ROMANIAN JOURNAL OF INFORMATION SCIENCE AND TECHNOLOGY
Volume 7, Numbers 1–2, 2004, 45–60

Towards Bulgarian Wordnet

Svetla KOEVA, Angel GENOV[1]
Bulgarian Academy of Sciences
E-mail: svetla@ibl.bas.bg
Georgi TOTKOV[2]
Plovdiv University

Abstract. The paper aims at presenting the Bulgarian wordnet which joined the family of the other wordnets relatively recently. The document itself is a report on the preparation and building of the Bulgarian wordnet in the framework of the BalkaNet lexical database. It describes the resources that have been used as the starting data on which Bulgarian wordnet is based, the techniques and tools developed, as well as the current version of the Bulgarian wordnet built in the course of the thirty six months of the project.

1. Introduction

The Bulgarian wordnet [6] has been under development for three years within the framework of the BalkaNet project [10]. The project BalkaNet (Multilingual Semantic Network for the Balkan Languages) aims at constructing a multilingual resource that encodes semantic relations among words in Czech, English and five Balkan languages – Bulgarian, Greek, Serbian, Romanian, Turkish. The list of semantic relations presented in the Bulgarian wordnet (as well as in the other Balkan languages) is based mostly on the Princeton WordNet lexical and conceptual relations [7], [2] and the EuroWordNet language internal relations [14].

In spite of its basic similarity to the English WordNet 2.0, the Bulgarian wordnet is not a mere translation of the English WordNet but has developed its own features, (e.g. with respect to artificial and language-specific concepts, the encoding of some additional grammatical characteristics, the marking of language-specific usage and different stylistic, morphological, or syntactic features, etc.).

[1]Co-author of the section 4.4.
[2]Author of the sections 4.1., 4.2., 4.3.

2. Bulgarian Wordnet – Brief Description

The Bulgarian wordnet models nouns, verbs, adjectives, and (occasionally) adverbs, and contains already 18 810 word senses (towards 1.03.2004), where 35 549 literals have been included (the ratio is 1.89). The distribution of synsets across different parts of speech is shown in Table 1:

Table 1. The distribution of synsets into parts of speech

	Nouns	Verbs	Adjectives	Adverbs	Total
Synsets	12 292	3 564	2 946	8	18 810
Literals	22 002	8 665	4 870	12	35 549

Following the standards accepted in the BalkaNet project the structure of the Bulgarian data base is organized in an XML file.

<SYNSET>
<ID>...</ID><POS>...</POS>
<SYNONYM><LITERAL>...<SENSE>...</SENSE></LITERAL>
</SYNONYM>
<ILR>...<TIPE>...</TYPE></ILR>
<DEF>...</DEF><BCS>...</BCS>
</SYNSET>

Every synset encodes the equivalence relation between several literals (at least one has to be present), having a unique meaning (specified in the SENSE tag value), belonging to one and the same part of speech (specified in the POS tag value), and expressing the same lexical meaning (defined in the DEF tag value). Each synset is related to the corresponding synset in the English WordNet 2.0 via its identification number ID. The common synsets in the Balkan languages are encoded in the tag Base Concepts – BCS. There has to be at least one language-internal relation (there could be more) between a synset and another synset in the monolingual data base. There could also be several optional tags encoding usage, some stylistic, morphological or syntactical features, a stamp marking the person who worked out the particular synset, as well as the last time it was edited.

3. Bulgarian Wordnet – Methodology of Developing

The development of the Bulgarian wordnet has been following a methodology which effectively combines automatic and manual procedures for translation, checking, and correction of the synsets. The methodology used is based partially on the merge model: the selection is done in our linguistics resources, briefly described in section 4; synsets (and some language-internal relations) are first developed separately, afterwards the equivalence relations to English WordNet are assigned and checked by a human expert. The merge approach includes different semi-automatic verifications of coded semantic relations in existing Bulgarian language resources.

The results of the automatic assignment of translated literals, additional synonyms, glosses, and hypernyms showed that the different types of automatic assignment rate differently as regards their correctness and effectiveness. The obtaining of

Bulgarian translation equivalents of the English literals proved a necessary step in the work on the BulNet. While some of the assigned candidates were not correct, the procedure enabled the linguist to perform selection of the candidates instead of looking them up in the dictionary for him/herself after s/he had decided what the meaning of the synset was. The selection of the suitable candidates, the deletion of incorrect candidates, and the addition of missing synonyms was performed using the VisDic tool [9]. Generally the relation of synonymy is set up on the basis of a lexicographer's language intuition and on existing Bulgarian dictionaries of synonyms, then it is checked against the synonymic definitions in the Bulgarian explanatory dictionaries for details, and is finally verified by substitutions in corpora examples or by implementation of the standard tests.

