BULGARIAN SENSE TAGGED CORPUS

Svetla Koeva, Svetlozara Lesseva, Maria Todorova

Department of Computational Linguistics, IBL - BAS
52 Shipchenski prohod, Bl. 17, Sofia 1113, Bulgaria
svetla@ibl.bas.bg, zara@ibl.bas.bg, maria@ibl.bas.bg

Abstract

The Bulgarian Sense Tagged Corpus is derived from the "Brown" Corpus of Bulgarian and annotated with word senses from the Bulgarian WordNet. The paper gives a brief account of the already available and currently developed language resources and tools which enabled the compilation and annotation of the Bulgarian Sense Tagged Corpus. We briefly describe the adopted methodology for constructing and preprocessing the source corpus of 63 440 words: all words were lemmatised, PoS-tagged and linked to the corresponding sets of senses in the Bulgarian WordNet. The paper also presents the annotation criteria underlying the sense selection process and outlines the general directions of expansion and modification of the Bulgarian WordNet. At the present stage 45 562 words (single words and multi-word expressions) are semantically annotated. The chief intended application of the Bulgarian Sense Tagged Corpus is to serve as a test and / or training dataset for word sense disambiguation with the further aim of developing a Bulgarian - English bi-directional machine translation system.

1. Introduction

The main objective of this paper is to present the Bulgarian Sense Tagged Corpus (BulSemCor) derived from the "Brown" Corpus of Bulgarian and annotated with word senses from the Bulgarian WordNet (BulNet)[1]. The chief intended application of the Bulgarian Sense Tagged Corpus is to serve as a test and / or training dataset for word sense disambiguation with the further aim of employing the results in the implementation of a Bulgarian-English bidirectional machine translation (MT) system. The paper also gives a brief account of the already available and currently developed language resources and tools which enabled the compilation and annotation of BulSemCor.

It is generally acknowledged that statistical approaches (completely or partially underlying any disambiguation process) can be efficiently combined with data derived from annotated corpora for testing and / or training as within the so-called supervised corpus-based methods. The statement that "a logical next step for the research community would be to direct efforts towards increasing the size of annotated training collections, while deemphasizing the focus on comparing different learning techniques trained only on small training corpora" (Banko & Brill, 2001) fully confirms our understanding of how part of speech and word sense disambiguation (WSD) should be handled. Therefore, effort should be concentrated in the devising of a balanced combination of the currently employed methods that will be able to yield a strong positive impact on the effectiveness of WSD.

In the compilation of BulSemCor we generally follow the methodology adopted for the English semantically annotated corpus – SemCor, created at the Princeton University (Fellbaum et al., 1998). The latter is a subset of the Brown Corpus of Standard American English containing almost 700 000 running words. All the words in SemCor are PoS-tagged, and more than 200 000 content words are additionally lemmatized and tagged with Princeton WordNet senses.

[1]The investigation is developed under the national funded project *"BulNet – Lexical-semantic Network of the Bulgarian Language"*.

2. Bulgarian resources

Likewise, our target corpus for semantic annotation is a subset of the "Brown" Corpus of Bulgarian (BCB) (Koeva et al., 2005a). BCB consists of 500 corpus units of approximately 2000 words each, distributed proportionally to language use in 15 categories, thus forming an overall of 1 001 286 words. The methodology of the developing of Brown Corpus of Bulgarian is as close as possible to the original Brown corpus in terms of structure and content, but still it differs in some respects: some categories either partially or not at all represented in contemporary Bulgarian language use were replaced by more appropriate ones. The sub-corpus for sense annotation preserves the original structure of BCB by including a section of each BCB unit sampled according to the density of high frequency words.

The linguistic database which serves as a source for introducing and resolving ambiguity is the Bulgarian WordNet - BulNet (Koeva, 2004a). Synsets (as basic structural units of wordnet) are equivalence sets containing a number of obligatory elements: literals (single words and multi-word expressions (MWE) with the same referential meaning, expressed by an interpretative definition, usage examples and language notes. The synsets are interrelated in a lexical-semantic network – wordnet, by means of a set of semantic relations such as hyperonymy, antonymy, meronymy, etc. EuroWordNet (EWN) extended the Princeton WordNet (PWN) with cross-lingual relations, which were further adopted in BalkaNet (BWN) (Stamou et al., 2002) and by the Bulgarian WordNet as part of it. The equivalent synsets in the different languages are mapped to the same Inter-Lingual Index (ILI), thus connecting the individual wordnets in a global lexical-semantic network. The Inter-Lingual Index is based on PWN and is consecutively synchronized with new PWN versions.

