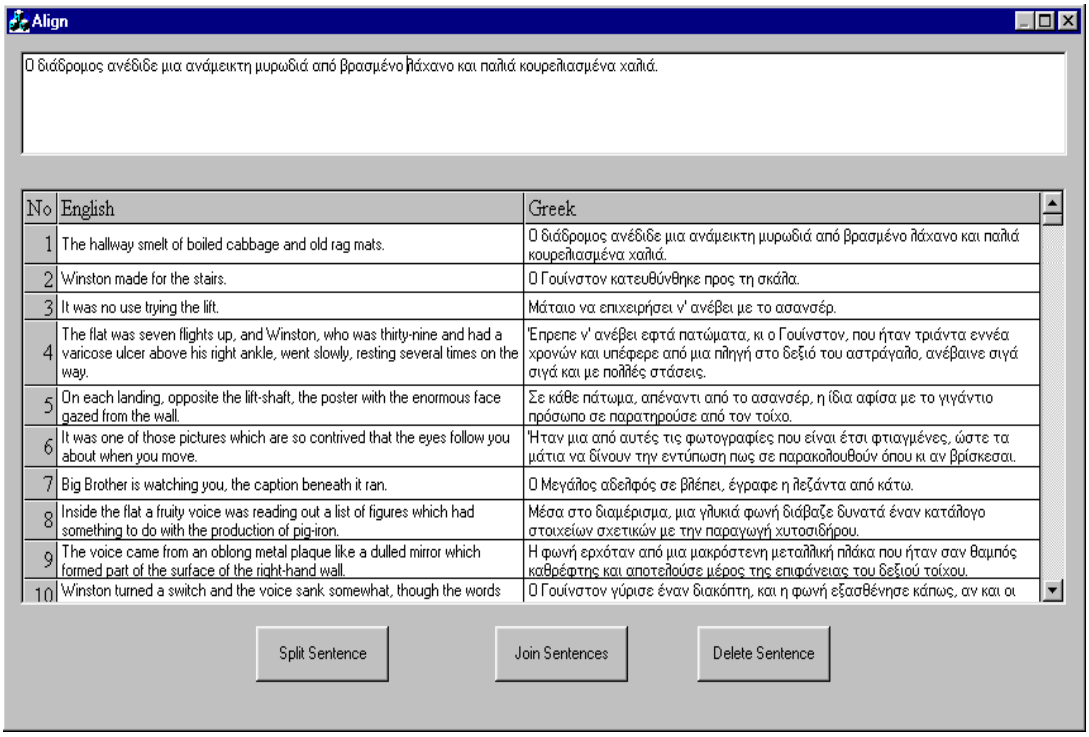


Figure 2: The sentence alignment tool

Annotation and Lemmatization

After the Greek text had been aligned to the multilingual text, it was necessary to annotate the words in the text with their grammatical attributes and to lemmatize them i.e. for each word find its citation form.

In order to achieve that, we used a lemmatizer for the Greek language (Kornilakis et al., 2004) whose function is, when given as input a word in Greek, to analyze the word and to find its dictionary citation form (lemma). The lemmatizer can deal with the inflection of nouns, adjectives and with the conjugation of verbs that do not alter their stem and can also deal with cases of irregular inflection. Furthermore it can handle stress movement, a common phenomenon in the Greek language. In order to achieve these, the lemmatizer keeps an amount of lexical information, which is kept in three lists: a list of words, a list of inflectional information and a list of irregular forms. The operation of the lemmatizer is based on the principle of removing the ending of the input word (stemming) and then subjecting the stem of the word to certain transformations (such as stress movement) specified in the list of inflectional information, in order to obtain possible lemmas of that word. Then the lemmatizer searches for these lemmas in the

wordlist to check if the lemmatization is valid. Through this process, one or more lemmas are found for the input word.

The lemmatization of the text simply consisted of running the 1984 through the lemmatizer, which assigned a lemma to each word in the text. In cases of words that the lemmatizer could not handle, such as proper names or words not appearing in the wordlist, a human annotator that worked interactively with the lemmatizer entered the correct lemma for the word manually. In cases where more than one lemma were possible for an input word, all possible choices were kept in order to be manually disambiguated at a later processing stage. Fortunately, this case was not so often so as to present an insurmountable problem. The annotation was done in a similar way, by enhancing the list of inflectional information with morpho-lexical information based on the ending of the input word and on the discovered lemma.

After the automatic process of the lemmatization and of the annotation was finished, we performed a manual validation of the automatically produced results. This was a time consuming process that included the checking of the correctness of the lemma and the morpho-lexical information for each word in the corpus, as well as the manual selection of the correct lemma for cases in which the lemmatizer has produced more than one possible lemma for a word.

Greek 1984 Corpus Statistics

In table 1 we present the characteristics of the Greek text of *Nineteen Eighty-Four* in comparison to the same data for the rest of the languages which are common in both Multext-East and BalkaNet. Data for the language except Greek were taken from (Dimitrova et al., 1998). It can be seen that the numbers are comparable for all languages.

The annotated text follows the specification given in the Multext-East project. In table 2 we give the attributes for each part of speech and the number of words that belong to that part of speech in the corpus. A sample sentence from the corpus, as it has been annotated for Greek, appears in Figure 3. In fact, it is the sentence that was given in Figure 1 for English, Romanian and Czech.

Language	Greek	English	Bulgarian	Romanian	Czech
Tokens	93299	118102	101173	118063	100358
Words	81316	103997	86020	101508	79862
Distinct Words	12972	9745	16348	15225	19115
Distinct Lemmas	6375	7260	8517	7433	9161

Table 1: Characteristics of the multilingual corpus for the various languages

```
<tu id="Ozz.3751">
<seg lang="gr"><s> <w lemma="ακούω" ana="V-is3s-p------e">Ακούστηκε</w>
 <w lemma="πάλι" ana="R-p">πάλι</w> <w lemma="ένας" ana="Ti">ένας</w>
 <w lemma="πάταγος" ana="Ncms">πάταγος</w><c>.</c></s></seg>
</tu>
```

Figure 3: Sample sentence of the Greek corpus.