The automatic assignment of hypernyms (and all other relations included in the English WordNet 2.0) proved very helpful, too yet, a linguist was free to correct the proposed candidates. The English hypernyms which are already encoded in the Bulgarian data base are mapped to Bulgarian synsets, but the hypernymy relations may be different in the two languages due to non-lexicalized items. A lexicographer decides whether the hypernymy relation is really true for Bulgarian according to his language intuition, existing Bulgarian explanatory dictionaries, and the implementation of the standard verification tests and if necessary the use of empty synsets eliminates such differences. The other English relations are also assigned automatically to the corresponding Bulgarian synsets, if both synsets that form the relation are presented in the Bulgarian data base. Again a lexicographer decides whether the particular relation exists in Bulgarian depending on his language intuition, existing specialized dictionaries, and verification tests.

Automatic assignment of additional Bulgarian synonyms and interpretation definitions to literals from the machine-readable versions of the available Bulgarian dictionaries could be applied in principle but overgeneration and the low correctness rate of the proposed candidates posed certain problems. The results enabled the team to realize the possibilities and hindrances to using certain machine-readable resources in structuring the synset entries in the Bulgarian wordnet and to apply them accordingly. Thus, the practice of assigning translation equivalents of literals, hypernyms (and other relations) is established, whereas the automatic mapping of glosses and additional synonyms can be carried out for experimental purposes for the time being.

At all stages manual correction by linguists proved necessary for the selection and ordering of the candidates, for the structuring of the definitions and checking the correctness of the semantic relations.

4. Language resources and tools

In order to make the developing of BulNet as reliable as possible, we have gathered a number of machine-readable dictionaries (synonymous, bilingual, explanatory, frequency, etc.) and tools that the lexicographers need in making the decisions necessary in building the Bulgarian data base.

The methodology followed has been performed in six steps:

- Research and gathering of lexical resources and methods used by previous projects;
- Collection, creation and systematization of electronic linguistics resources such as Bulgarian Dictionary of Synonyms, Explanatory Dictionary of Bulgarian, Bilingual Bulgarian – English Dictionary, etc.;
- Compilation, collection and systematization of tools developed in the field of NLP;
- Evaluation of the available linguistic and software resources and the possibilities they offer for semi-automatic developing of the Bulgarian wordnet;
- Design and working out of software tools for creation and editing the BulNet prototypes;
- Development of software tools for wordnet queries and validation of the completeness and consistency of the data base.

4.1. Language Resources

Lexical resources used in natural language processing have evolved from manually-created lexical entries to machine-readable lexical databases and large corpora. Much effort is being applied to the creation of electronic lexicons and electronic linguistic resources in general.

Morphological Dictionaries

Two Bulgarian morphological dictionaries have been developed independently from each other. The Grammatical Dictionary of the Bulgarian language contains over 84 000 citation forms that represent the basic vocabulary of the Bulgarian Literary Language [3] and take into consideration sound alternations. The dictionary allows automatic analysis (and synthesis) of word forms the total number of which amounts to approximately 1 100 000. In the Grammatical Dictionary a given lemma is associated with the name of a Finite State Transducer – FST, which recognizes all corresponding forms of that word and assigns to them the relevant grammatical information. Another implementation is the large word-form lexicon [12] (more than 80 000 entries) containing triples [word-form, lemma, morpho-syntactic code].

Dictionary of Synonyms

The extremely useful lexical resource we relied on is the Bulgarian Dictionary of Synonyms [8], which was digitized and encoded as an MS Access database. A great number of the headwords are polysemous which accounts for the fact that each separate meaning generates a separate synonym set. The relations of synonymy in the dictionary are presented in more than 34 000 synonymy sets. Every word belongs to 1.4 synsets, and every word has 3.4 word synonyms and 0.2 phrase synonyms (on average).

Bilingual Dictionaries

At present our electronic English-Bulgarian Dictionary consists of 58 000 English headwords and corresponding Bulgarian (one or more) translation equivalents. The dictionary provides also information for part of speech and transcription. The electronic Bulgarian-English Dictionary consists of 42 500 Bulgarian headwords and respective English (one or more) translation equivalents. Both dictionaries were additionally corrected and edited.