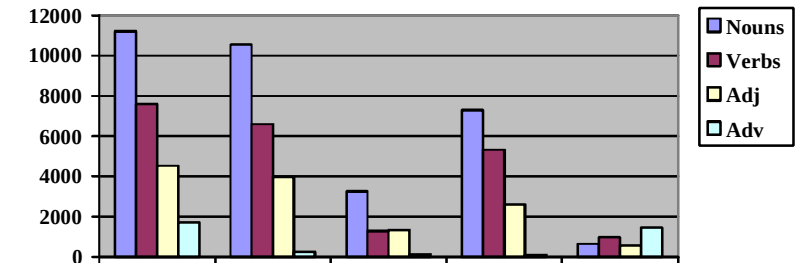
At the moment BulNet consists of 27 045 synsets (synonym sets), containing 57 496 literals, and the average number of literals per synset is 2.12. The language-internal relations encoded in the Bulgarian WordNet are seventeen (following the Princeton WordNet), their occurrences are 48 371, the average number of relations per synset is 1.79.

3. Development and pre-processing of the source corpus

The annotation corpus consists of 500 excerpts (clippings) of approximately 100 words each, selected according to a criterion for well-balanced density of highest frequency Bulgarian open-class lemmas located in BCB. The calculation of the frequency list is based on the occurrences of content words in two Bulgarian POS disambiguated corpora – 71 876 words from Orwell's *1984* and a selection of 328 964 words from three thematic domains – economy, law and politics (400 840 words altogether).

The task of constructing the corpus for annotation consisted in the selection of a 100-word excerpt (clipping) from every file in the "Brown" Corpus of Bulgarian such that would contain the highest density of words from the frequency list. The selection procedure involved several experiments with frequency lists of different sizes derived from the original one by consecutively excluding words occurring one, two and three times. Relative weights were assigned to the lemmas featuring on the lists which were further modified proportionally to the frequency of the lemmas' occurrence both in BCB and in BulNet, so that less frequent words have greater weights. Further, additional weights were calculated according to part of speech as follows: 0.4 to nouns, 0.3 to verbs, 0.2 to adjectives and 0.1 to adverbs in order to provide a better balance in the proportion of nouns and verbs in comparison with adjectives and adverbs. After the clippings' selection the following statistics was made: nouns, verbs, adjectives and adverbs were divided in two groups depending on their occurrence on the frequency lists. For each group the number of words encountered in the wordnet and the number of multiple-sense words was calculated. Subsequently, the clippings that had the best lexical coverage estimated in terms of the greatest number of different lemmas in combination with the greatest number of corresponding wordnet senses were selected for the corpus.

The resulting corpus was further enlarged by expanding the clippings to the left and right sentence boundaries, thus amounting to a total of 63 440 words. The word forms in the source corpus were lemmatized, PoS-tagged and linked to the corresponding sets of senses in BulNet, if available. 6 031 lemmas were automatically linked to only one sense, 3 704 lemmas left without a sense matched in BulNet and 15 343 lemmas received more than one sense. Figure 1 below shows the distribution of open class lemmas in the resulting corpus across part of speech and the coverage of the same lemmas in the Bulgarian WordNet. Outside these figures remain the function words which had to be additionally encoded. In the course of annotation single-sense entries are subject to validation and possibly new senses for such lemmas are encoded where needed; for the lemmas not having a corresponding entry a new synset denoting the appropriate sense is to be included in BulNet (or the sense of an already existing synset has to be revised) and then associated with the word; for multiple sense lemmas the particular sense used in the context has to be picked up, or if not available - encoded.