POS	**Attributes**	**Appearances**
Noun	Type, Gender, Number	17047
Verb	Mood, Tense, Person, Number, Voice, Aspect	14985
Adjective	Degree, Gender, Number	6394
Pronoun	Type	7542
Article	Type	11329
Adposition	Type	6298
Conjunction	Type	5123
Numeral	Type	1041
Particle	Type	4926
Interjection	-	9
Abbreviation	-	21

Table 2: The parts of speech that can be found in the corpus, their attributes and their frequency.

References

Dimitrova, L., Erjavec, T., Ide, N. Kaalep, H., Petkevic, V., Tufis, D. (1998) Multext-East: Parallel and Comparable Corpora for Six Central and Eastern European Languages. *Proceedings of ACL/COLING98,* Montreal, 315-19.

Erjavec, T. and Ide, N. (1998) The MULTEXT-EAST Corpus, *Proceedings of the First International Conference on Language Resources and Evaluation*, Granada, pp. 971-974

Kornilakis H., Grigoriadou M., Galiotou E., Papakitsos E.: Using a Lemmatizer to Support the Development and Validation of the Greek WordNet. Second Global Wordnet Conference, Brno, Czech Republic, January 2004.

Tufis, D., Ide, N., Erjavec, T. (1998) Standardized Specifications, Development and Assessment of Large Morpho-syntactic Resources for Six Central and Eastern European Languages. *First International Language Resources and Evaluation Conference,* Granada, Spain, 233-40.

Statistical data of the Greek Wordnet (as of April 25th 2004)

In the subsequent tables all statistics performed by DBLAB concerning the coverage and POS-distribution of Greek Wordnet are summarized.

FUNDAMENTALS	NUMBER OF OCCURENCES
Synsets	18.677
Literals	24.811
Liter/synset	1.33
Liter/word	1.34
ILR	24.582
ILR per synset	1.33
Non lexicalized concepts	46
Definitions	18.649

Table 1: *Greek Wordnet's overall statistics*

SYNSET TYPE	NUMBER OF OCCURENCES
BC 1	1.218
BC 2	3.462
BC 3	3.826
Nouns	14.480
Verbs	3.539
Adjectives	635

Table 2*: Overall statistics of synsets'Pos and BC distribution in Greek Wordnet*

RELATION TYPE	NUMBER OF OCCURENCES
Also_ see	210
Be in_state	143
Verb_group	424
Derived	64
Holo_member	1324
Holo_part	2708
Holo_portion	162
Hypernym	18.521
Holo_substance	57
Causes	76
Near_antonym	693
Similar_to	46
Subevent	132
Antonym	22
Total	**24.582**

Table 3: *Overall statistics of the lexical relations encoded in Greek Wordnet*

3.4 Romanian wordnet

Towards the final stage of the Romanian wordnet

The last stage in developing the Romanian wordnet consisted in two phases: enrichment of the number of synsets and improvement of its quality.

Enrichment of the Romanian wordnet

With this task we aimed at covering the POS distribution agreed by all the consortium members.

The state of the Romanian wordnet before the final step is presented in the below table:

POS	Nouns 65%	Verbs 25%	Adjectives 5%	Adverbs 5%
Number of synsets (POS) agreed to be implemented	10400	4000	800	800
Number of synsets (POS) in Wordnet	10727	2930	844	200
Number of synsets to be implemented		1070		600

Table 1. Pre-final status of the Romanian wordnet

As you can see from the table above we had already covered the number of synsets for nouns and adjectives but we had not implemented all verb and adverb synsets. For the selection of the synsets to be implemented we took into consideration the semantic validation task, trying to cover as many senses as possible of the words in *1984*. The procedure consisted in the following steps:

1. From the English version of *1984* we extracted all the verbs and adverbs;
2. We generated all the PWN synsets which contain these words;
3. We identified the synsets that were not mapped onto Romanian synsets.

In the table bellow we present the degree of coverage of the English synsets in our wordnet:

	VERBS	ADVERBS
Number of synsets containing the words in *1984*	5466	864
Percentage of coverage in the Romanian wordnet	42%	22%

Table 2. Degree of coverage

We tried to implement all the adverb senses and a number of 1436 verbal synsets. We took care that all the hyperonyms of the verbal synsets selected are either already mapped or among those selected now.

Quality improvement

With a large team of lexicographers working in parallel and due to the very fine-grained sense inventory of the PWN, sense assignment conflicts were not surprising in our merge approach. Even if, idealistically, there had been only one lexicographer developing the Romanian wordnet, conflicts have come up, given the fact above. Detecting sense assignment conflicts is simple, but eliminating them requires significant efforts. There were four types of sense assignment conflicts, generated by the much finer granularity of PWN as compared to EXPD&SYND:

- sense distinctions in PWM with a metonymic flavor (e.g quality for the act) represent by far, the most frequent source of sense assignment conflicts in our wordnet:{dishonesty[2], knavery[1]}(GLOSS: lack of honesty; acts of lying or cheating or stealing) and {dishonesty[1]}(GLOSS: the quality of being dishonest).
- an English hyperonym and one of its hyponyms have as a Romanian equivalent the same literal with the same sense identifier: the synset {end[2], ending[3]} (GLOSS: the point in time at which something ends) and its hyponym { stopping point[1], finale[1], finis[1], finish[5], last[1], conclusion[3], close[1] }(GLOSS: the temporal end; the concluding time) are given sfârşit(1.1.3) as a Romanian equivalent.
- two English co-hyponyms were given the same equivalent in Romanian: for {mister[1], Mr[1]} (GLOSS: a form of address for a man) and {sir[1]} (GLOSS:

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

the EXPD gloss of a Romanian literal covers the meaning of two English synsets, themselves not very well differentiated: țâr(2.1) as compared to {herring[1]}(GLOSS: valuable flesh of fatty fish from shallow waters of northern Atlantic or Pacific; usually salted or pickled) and {kipper[1], kippered herring[1]}(GLOSS: salted and smoked herring).