Explanatory Dictionary

The reference dictionary we used is the Explanatory Dictionary of Bulgarian [1]. This authoritative lexicographic source for contemporary Bulgarian was digitized, corrected, edited and converted into a lexical database (Access encoded). The Explanatory Dictionary of Bulgarian (in its electronic form) consists of 52 434 Bulgarian headwords and corresponding Bulgarian (one or more) explanatory notes. The dictionary provides also information for part of speech, grammatical features, meaning of the word (Bulgarian glosses), stylistic and expressive characteristics, etc.

Frequency Dictionaries

Frequency Dictionaries [4] are extracted automatically from corpora with a program that:

– identifies tokens;

– analyses statistically word forms.

The resulting frequency dictionaries differ from each other according to the corpora that they exemplify.

Corpora

The large Bulgarian corpus [5] consists of approximately 33 000 000 words collected from texts published mainly in electronic form (some texts are scanned).

The texts were selected from different genres and types of prose and poetry. The range of texts in prose covers periodicals, fiction, science fiction, administrative documentation, and scientific texts. The approximate proportion of texts in the corpus is 30 percents literary texts, 50 percents journalistic texts, and 20 percents administrative texts. The major part of the corpus consists of original Bulgarian texts.

The development of Bulgarian structured linguistic corpus [5] was one of the stages of the BalkaNet project. The corpus consists of 1 000 805 words extracted from original Bulgarian texts published mainly in electronic form. The corpus is divided into 500 text units – approximately 2 000 words each, taking into account sentence boundaries. The texts were sampled from 15 different text categories following the model of the Brown corpus.

4.2. Software Tools Assisting the Development of BulNet

Our main purpose was to propose methods and tools for semi-automatic mapping of Bulgarian synsets to English WordNet entries and, hence, for building the BulNet. With reference to this, a system for automatic improvement of the Dictionary of Synonyms [13] and a system for automatic assignment of Bulgarian translation equivalents [12] were developed.

4.2.1. Automatic Improvement of the Dictionary of Synonyms

The automatic improvement of the Dictionary of Synonyms includes finding out different synsets representing one concept, synsets that contain words referring to two or more different concept (mixed synsets), incomplete or incorrect synsets, etc. In order to locate and remove various types of errors, gaps and discrepancies, a distance between an arbitrary couple of synonym sets is introduced and a special editor SplitMerge for splitting/merging synonym sets was created and tested.

The synonym row contains synonyms of one meaning of a given lexeme. The lexeme itself stands in the left-hand part of the row and is called leading lexeme, while its corresponding synonyms occupy the right-hand part of the synonym row. A combination of synonym rows with one and the same leading lexeme will be called synonym paradigm of the lexeme. The following possible inaccuracies in the compilation and editing of synonym rows in the Bulgarian Dictionary of Synonyms may give rise to contradictory descriptions of a concept or the lack of an equivalent row in some synonymy paradigm of the words in the right-hand part of a row:

1. Encoding one concept in more than one row – such rows must be merged into one.
2. Encoding more than one concept in one row ("merging of rows") – such a row must be divided into two (or more) rows.
3. Omitted, added, or misspelled lexeme within the row – such a word must be added, deleted, or corrected (the misused one should be deleted and the right one should be added).
4. Omitted row – such a row must be added.

The objective is to edit the rows denoting the different meanings of the lexemes so that the row descriptions should be complete and should not contain contradictions. We consider that a row description is contradictory if the synonymy paradigm of at least one lexeme in its right-hand part does not contain an equivalent row [13].

When the expert examines the synonymy rows with the SplitMerge editor s/he has to apply one of the following functions: unification of rows, division of rows, movement of lexemes from one row to another and deletion of lexemes. The main menu of the program is shown on Figure 1.

The automatically processed Bulgarian Dictionary of Synonyms at first contained 34 907 synonym rows with 25 103 simple words and 2 272 phrases. After the processing, 14 444 synonym sets are distinguished, each representing different concept and containing the average number of 4.94 words and 0.22 phrases. This result could be successfully used for validation of the Bulgarian wordnet developed.

4.2.2. System for Assignment of Bulgarian Translation Equivalents

A special system – Bulgarian wordnet Extractor, was created to calculate statistics, manage and process rules for forming Bulgarian synsets corresponding to English

WordNet 2.0. The Wordnet Extractor can be used for assessment of the roles of the commas, full stops, semicolons, square brackets, and slashes in a given row from the English Bulgarian Dictionary. Some of the translation equivalents may not be a single word but a phrase, and that phrase could in turn contain commas which are not to be treated as separators. The goal was to design an efficient algorithm that determines which of the commas are separators for the different translation equivalents and which are not.