	All words	WN senses	1 sense	>1 sense	No sense
Nouns	11225	10557	3262	7295	668
Verbs	7606	6612	1291	5321	994
Adj	4525	3950	1335	2615	575
Adv	1722	255	143	112	1467

Figure 1: Distribution of content-word lemmas across POS and coverage in BulNet

4. Annotation tool

Sense annotation is conducted with the annotation tool Chooser developed at the Department of Computational Linguistics (DCL)[2]. Chooser was designed as multi-purpose multi-functional platform aimed at performing various tasks that require corpora annotation as well as at enabling automatic analysis and manual disambiguation of large volumes of text (Koeva et al., 2005a). Figure 2 below shows Chooser's layout. The application's visualisation and editing functionalities provide text display management and a number of other functions such as: text navigation according to various strategies, selection of particular options available for particular language units, group selection of adjacent or distant units (such as multi-words expressions, expressions whose constituents can be intervened by other words, etc.). The corpus for annotation is displayed in the left top window, the synchronization with the other windows is instantly initiated on navigating along the text. On selecting a current word (coloured red on the picture) the definitions of the senses available in BulNet for the word are displayed in the bottom window.

[2] Borislav Rizov from DCL has programmed the annotation tool.

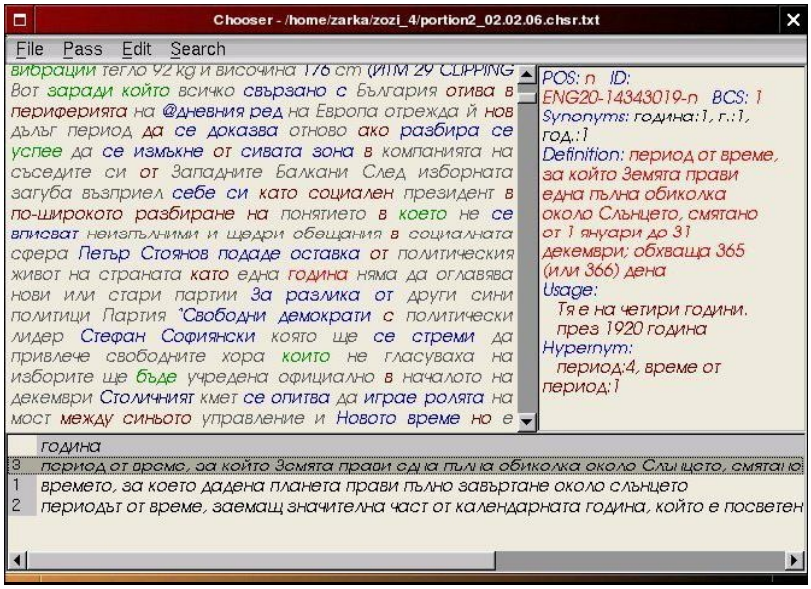


Figure 2: Layout of the annotation tool

The right upper window shows all the information for the sense corresponding to the selected item in the bottom window including the synonym set, definition, usage examples and the relations available in BulNet. The set of choices available for a given word can be ordered according to different criteria, presently the adopted one is the frequency of selection of the sense in the process of word sense annotation.

The application design envisages a number of pass strategies such as: passing all language units in the corpus, stopping at language units which are associated with certain information in the linguistic database, passing only ambiguous language units, language units that have been modified in the database since the last selection of the same item by the current user, or stopping at all occurrences of a particular item.

A major asset of the annotation framework is that it handles single as well as multiword expressions (MWE) referring to a single concept (i.e. *New York, nitric acid,* etc.) regardless of their structural or constituent variations, treating such expressions as single strings with spaces at certain places. The tool allows selection of adjacent as well as of distant MWE constituents, thus managing specific syntactic properties of multiword expressions - intervening words between constituents, varying constituents, different word order, syntactic transformations, etc. This has been pointed out as one of the main problems of encoding of MWE in WordNet (Fellbaum, 1998). The components of a MWE in the corpus are associated with their lemmas through which the tool makes correspondence to the relevant literal in BulNet.

Chooser is a multiple-user platform that performs dynamic interaction between the local users. User communication is implemented by means of a server that takes care of a number of activities in two principal directions:

- Interaction between the local users and the linguistic database;
- Interaction between the local users.