Besides these categories of "objective" sources of sense assignment conflicts, we discovered several errors due to lexicographers' wrong decisions in equivalence mappings. For instance the Romanian synset {petală [1]} has been wrongly mapped on both {floral leaf [1]} and {petal [1]} where only the second equivalence is valid.

For the correction of the conflicts, two alternatives are possible:

- 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 also be easily traced and corrected.

During the developing of the project we adopted a two-way strategy for syntactic validation:

- 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.

During this phase of the project, only the latter strategy was used.

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>fiinţă<SENSE>1</SENSE></LITERAL>
            <LITERAL>vieţuitoare<SENSE>1</SENSE></LITERAL>
            <LITERAL>vietate<SENSE>1</SENSE></LITERAL>
      </SYNONYM>
      <DEF>Tot ceea ce are viaţă</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 viaţă</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 8.500 synsets in BCS 1, 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:

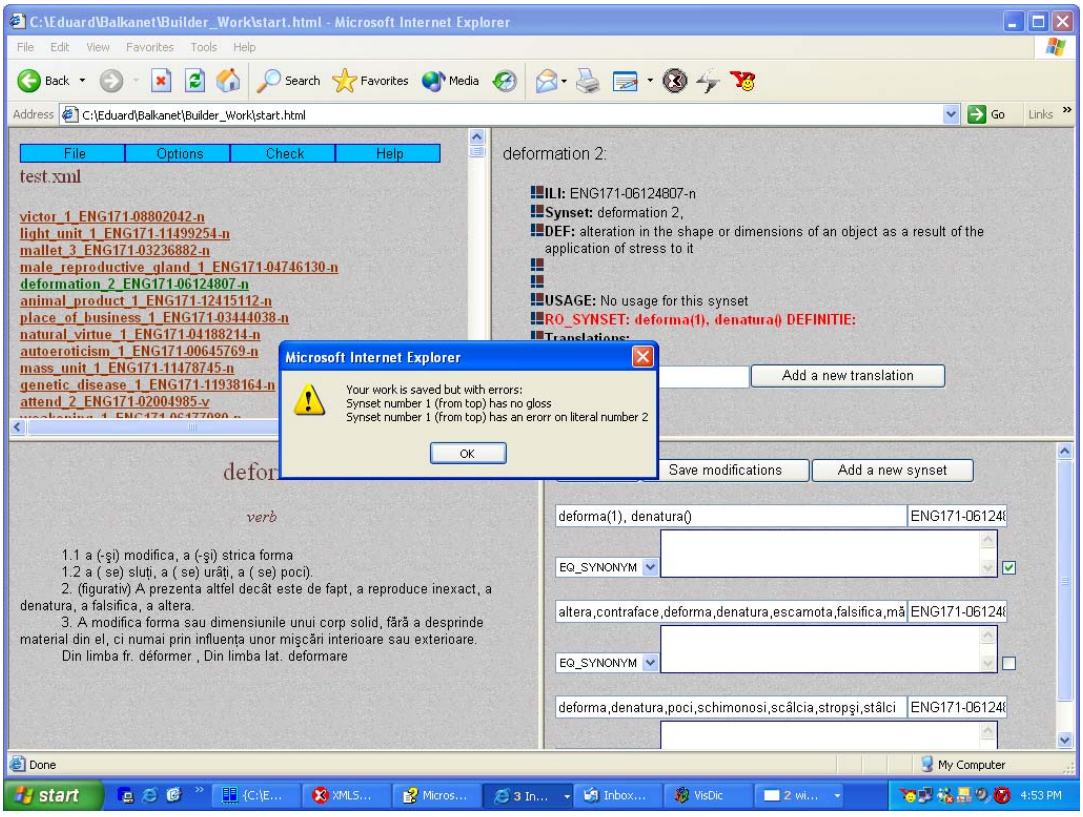


Figure 1. WNBuilder Interface. Error message after saving a synset.

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:

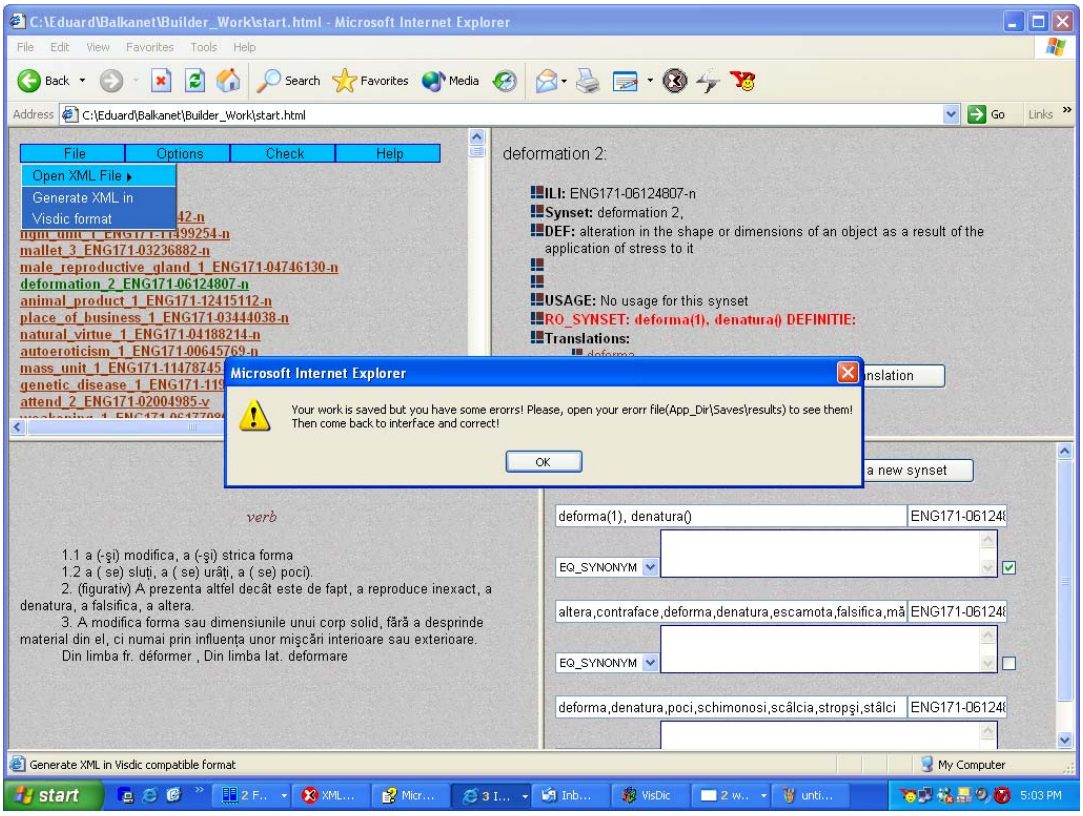


Figure 2. WNBuilder Interface. Error message when exporting work in XML format.

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, but only the second one, namely WN-Correct-2, was used during this phase, following the steps below:

- 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 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.

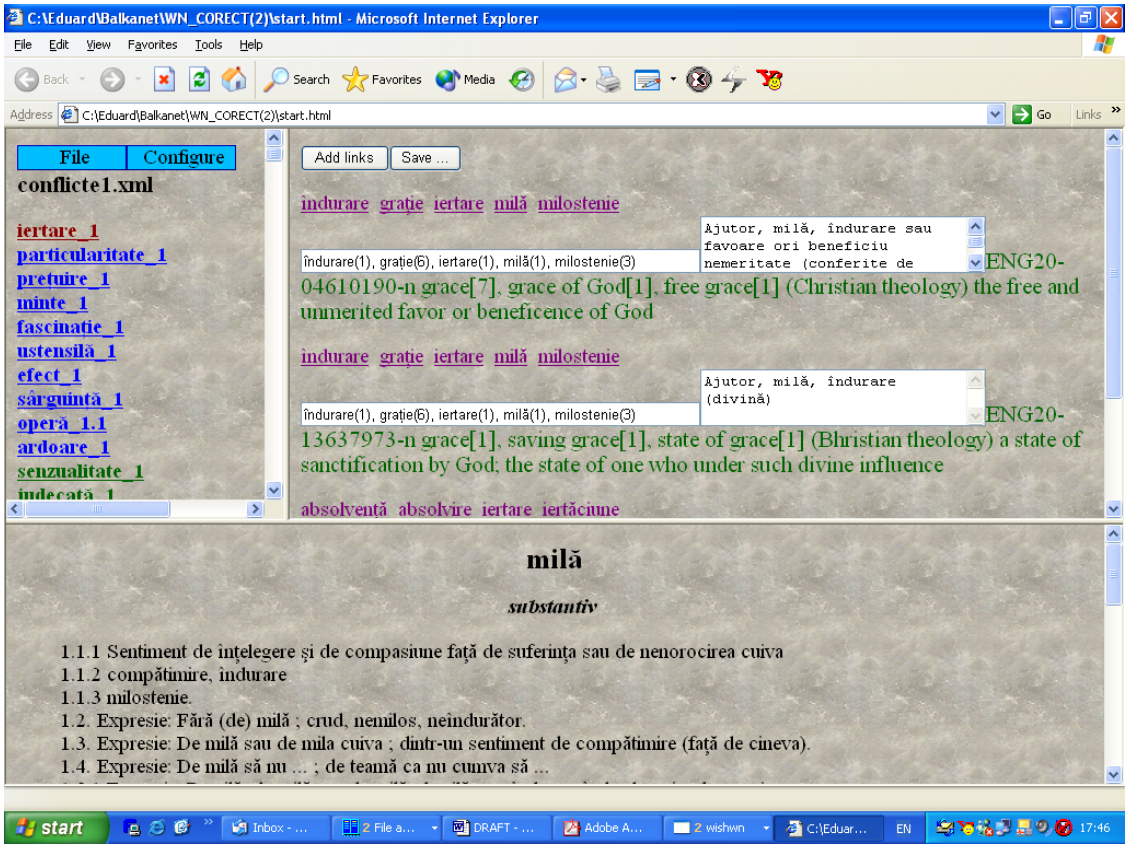


Figure 3. Wn-Correct-2 interface

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)*.

Current status of the Romanian wordnet

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

Noun synsets	Verb synsets	Adj. synsets	Adv. Synsets	Total
10725	4173	844	833	16575

Table 3: POS Distribution of the Synsets

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

Hypernym	14867	category_domain	579
near_antonym	1576	also_see	394
holo_part	1005	subevent	169
similar_to	896	holo_portion	107
verb_group	980	causes	122
holo_member	779	be_in_state	546

Table 4: Internal relations used in the Romanian wordnet.

The table below shows the average synset length and the average senses per literal for Romanian wordnet.

Language	Synsets	Token literals	Type literals	Average synset length	Average senses/lit
Romanian	16575	29299	17527	1,76	1,67

Table 5. Average synset length and average senses per literal.

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.

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 6594 synsets, covering almost completely sets BC1 and BC2. Also, one third of the BC3 set has already been covered, paying special attention to those that fill gaps in BC1 and BC2 as well an the ones 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. 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	1218	1219	99.9%
BC2	3120	3508	88.9%
BC3	1149	3788	30.3%
other	1107		
total	6594		

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	4859	73.6%	7884	1.62	71	"mesto", 11 senses
verbs	1495	22.6%	2975	1.99	98	"drzxati", 13 senses
adjectives	232	3.5%	300	1.38	3	"velik", 8 senses
adverbs	10	0.3%	10	1.00	0	
total	6596	100.0%	11169	1.73	172	

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

Hypernym	6112
near_antonym	426
holo_part	302
verb_group	133
holo_member	718
be_in_state	109
Subevent	56
Causes	45
derived	98
other	
Total	10518

Out of 6594 synsets, the majority of 6413 (97.2%) now have glosses, and for 182 synsets only glosses remain to be added.