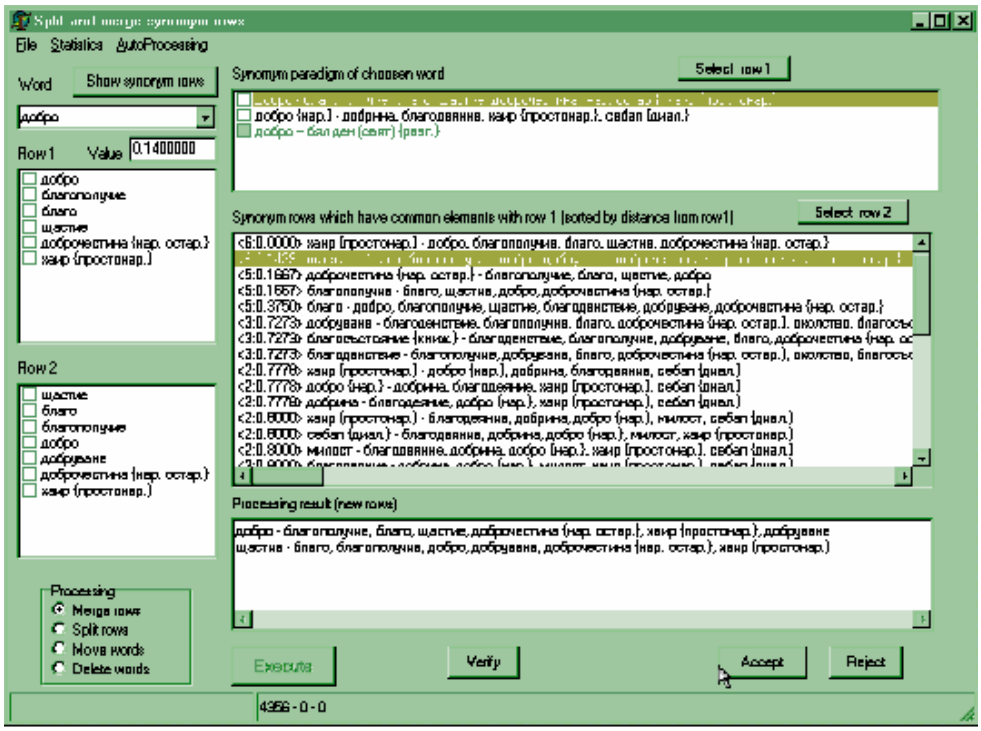

Fig. 1. Main Menu of the SplitMerge program.

In order to automatically extract the different descriptions, a set of rules based upon the morphological characteristics of the Bulgarian words were found and examined. The column of Bulgarian words – in the English Bulgarian Dictionary, was processed by the Bulgarian Morphological Analyzer in order to obtain and place their morphological characteristics in a separate column.

The statistics shows that, for example, when an English word is translated in Bulgarian and the Bulgarian translation consists of two nouns separated by comma (Nc, Nc) in 4 123 of 4 138 cases (more than 99.6 percents) the English word is also a noun which means that the two Bulgarian words are nothing else but two different descriptions of the English one. Only the cases when the part of speech does not match are questionable and need to be checked manually.

The process of the synset extraction can be divided in three steps:

1. Creating rules for automated synset extraction.
2. Automated and semi-automated processing of the English Bulgarian Dictionary with the created rules.
3. Manual processing of the synsets with complex or questionable patterns.

The functional capabilities of the system for edition of the rules (Figure 2) are: automatic synthesis of the "proper" rules, starting with the most reliable ones; options for additional editing of the synthesized rules; representation of all the rows in the English Bulgarian Dictionary corresponding to the edited pattern in "viewing mode"; options for making changes in the respective rows in English Bulgarian Dictionary in "edit mode" and options for successive processing of rows from English Bulgarian Dictionary (one by one or in group) in "apply mode".