Chooser's database is permanently updated with enlarged BulNet versions contributed by the individual annotators. Thus, newly-entered and edited senses are imported for selection by the annotators, and passed along via the option of navigation along unselected items. By taking care of the frequency update and the respective reordering of the senses the tool keeps track of and stores valuable information about language data while facilitating the annotators' work.

5. Development of the Bulgarian Sense Tagged Corpus

After the processing of the source corpus, annotation of the language units in the corpus with the correct senses in the Bulgarian WordNet is performed.

For tagged words in the corpus the following outputs are produced:

- For nouns, verbs, adjectives and adverbs - Word, Lemma, Sense identification including the ID number and POS of the corresponding sense in BulNet.
- For multi-word expressions - MWE, Lemma, Sense identification including the ID number and POS of the corresponding sense in BulNet.
- For function words - Word, Lemma, Sense identification.

For example the Bulgarian sentence *Obichal mekiya kaliforniyski klimat* (He loved the soft Californian climate) will be annotated as follows:
Obichal{obicham#ENG20-01723774-v}
(love:have a great affection or liking for)
mekiya{mek#ENG20-00411931-a}
(mild: mild and pleasant)
kaliforniyski{kaliforniyski#ENG-02893758-a}
(Californian: of or related to or characteristic of California or its inhabitants)
klimat{klimat#ENG20-13692717-n}
(climate: the weather in some location averaged over some long period of time)

Sense Tagged Corpus	
New senses added in BulNet	5 328
Annotated units	45 562
Annotated single words	40 255
Annotated MWE	2 177
Words left for annotation	17 878

Table 1: Current state of BulSemCor

The current state of BulSemCor includes 45 562 semantically annotated single words and multi-word expressions (Table 1), of which 40 255 are single words and 2 177 - MWE. The average length of MWE is 2.19 words.

All the words initially assigned only one sense are considered annotated after the validation of the sense mapping. In the course of the annotation 5 328 new synsets have been added in BulNet so far. New senses are encoded where a corpus occurrence is not mapped in BulNet at all, or if the available candidates (single-sense or multiple-sense BulNet entries) do not match the meaning found in the corpus.

6. Annotation criteria

The annotation of the senses consists in the association of word occurrences in the corpus – single words and multiword expressions - with the appropriate senses in BulNet[3]. Coverage is ensured through the evaluation of the encoded data against the empirical evidence from the corpus and the respective revision and enlargement of BulNet with new senses. New literals and synsets are either such found in PWN or ones having no equivalent in PWN. In the latter case new BulNet-specific entries are created. Apart from this, optimisation of the encoded language data in BulNet is performed.

6.1. Selection of senses

Under this heading we discuss the implicit consistency criteria involved in the annotators' choice of a given sense of a graphical word (literal) from among the available candidates in BulNet. These procedures (or analogous ones) are extensively applied where a word in a language has a number of closely related senses:

6.1.1. Consistency with the other (if any) members of the synset

In deciding which is the most appropriate among the candidate senses, the first thing to be considered is the relation of equivalence defined between the members of a synset. This means that if an instance of a word in the corpus is semantically equivalent to an instance of another word in the same context, it is most likely that the correct sense is the one that corresponds to the synset where the two items appear as synonyms. Of course, cross-check with other criteria is performed even in this case, to avoid possible errors due to incompleteness in the database.

3 Ekaterina Tarpomanova and Hristina Kukova from the Department of Computational Linguistics (IBL-BAS) and Katya Alahverdzhieva and Nikolay Radnev, students at Sofia University, also worked in different capacity as annotators.

6.1.2. Consistency with the interpretative definition covering the general meaning of the synset

The interpretative definition (gloss) associated with the synset encodes the meaning of all the members of the synset in an explicit way, hence it is a principal clue in choosing between senses.