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 70.000) , dictionary of word forms (DELAF - around 1.000.000) 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 8022 (71.8%), while this information needs to be added to another 3147 literals (28.2%).

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.

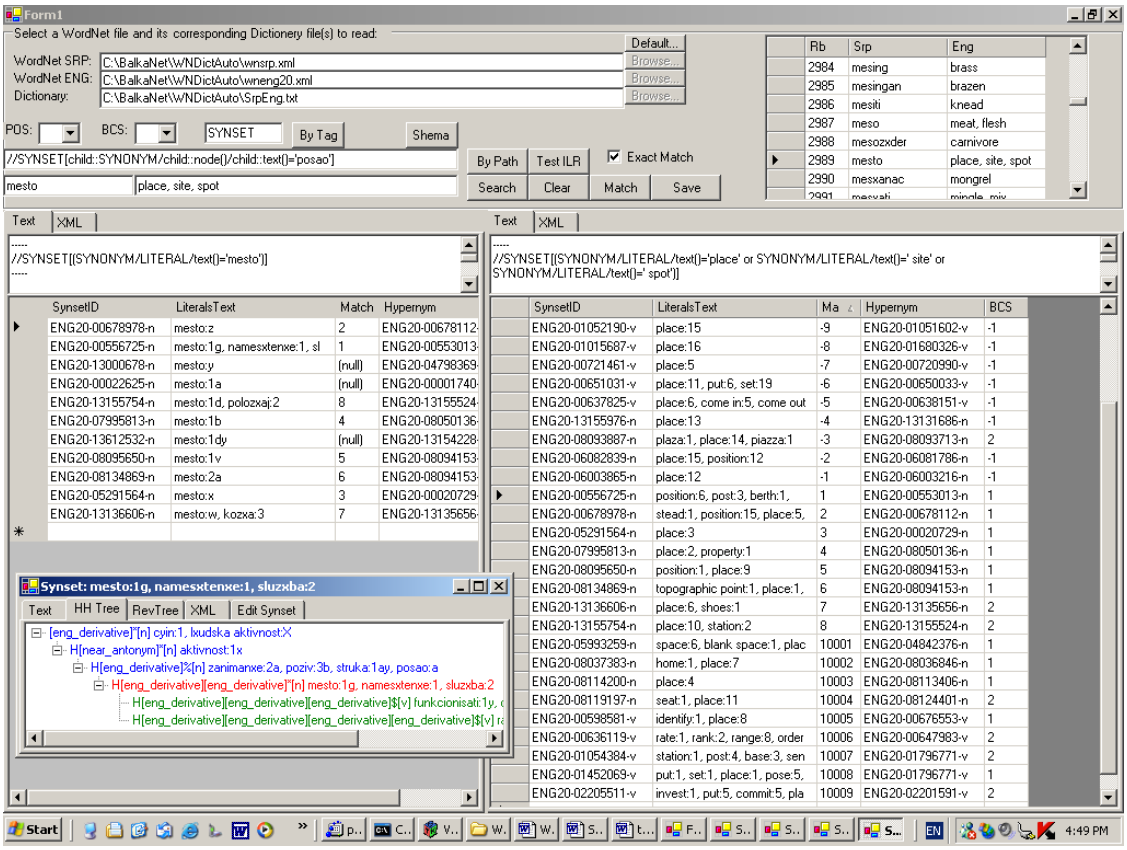
The results obtained by validating a representative group of synsets fully approve the usability of corpora approach to the validation of wordnet synsets. Besides the reestablishment of synsets themselves, this approach enables the establishment of relations between various derivatives, either by including 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, 319 synsets have been checked against corpora, and as a result 386 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).



4. For the purposes of semantic cross-lingual validation a tagged, lemmatized and disambiguated Serbian version of 1984 has been completed.

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 presented them in a paper submitted to 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.

3.6 The Turkish wordnet

Validation Tasks

Syntactic Quality

We ensured the syntactic quality of the latest version (9 April 2004) 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 corresponds to an unlexicalized concept, it has at least one <LITERAL> tag, together with its subtag <SENSE>. Otherwise, it has the special <NL>yes</NL> tag. There are no empty tags.

Structural Quality

Gaps and BalkaNet Common Sets: In line with the common decision taken by all partners, each wordnet (except Serbian) should have the synsets in BCS1, BCS2, and BCS3. We, as the Turkish team had already finished BCS1 and BCS2, and at the end of 2003, we also finished BCS3. We obeyed the rule that the wordnets should not contain any "gaps". We found 125 gaps to be added to the latest version.

Closed-world Assumption and Dangling Relations: Due to the closed-world assumption we adopted, if a relation is defined between ILI 1 and ILI 2, where ILI 1 is contained in the wordnet, then ILI 2 should also be contained in the wordnet. All such relations have been identified with the help of a small Perl script and the ones that have missing synsets have been deleted from the file.

VisDic Tests: We applied VisDic's duplicate ID test and duplicate synset literal test on our wordnet. We identified 24 duplicate ID's and two duplicate synset literals. We corrected these mistakes in the current version. Another test VisDic offers allows us to identify duplicate semantic links. In our recent wordnet we had 56 duplicate links, all of which were not errors but instances of relations "verb_group", "similar_to", and "also_see".