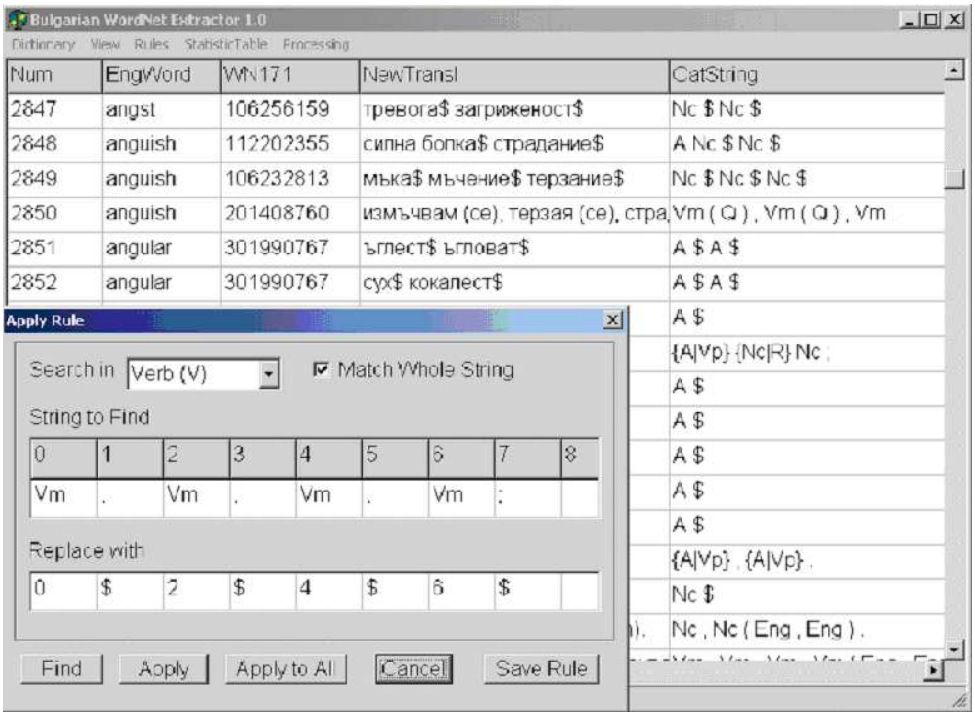

Fig. 2. Wordnet Extractor.

Now, approximately 60 000 rows are automatically processed with 1 200 rules. Two lexical corpora are obtained after the synthesis and the application of the rules on the English Bulgarian Dictionary by Wordnet Extractor. The first lexical corpus consists of separate rows in which the Bulgarian phrases corresponding to one synonym row from the English Bulgarian Dictionary are organized and the second lexical

corpus contains rows that have not been processed by the system. The first lexical resource (more than 50 000 different synonym rows) practically contains a prototype that could be used for further validation of Bulgarian wordnet.

4.3. Wordnet Logic

The encoding of different semantic relation in a multilingual system and the validation of the systems completeness and consistency require more complex mechanisms than the existing ones. For these purposes a specific logic for wordnet [11] was developed. This logic provides sufficient expressive power for all important verifications, queries and consistency, and completeness proofs required for wordnet applications. The syntax and semantics of the wordnet logic is presented in the paper "Bulgarian Wordnet – Structure and Validation", included in this volume.

Bellow we describe some examples of the basic queries and show how they are expressed in the wordnet logic.

4.3.1. Inquiry Unit Queries

- For a given Language returns duplicated literals inside one synset.

$$p^{WS} \;\&\; \langle\equiv\rangle q^{WS} \;\&\; \langle Lit\rangle q^{WS}$$

- Return all ID numbers of Base English synsets which are not included in Bulgarian wordnet

$$p^{Ili} \;\&\; English \;\&\; Base \rightarrow [Ili]\neg Bulgarian$$

- Return all ID numbers of Base English synsets, which in Bulgarian are marked as not Base.

$$p^{Ili} \;\&\; English \;\&\; Base \rightarrow \langle Ili\rangle(Bulgarian \;\&\; \neg Base)$$

$$p^{Ili} \;\&\; Bulgarian \;\&\; Base \rightarrow \langle Ili\rangle(English \;\&\; \neg Base)$$

- Return the Hyperonymy loops inside the wordnet structure

$$p^{\equiv} \;\&\; \langle Hyp^{+}\rangle p^{\equiv}$$

- Return for a given Language and given ID number the corresponding Synonymy Word senses, check if it is a base synset, check if it has a gloss etc.

$$Bulgarian \;\&\; 00001022 \;\&\; p^{WS}$$

$$Bulgarian \;\&\; 00001022 \;\&\; Base$$

$$Bulgarian \;\&\; 00001022 \;\&\; Glos$$

- Return for a given Language the ID numbers of all synsets containing a specified Literal.