6.1.3. Consistency with the relative position of the synset in the overall wordnet structure

Unlike the previous criteria which establish the association between an instance of a word and a synset in BulNet according to the linguistic information contained in the synset, this one employs the degree of relatedness between pairs of synsets and is hence very helpful where a word has a number of closely related meanings. Similarity may well be signaled by identical or very close synonym sets and definitions. However, distinctness between similar lexical items will (or at least should) be observed in the different set of relations defined for a synset. Relatedness involves relations of similarity between semantically similar items, as well as other types of semantic relations (meronymy, antonymy, etc.) between dissimilar units (Budanitsky & Hirst, 2001). Hence, the exploration of the set of semantic relations encoded for the examined synset may provide helpful clues for the annotators.

The following examples illustrate the interaction of the three criteria:

Synonyms: *{nature:1}*

Definition: *the essential qualities or characteristics by which something is recognized*

Usage: *it is the nature of fire to burn*

Hypernym: *{quality:1}*

Synonyms: *{nature:3}*

Definition: *the natural physical world including plants and animals and landscapes etc.*

Usage: *they tried to preserve nature as they found it*

Hypernym: *{universe:1, existence:2, creation:6, world:2, cosmos:1, macrocosm:1}*

On looking at the Bulgarian counterparts one can see that the Bulgarian synset corresponding to *{nature:1}* has two members – *{estestvo:1, priroda:3},* and that corresponding to *{nature:3}* – one – *{priroda:2}*. The two senses are further distinguished by the glosses and the hyperonyms defined for the synsets. The usage examples also account for the distinction between the senses.

6.1.4. Consistency with the usage examples

Besides illustrating the context of use of a word, usage examples provide a quick way of scanning through and checking different senses of a word as well as of potential candidates for encoding. They are especially helpful in

cases of similar synonym sets and / or unclear definitions as in the example given below:

Synonyms: *{disorder:1, upset:3}*

Definition: *condition in which there is a disturbance of normal functioning*

Usage: *the doctor prescribed some medicine for the disorder*

Hypernym: *{condition:1, status:2}*

The definition and the synonyms do not at first sight help to infer the meaning of the synset in this example. Scanning the usage examples, together with hyponyms, such as *{immunological disorder:1}*, *{cardiovascular disease:1}*, etc. help the annotator grasp the meaning at once.

6.1.5. Consistency with grammatical features accounting for sense distinctions

Certain sense distinctions may be suggested by grammatical differences. For example, the plural form of a noun signifying a member of a nation may stand for the relevant nation as well, as in *The Brits are a great nation* where the sense assigned to *the Brits* corresponds to:

Synset: *{British:1, British people:1, the British:1, Brits:1}*

Definition: *the people of Great Britain*

Hypernym: *{nation:2, land:8, country:3 a people:1}*

whereas the phrase *two Brits* in *Two Brits were rescued* is semantically equivalent to:

Synset: *{Britisher:1, Briton:1, Brit:1}*

Definition: *a native or inhabitant of Great Britain*

Hypernym: *{European:1}*

Hence, on coming across similar instances in the corpus, one should bear in mind this distinction and correctly assign the appropriate sense. It should be noted that since different lemmas (sg. and pl.) correspond to the considered literals, as well as to their Bulgarian counterparts, and lemmas are the mediator between the BulNet entries and the annotation tool, the appropriate lemmatization of the words in the corpus is a prerequisite for the generation of correct lists of choices. To be more particular, if the *Brits* in the first sentence is lemmatized as *Brit,* the synset featuring *Brits* will not be on the list of choices at all. The functionality of the annotation tool allows the system's update on manual corrections in the corpus if necessary and new set of choices is subsequently generated.

6.1.6. Appropriateness with respect to the available senses encoded in PWN

While the criteria (1-5) refer to the exploration of the senses already encoded in BulNet, this one applies mainly to the cases where the corpus occurrence might not be an instance of any of the senses present in the Bulgarian database.