"1984" Corpus

At the time of the preparation of Deliverable 6.1, we had finished scanning and optically recognizing the Turkish translation of George Orwell's novel 1984 for wordnet

validation purposes. We had also aligned all sentences at the sentence level, using the alignment tool of the TRADOS translation memory software. There were 6,259 Turkish-English pairs aligned. We then compared this parallel corpus with the seven-language parallel corpus (final product of the MULTEXT project) which was uploaded on the IS by our Romanian partners. The main parallel corpus had 5,463 aligned sentences, since only the intersection of the one-to-one alignment of the six languages with English was taken into account. With the help of a Perl script, exactly matching lines of the main parallel corpus and our EN-TR aligned corpus were calculated; 4,073 (74.5 %) lines were assigned sentence IDs from the MULTEXT corpus. Manual checking of non-matching lines showed that some of the sentences in the MULTITEX corpus were divided into two separate sentences in the aligned EN-TR corpus, hence marked as no-ID, although both of them could have the same ID taken from the main corpus. A new Perl script was used to identify such cases and mark them not as exact IDs but as candidates, since the decision process requires manual revision to eliminate possible errors due to the loose matching algorithm used in the second script. The second pass gave us a set of 214 (4%) candidates. 796 (14.6%) of the sentences do not exist in the original MULTITEXT corpus, therefore the lines that possibly had an ID but did not have an ID at the end of the process was only 380 (6.9 %). The next step was to morphologically analyze every word in the corpus.

We finished the morphological analysis of the Turkish 1984 corpus. We first passed the Turkish file through our morphological analyzer, which gave us 402 unique unknown words. Most of these unknown words were proper names. They were manually checked and assigned lemmas and POS tags. Currently, the total number of unique unknown words is 163. The next step was to pass the morphologically analyzed file through our POS-tagger. We do not have a Turkish POS-tagger that assigns one POS to each word. Instead, we use a Perl script that extracts the POS of a given word from the morphological analysis result. Wherever ambiguity exists, filters designed using statistical observations were used to eliminate improbable POS assignments. Special filters prepared for the 1984 corpus were added to lower the ratio of ambiguous analyses. At the end of the process, the ratio of POS tags per word fell to 1.28. Further enhancements required manual checking. The final format of the file consisted of

sentence tags with IDs taken from the MULTEXT corpus; sub-tags for each word in the sentences, with attributes “lemma” and “ana”lysis. The analyses were only assigned a POS tag although full analyses were available, since the format of our morphological output and the format of the MULTEXT project do not match.

After finishing the “1984” corpus we again applied statistical methods comparing “1984” and TWN. We built the frequency list of the words in the “1984” corpus. First we omitted function words and passed the list through our morphological analyzer and a POS tagging script. We also morphologically analyzed the then current Turkish WordNet and obtained synset member roots with POS tags attached to each one. The final results gave us a 87.4% coverage of TWN roots on the “1984” corpus.

Similarly, we tested our wordnet by using a Turkish frequency wordlist derived from a 13-million-word corpus. We used the same method and obtained the following results: When we deleted function words and took the 50,000 most frequent words, coverage was 85.94%. When we took the first 20,000 words, the rate rose to 86.45%. We then limited our list to the 1,000 most frequent words of the corpus, and coverage reached 87.32%.

Added Synsets

In line with the decision taken at the 5th Progress Meeting in Bucharest, we added some new synsets to our wordnet. We tried to select those synsets that are “important” in Turkish. The first step in this process was to collect concepts from the following most fruitful domains: administrative system (provinces, municipalities, officers), religious objects, religious practices, wedding traditions, architecture (buildings, parts of buildings, styles), food, animals, plants, fish, traditional clothes, traditional occupations, traditional arts, handicrafts, traditional music (genres, dances, instruments) and tools (special types of scissors, knives, cooking utensils, farming equipment etc.). We collected Turkish specific concepts without considering whether they were represented in PWN or not. When we reached the stage of integrating the new synsets into our wordnet, we first manually checked all the candidates to see if they already have ILI numbers or not, and separated synsets into two groups according to this criterion:

- **Turkish-specific synsets which exist in Princeton WordNet:** Some of our candidates such as "*lokum*" (Turkish delight) and "*Ankara kedisi*" (Angora cat) were already represented in PWN. We included such synsets (about 200) in our wordnet, with their gap hypernyms where necessary.
- **Turkish-specific synsets which do not exist in Princeton WordNet:** The second group consists of Turkish synsets that are not represented in PWN. We added 300 Turkish-specific synsets with their TUR-ILIs automatically assigned by VisDic. These synsets contain glosses in both Turkish and English, so that the partners can compare them and mark common concepts as "BalkaNet synsets". In order to make the comparison process easier, we also provided pictures of Turkish-specific objects where available. The total number of pictures is 127.

Statistical Data Regarding Turkish Wordnet (as of April 9th, 2003)

FUNDEMENTALS	NUMBER OF OCCURRENCES
Synsets	12,148
Literals	16,707
Nonlexicalized Literals	991
Definitions	4,608
POS tags	12,148
Literal/synset	1.49

SYNSET TYPE	NUMBER OF OCCURRENCES
BCS1	1,218 (100%)
BCS2	3,471 (100%)
BCS3	3,782 (100%)
Nouns	9,185 (75.6%)
Verbs	2,566 (21.1%)
Adjectives	397 (3.3%)
Turkish-Specific Synsets	300 (2.4%)

RELATION TYPE	NUMBER OF OCCURRENCES
Hypernym	11,478
Holo_member	1,026
Holo_part	1,554
Holo_portion	218
Causes	100
Be_in_state	581
Near_antonym	1,437
Subevent	127
Also_see	251

Verb_group	896
Similar_to	31
Category_domain	439
Usage_domain	3
TOTAL	**18,141**

Ongoing Tasks

Missing frequent words in TWN: While we were working on the projection of the 1984 corpus and our 13-million-word corpus, we also extracted the list of our future synsets. Frequent words in the "1984" corpus and our 13-million-word corpus that are not represented in Turkish wordnet will be added as soon as possible. Some of these frequent words correspond to basic concepts such as "million", "billion" and "trillion". We think that these concepts are important for all Balkan languages, so we will distribute a list of the relevant ILI records to all partners.

Synsets from the domain of law: In line with the common decision taken at the 5th Progress Meeting in Bucharest, we will translate more than one hundred synsets that are marked as "legal synsets" in the SUMO-Wordnet ontology. It has been observed that the legal synsets in the SUMO-Wordnet ontology might not be adequate for domain classification purposes. So, additional synsets from the legal domain will be proposed to all partners, using an English law dictionary provided by a translation agency in Turkey.