$$Bulgarian \ \& \ kaca \ \& \ p^{Ili}$$

- Return for a given Language the ID numbers of all base synsets.

$$Bulgarian \ \& \ Base \ \& \ p^{Ili}$$

- Return for a given Language the ID numbers of all synsets that are in direct hyperonymy / hyponymy relation with the given ID.

$$Bulgarian \ \& \ 00001022 \ \& \ \langle Hyp \rangle p^{Ili}$$

$$Bulgarian \ \& \ 00001022 \ \& \ \langle \widetilde{Hyp} \rangle p^{Ili}$$

- Return for a given Language the ID numbers of all synsets that are in transitive hyperonymy / hyponymy relation with the given ID.

$$Bulgarian \ \& \ 00001022 \ \& \ \langle Hyp^{+} \rangle p^{Ili}$$

$$Bulgarian \ \& \ 00001022 \ \& \ \langle \widetilde{Hyp}^{+} \rangle p^{Ili}$$

- Return for a given Language the ID numbers of all synsets which have no Hyperonyms / Hyponyms.

$$Bulgarian \ \& \ p^{Ili} \ \& \ \neg \langle Hyp \rangle q^{WS}$$

$$Bulgarian \ \& \ p^{Ili} \ \& \ \neg \langle \widetilde{Hyp} \rangle q^{WS}$$

- Return for a given Language the ID numbers of all synsets which have no Gloss defined.

$$Bulgarian \ \& \ p^{Ili} \ \& \ \neg Glos$$

- Return for a given language the ID numbers of all synsets which have different hyperonyms in another language

$$p^{\equiv} \ \& \ Bulgarian \ \& \ \langle Hyp \rangle q^{Ili} \rightarrow \neg \langle Ili \rangle (English \ \& \ \langle Hyp \rangle q^{Ili})$$

$$p^{\equiv} \ \& \ English \ \& \ \langle Hyp \rangle q^{Ili} \rightarrow \neg \langle Ili \rangle (Bulgarian \ \& \ \langle Hyp \rangle q^{Ili})$$

- Return all literals with more than k senses.

$$p^{Lit} \ \&_{i=1}^{k} \ \langle Lit \rangle p_i^{WS} \ \&_{i<j} \ \langle Lit \rangle (p_i^{WS} \ \& \ [\equiv] \neg p_j^{WS})$$

4.3.2. Wordnet Validator

The Wordnet Validator is a Web-based system for validation (and correction) of the wordnet completeness and consistency. In the Wordnet Validator the predefined queries as the exemplified above are used. The system works with the adopted XML-file format.

The Wordnet Validator has the following main functions:

1. Automatic correction of XML syntax;
2. Validation of wordnet completeness and consistency;
3. Search for a given synset;
4. Visualization of semantic trees.