For example, *nature* in the sentence *Nature has taken care of us for centuries and we are still discovering her many wonders* is not an instance of any of the senses encoded for *nature*, discussed above. On exploring PWN one can see that the sense *nature:2* corresponds precisely to the meaning of the word in the sentence:

Synonyms: *{nature:2}*

Definition: *a causal agent creating and controlling things in the universe*

Usage: *Nature has seen to it that men are stronger than women.*

Hypernym: *{causal agent:1, cause:4, causal agency:1}*

Appropriateness of the choice is also considered with respect to specific cases of language use. It is not infrequently the case that a more general and a terminological sense are overlapping. Therefore, special consideration to the type of annotated text should be involved in choosing between senses such as:

Synonyms: *{water:1, H2O:1}*

Definition: *binary compound that occurs at room temperature as a clear colorless odorless tasteless liquid; freezes into ice below 0 degrees centigrade and boils above 100 degrees centigrade; widely used as a solvent*

Hypernym: *{binary compound:1}*

Hypernym: *{liquid:3}*

Synonyms: *{water:6}*

Definition: *a fluid necessary for the life of most animals and plants*

Usage: *he asked for a drink of water*

Hypernym: *{food:1, nutrient:1}*

6.2. Expanding the knowledge base - BulNet

The knowledge base BulNet is expanded in two principal directions: encoding of new entries found in PWN where a relevant occurrence in the corpus requires that in compliance with criterion 6.1.6, and encoding of BulNet-specific entries which fall into several categories:

6.2.1. Culture-specific concepts

In the development of the individual wordnets the BWN adopted the hierarchy of concepts and the structure of the relations established in the construction of the English WordNet. Hence, a strong rule for the preservation of the PWN structure has been strictly observed as a way of ensuring a proper cross-lingual correspondence and navigation via the ILI. Naturally, not all concepts stored in the ILI are lexicalized in all languages, and besides, there are language-specific concepts that might have no ILI equivalent. The structure preservation rule requires that in the first case empty synsets be created (called non-lexicalized synsets) in the wordnets of the languages that do not lexicalize the respective concepts. Thus, the non-lexicalized synsets preserve the hierarchy and cover the proper cross-lingual relations. In the second case culture-specific concepts not featuring in the English database are encoded such as:

Synonym:*{bogomilstvo:1}*

Definition: *an orthodox heretic sect founded by the Bulgarian priest Bogomil*

Hypernym: *{heresy:2 unorthodoxy:2}*

The adopted methodology for the incorporation of such concepts involved the further extension of the ILI with new records. The language-specific concepts shared among Balkan languages were linked via a BILI (BalkaNet ILI) index (Tufis et al., 2004). The initial set of common Balkan specific concepts consisted mainly of concepts reflecting the cultural specifics of the Balkans (family relations, religious objects and practices, traditional food, clothes, occupations, arts, important events, measures, etc).

6.2.2. Language-specific instances of lexicalization

Beside culture-specific concepts the semantic annotation involves the encoding of single words or

MWEs that are not lexicalized in English, for example: *stamva se* whose English counterpart is *get dark* and is not present in the PWN, as contrasted with its antonym *samva se - dawn* which is lexicalized in English. A systematically occurring case is presented by ingressive verbs in Bulgarian formed with a prefix which correspond to compositionally formed expressions in English.

There are four morpho-semantic relations included in PWN and mirrored in EWN and BWN, *Be in state*, *Derivative*, *Derived* and *Participle* (Koeva, 2004). These relations semantically link synsets although they can actually be encoded between pairs of literals (graphic and compound lemmas). There are systematic morpho-semantic differences between English and Slavic languages such as certain derivational mechanisms for forming classifying adjectives, gender pairs and diminutives. The Slavic languages possess rich derivational morphology which has to be incorporated into the strict one-to-one mapping with the ILI.

A productive derivational feature is the formation of classifying adjectives from Bulgarian nouns with the general meaning *'of or related to the noun'*. For example, the Bulgarian adjective

Synset: *{stomanen*:1}

Defintion: *made of or related to steel*

is expressed in English by the respective noun used attributively (rarely at the derivational level, consider *wooden ↔ wood, golden ↔ gold*), thus the concepts exist in English and the mirror nodes have to be envisaged.

The gender pairing is a systematic phenomenon in Bulgarian and other Slavonic languages that display binary morpho-semantic opposition: male ↔ female, and as a general rule there is no corresponding concept lexicalized in English. The derivation is applied mainly to nouns expressing professional occupations. For example, the masculine nouns in the Bulgarian synset *{prepodavatel:2, uchitel:1, instruktor:1}* corresponding to the English:

Synset: *{teacher:1, instructor:1}*

Definition: *a person whose occupation is teaching*

have their female gender counterparts *{prepodavatelka, uchitelka, instruktorka}* with a feasible definition '*a female person whose occupation is teaching*'.