Adjectives: A study conducted by a Ph. D. student using Turkish WordNet showed that the usefulness of our Wordnet could be significantly improved by adding a limited number of adjectives. These adjectives will be added to Turkish Wordnet as soon as possible and the relevant ILI records will be distributed to the partners.

Language-specific concepts: As soon as all partners finish defining their language-specific concepts and upload their lists on the Information Server, we will make an effort to merge these lists and arrive at a set of "BILI records".

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) ≥ 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;

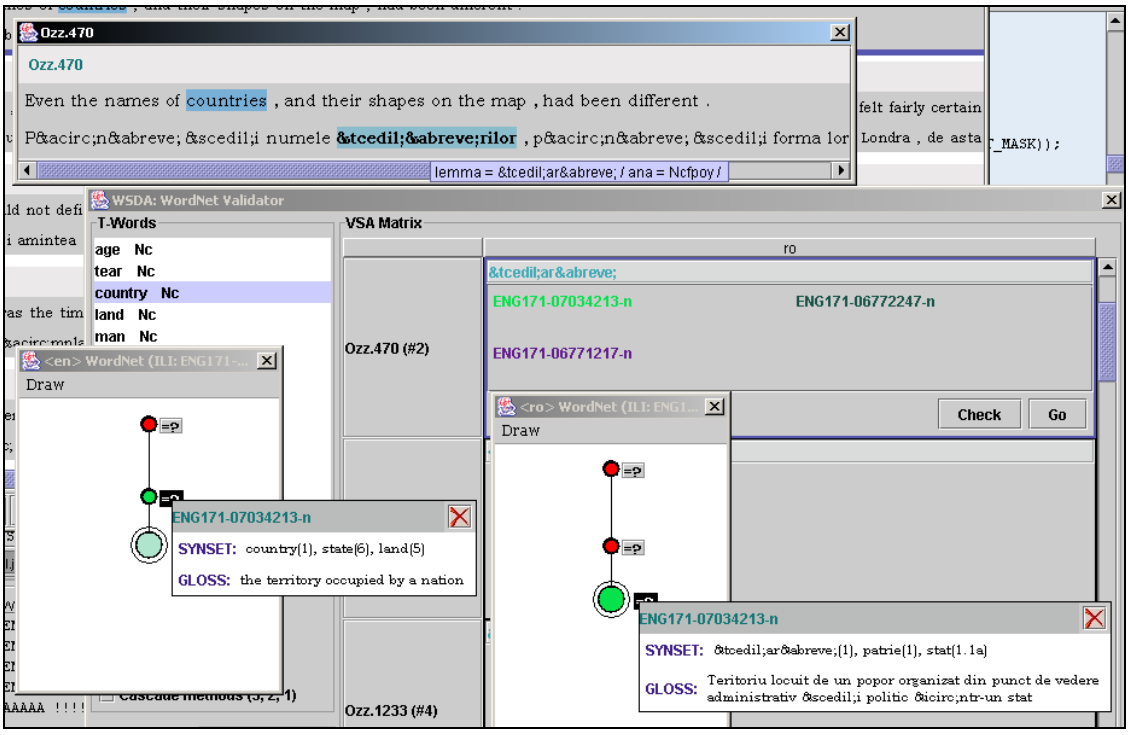


Figure 2

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;

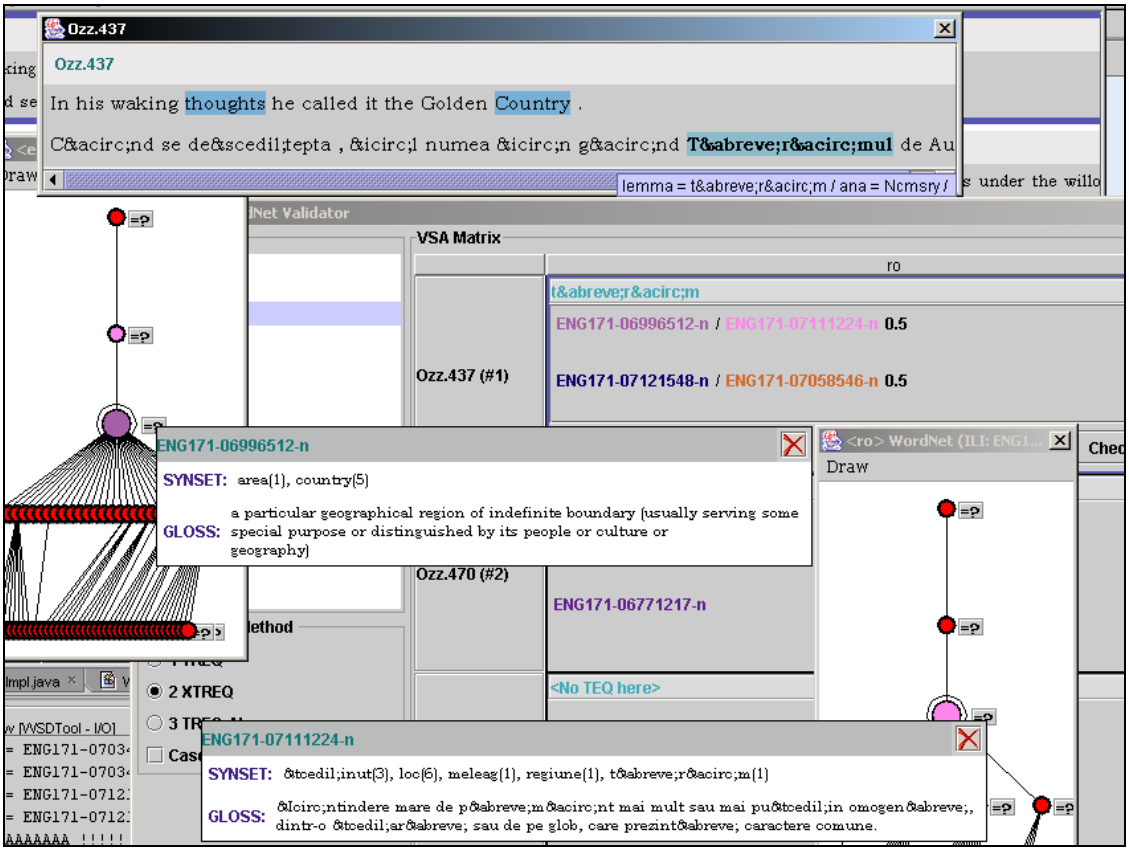


Figure 3

*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 δ_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).