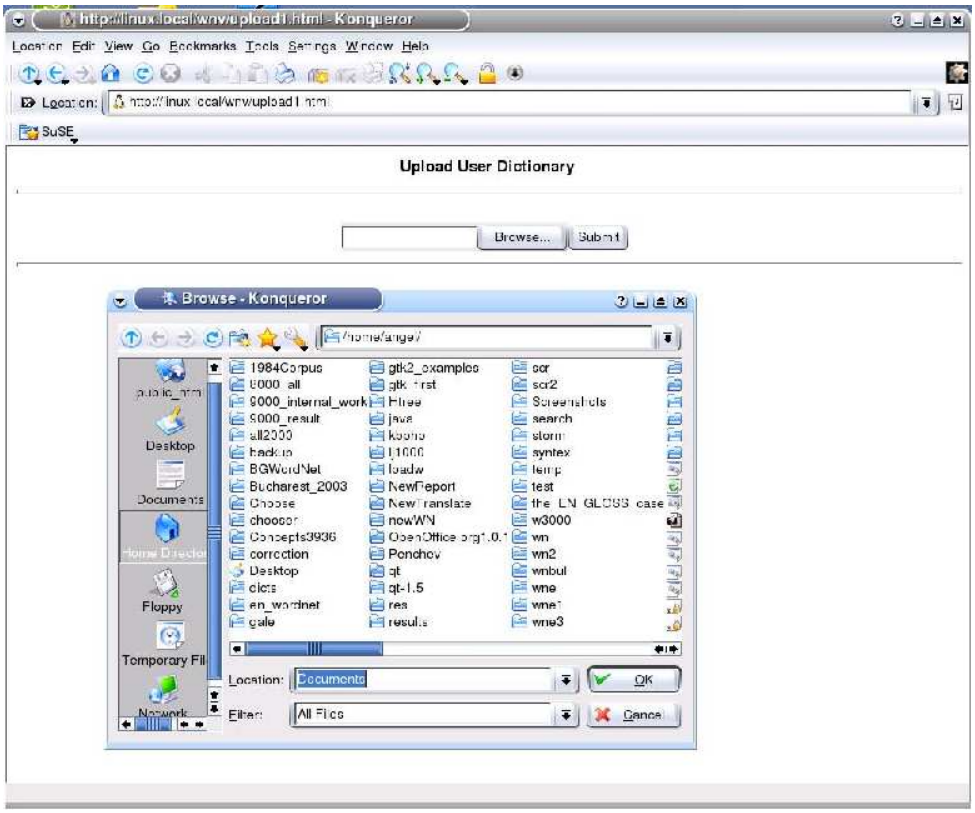

Fig. 3. Wordnet Validator Browse Function.

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 3). The

browsed language is accepted if it corresponds to several conditions: an appropriate XML format, no empty ID tags and no duplicated ID's.

Our approach to the validation of wordnets includes three separate levels presenting different degrees of complexity and significance and different possibilities for automatic data correction. The three levels include checking the syntax of the XML files, checking the completeness of the wordnets, and validating the consistency in defining the semantic relations.

In the following cases the automatic correction function 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) – validating the presence of all members from the chosen up to now Base Concepts within the framework of the BalkaNet project; check "dangling" relations – checking whether both members of the defined relation are present in the wordnet; check "gaps" – verifying that for a certain synset included in the wordnet all levels of hypernyms up to the top of the tree are present.

- *Verifying the consistency of the data*: check ID format; check synsets without DEF tags; check synsets without literals; check duplicated relations – checking whether there are duplicated relations between two synsets in a language; check differences in relations – finding all synsets whose hypernyms differ from the corresponding ones in the second selected language, this check may be broadened to cover all relations; check loops – verifying for lack of hypernym cycles, as well as any relation 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.

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.

5. Bulgarian Wordnet – Current State

Bulgarian wordnet contains 18 810 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 35 549 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 27 244 – 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 included in the Bulgarian wordnet are seventeen (following the Princeton WordNet), their occurrences are 31 980, the average number of relations per synset is 1.7. The figures representing the current state of the Bulgarian wordnet are exemplified in the Table 2.

Table 2. Statistical data characterizing BulNet

	Nouns	Verbs	Adjectives	Adverbs	Total
Synsets	12 292	3 564	2 946	8	18 810
Literals	22 002	8 665	4 870	12	35 549
Literals/synsets	1.79	2.43	1.65	1.5	1.89
Graphic words	18 026	5 523	3 647	12	27 244
Literals/words	1.22	1.57	1.34	1	1.3
ILR	19 488	8 343	4 146	3	31 980
ILR per synset	1.59	2.34	1.4	0.38	1.7
Definitions	12 292	3 564	2 946	8	18 810

Each synset included in the wordnet is part of a semantic tree which consists of chains of hyponymy and hypernymy relations. Those hierarchical structures implement and apply inheritance and finish with very complex entities (tops) which are unobservable and cannot be grasped, heard, seen, or felt as an independent physical thing. The leaves of the tree are concrete, perceivable by the senses and located at any point in time. 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 3). The hierarchies for both nouns and verbs are quite deep. The average density for Bulgarian noun treee is 1 365.78 (in English WordNet 2.0 it is 8 854.33), and the average density for Bulgarian verb trees is 9.16 (compared to 24.38 for English WordNet2.0).

Table 3. Number of tops per Bulgarian nouns and verbs

WN	N nodes	Tops N	V nodes	Tops V
Eng 2.0	79 689	9	13 508	554
BulNet	12 292	9	3 564	389

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, PARTICIPLE, and some extralinguistic ones: REGION DOMAIN, USAGE DOMAIN, CATEGORY DOMAIN.

The hypernym-hyponym relation rates highest in terms of number of occurrences – 15 855 in 18 810 synsets – approximately 84 percent of the total number of synsets are assigned hypernyms. The distribution of the BulNet relations in comparison with the English WordNet 2.0 is shown in Table 4.