Diminutives are another language-specific derivational class used to express concepts that relate to small things. The diminutives display a sort of morpho-semantic opposition: big ↔ small, however sometimes they may express an emotional attitude, too. Thus the following cases can be found with diminutives: standard relation big thing ↔ small thing ↔ small thing to which an emotional attitude is expressed, consider *{stol:1}* corresponding to English :

Synset: *{chair:1}*

Definition: *a seat for one person, with a support for the back*

and *{stolche}* with an feasible meaning '*a little seat for one person, with a support for the back*' and *{stolchence}* with a meaning '*a "dear" little seat for one person, with a support for the back*'.

6.2.3. Missing English senses and unaccounted systematic differences between senses

Cases where an English sense (attested in dictionaries and known to be the lexical equivalent of a particular Bulgarian sense) is not present in PWN fall into this category. A prominent example is presented by causative and inchoative verbs, which although in general are encoded in PWN and interrelated by means of the Causes relation, are sometimes either mingled or not represented at all:

Synset: *{modernize:2, modernize:1, develop:11}*

Definition: *become more technologically advanced*

The corresponding transitive verb used in *They modernized the cities* does not feature in PWN. In this case our approach is to encode the sense as a separate synset and link it through the Causes relation to its transitive or intransitive counterpart.

6.2.4. Closed word classes

For the purposes of WSD, BulNet is artificially being expanded to incorporate in a systematic way the classes of prepositions, conjunctions, pronouns, particles, modal verbs, etc. The distinction between the senses is based on the analysis of the syntactic evidence and the semantic features observed in the tagged corpus and the senses registered in different Bulgarian lexicographic and grammatical works.

The existing classifications of closed-word classes are sometimes overlapping, not precise enough or based on unclear criteria. This necessitated the elaboration of classifications for the different function word classes that give an adequate account for the sense distinctions found in language use. For example, high-granularity sense distinctions for the class of prepositions has been initiated, based on semantic roles, such as instrument, location, direction, addressee, etc. Thus, for example, one of the 22 senses encoded for one of the highly polysemous Bulgarian preposition *{na}* is defined in the following way:

Synset: *{na:4}*

Definition: *a preposition that introduces the receiver or addressee or beneficiary, etc. of the action*

For some of the closed-word classes, existing entries in PWN have to be considered, to ensure consistency between the Bulgarian and the English databases. The traditional classification of the Bulgarian pronominal system subsumes classes of words with adjectival or adverbial functions whose English equivalents are encoded as adjectives or adverbs, respectively. For example the senses of the Bulgarian demonstrative pronoun *takav* correspond to the synsets:

Synset: *{such: 2; such that: 1}*

Definition: *of a degree or quality specified (by the `that' clause)*

Sunset: *{such: 1; such as:1}*

Definition: *of a kind specified or understood*

Synset: *{such:3; so much:1}*

Definition: *of so extreme a degree or extent*

6.2.5. Proper names

Different types of proper nouns denoting unique entities are encountered in the corpus – person names, geographical names, names of institutions, companies, etc. Certain proper nouns, including anthroponyms signifying famous persons, are encoded in the English WordNet - for example:

Synset: *{Ploviv:1}*

Definition: *the second sized town in Bulgaria*

Regional and Bulgarian proper nouns of historical or social or political significance in case they are not

included in PWN are encoded either as Balkan-specific concepts (BILI) or as Bulgarian-specific concepts (BUL) - for example:

Synset: *{Ivan Vazov:1; Iv. Vazov:1; Vazov:1; Ivan Minchov Vazov}*

Definition: *a famous Bulgarian writer, publicist and public figure.*

Otherwise, they are linked to the general term according to their referent e. g. *John* is connected to the synset *{first name:1, given name:1, forename:1}.*

6.2.6. Multi-word expressions

Multi-word expressions are linguistic units consisting of more than one distinct lexeme. They are incorporated in the Bulgarian WordNet in a similar way as single words, their POS having the same value as the head word of the expression. Decisions for the encoding of MWE in BulNet are taken according to the consistency criteria. The statistical data coming from the wordnets shows that the distribution of multiword expressions among natural languages is approximately equivalent and covers one forth of the lexis.