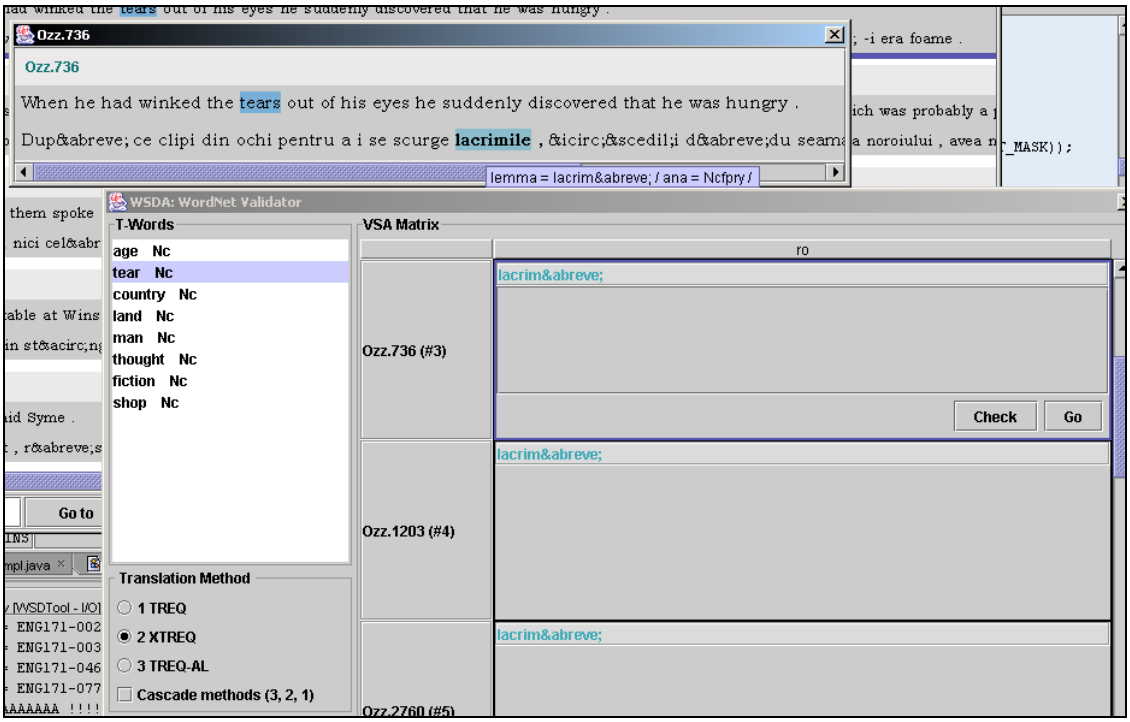


Figure 4

*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.*

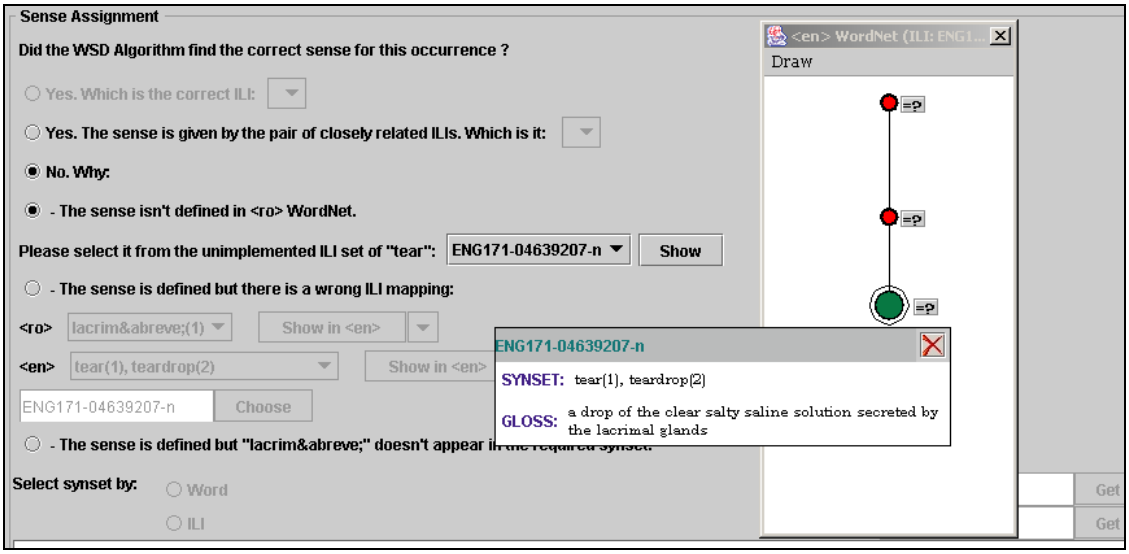


Figure 5

*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	3827	8516	

*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	21096(~72%)	4997(~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	0	no	0	0	0

* 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	24255
holo_part	1771
near_antonym	1772
similar_to	1138
category_domain	1106
verb_group	916
also_see	763
be_in_state	602
holo_portion	357
holo_member	1088
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
erb_group	888
holo_member	778

be_in_state	546
category_domain	508
also_see	333
subevent	139
holo_portion	107
causes	106
derived	28

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

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

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
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