Table 4. Distribution of the BulNet relations

ILR	POS/POS	EW2.0	BulNet
ALSO SEE	A/A V/V	3 240	732
BE IN STATE	A/N	1 296	589
BG DERIVATIVE	N/V	36 630	6 523
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	103
DERIVED	A/N	6 809	1 148
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	15 855
HYPONYM	N/N V/V	94 844	15 855
IS CAUSED BY	V/V	439	103
IS DERIVED FROM	N/A	6 809	1 148
IS STATE OF	N/A	1 296	589
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	1 824
PARTICIPLE	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	1 293
SUBEVENT	V/V	409	162
VERB GROUP	V/V	1 748	842
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	18 810
L NOTE		0	1 520
LITERAL		203 147	35 549
POS		115 424	18 810
SENSE		203 147	35 549
S NOTE		0	125
SYNONYM		115 424	18 810
SYNSET		115 424	18 810
USAGE		48 231	8 816

6. Conclusions

- The Bulgarian wordnet consists of 18 810 synsets (toward 01.03.2004)
- The distribution across part of speech is about 65.35 percents nouns and 18.95 percents verbs. There are 2 946 adjectives and 8 adverbs.
- The average synset in the Bulgarian wordnet contains 1.89 variants.
- Nouns have 9 top nodes, whereas verbs have 389 top nodes.
- The language-internal relations and the links to the ID are highly reliable, since they have been encoded manually.
- 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, PARTICIPLE, and some extralinguistic ones: REGION DOMAIN, USAGE DOMAIN, CATEGORY DOMAIN.
- The Bulgarian wordnet has a ratio of 1.7 language-internal relations per synset.
- The verification methodology is formulated and applied to the Bulgarian data – in result, the Bulgarian wordnet is complete and consistent according to the requirements and the specifications defined in the BalkaNet project.

References

[1] ANDREJCHIN, L. (ed.), *Bulgarian Explanatory Dictionary*, Sofia, Nauka i Izkustvo, 1999.

[2] FELLBAUM, C. (ed.), *WordNet: An Electronic Lexical Database*. Cambridge, Mass., MIT Press, 1998.

[3] KOEVA, S., *Bulgarian Grammatical Dictionary*, Organization of the language data, Bulgarian language, 1998, vol. **6**, 49–58.

[4] KOEVA, S., *Automatic generation of lexeme frequency dictionaries – For Words and Dictionaries*, Lexicology and lexicography' 98, Sofia, 2000, Academic Press, 109–117.

[5] KOEVA, S., PLACHKOVA, P., STOJANOVA, I., LESSEVA, S., *Bulgarian language resources – Cyrillic and Latin OCR correction using electronic dictionaries and sentence context*, Sofia, 2002.

[6] KOEVA S., *Bulgarian Wordnet – Bulgarian Studies*, Ohio, 2003.

[7] MILLER, G. A., *Introduction to WordNet: An On-Line Lexical Database*, International Journal of Lexicography, Miller G.A., Beckwidth R., Fellbaum C., Gross D., Miller K.J., **3**, No. 4, 1990, 235–244.

[8] NANOV, L., NANOVA, A., *Bulgarian Synonym Dictionary*, Sofia, Hejzal, 2000.

[9] PAVELEK, T., PALA K., *VisDic: A New Tool for Wordnet Editing*, in *Proceedings of the 1st International Wordnet Conference*, Mysore, India, January 21–25, 2002, 21–25.

[10] STAMOU, S., OFLAZER, K., PALA, K., CHRISTODOULAKIS, D., CRISTEA, D., TUFIŞ, D., KOEVA, S., TOTKOV, G., DUTOIT, D., GRIGORIADOU, M., *BALKANET: A Multilingual Semantic Network for the Balkan Languages*, *Proceedings of the International Wordnet Conference*, Mysore, India, January 21–25, 2002, 12–14.

[11] TINCHEV, T., MIHOV, S., KOEVA, S., GENOV, A., *Logic for WordNet*, Annuaire of Sofia University, **95**, 2003.

[12] TOTKOV, G., *Towards Building Bulgarian Wordnet: Language Resources and Tools*, in *Proceeding of the Intern. Conf. ICTP'2003*, Varna, June 23–26, 31–40.

[13] TOTKOV, G., IVANOVA, P., RISKOV, I., *Automated Improving and Forming WordNet Synsets on Conventional (non computer based) Synonym and Bilingual Dictionaries*, in *Comp. Ling. and its Applications* (A. Narin'iyani, ed.), DIALOGUE'2003, Protvino, June 2003.

[14] VOSSEN, P. (ed.), *EuroWordNet: a multilingual database with lexical semantic networks for European Languages*, Kluwer Academic Publishers, Dordrecht, 1999.