Following in part the existing literature, we adopt the following classification for phrases: free combinations of words, idioms and multi-word expressions. We consider a MWE a sequence of two or more words (including graphical words) that denotes a unique and constant concept. Idioms and idiomatic expressions are a lexicalized word group, whose meaning is not compositionally formed from the meanings of its components. Idioms can be part of the wordnet if they denote a unique concept, not a proposition.

An important class of syntactically-flexible MWEs are the so-called support (or light) verbs such as *do, give, have, take, etc.* which combine with certain nouns to express the same meaning as the corresponding lexical verb. They are either encoded as synonyms of the respective content verbs, for example *{ucahstvam:2, vzemam uchastie:1}*:

Synset: *{participate:1, take part:1}*

Definition: *share in something*

or annotated separately. Our approach is to follow the way the Princeton WordNet handles these expressions while at the same time considering factors such as the productivity (degree of collocativity) of the support verbs and taking into account whether it is the same or different support verbs that participate in the formation of semantically equivalent collocations in English and Bulgarian as a way to ensure correspondence between the senses of the support verbs in the two languages where appropriate.

An interesting case is presented by idioms. Some of them are the result of cultural interaction - the Bulgarian counterparts are loan translations of English expressions, for example *take the bull by the horns, close at hand, etc.* Others are functional equivalents and are therefore encoded in the synset representing the relevant meaning, for example *odera kozhata* and *svalyam rizata ot garba* are entered in the Bulgarian counterpart of the English:

Synset: *{overcharge:1, soak:2, surcharge:2, gazump:2, fleece:1, plume:1, pluck:3, rob:2, hook:2}*

Definition:*rip off; ask an unreasonable price*

Bularian idioms which have no idiomatic equivalents in English are encoded as hyponyms to an entry with roughly the same meaning. For example, the BulNet entry

Synset:*{med mi kape na sartseto:1}*

Definition: *be very delighted*

is encoded as a hyponym of *naslazhdavam se { delight:2, enjoy:5, revel:1}*

In the course of annotation the components of the multi-word expressions are grouped and linked to the corresponding wordnet synsets. The lemmas of the MWEs account for the grammatical features (e. g. adjective noun agreement) of the constituents and need not coincide with the lemmas of the individual words. For example: in *familna istoriya (family history)* the gender and person of the adjective *familna* agree with the feminine noun *istoriya* and is lemmatised both in the corpus and in the wordnet entry in its feminine singular form:

BulSemCor: familna{familna #ENG20-06112790-n 1144167285 5770 1} istoriya{istoriya#0 2000000000 5769 0}

Synset:Literal:*{familna istoriya:1}* Lemma:familna istoriya

The BulNet entries of MWEs reflect the neutral word order of the constituents where variations are possible as with idioms, collocations, etc. These features are handled at the stage of annotation.

6.2.7. Domain relations

With a view to modeling the tagged corpus into the Hidden Markov Model (HMM) WSD framework certain kinds of optimizations had to be implemented. A significant one was the association of adverbs with a semantic domain to which they pertain, following the PWN methodology of association of domain-specific words with the corresponding domain through the relation Category domain. All adverbs were linked to a synset corresponding to their semantic domain (such as time, location, manner, quantity, degree, frequency, etc.). For example the Bulgarian synset *{na zakrito}* corresponding to *{inside:1, indoors:1} 'within a building'* is connected through the relation Category domain to the Bulgarian equivalent of the synset *{location:1} - 'a point or extent in space'*. Further, grammatical peculiarities and syntactic function of certain items such as intensifiers, quantifiers, etc. are accounted through linking these items to the relevant domain synset by means of the relation Usage domain.

7. Evaluation

The evaluation of the Bulgarian Sense Tagged Corpus at this stage is performed manually by a second annotator. Further strategies of evaluation have to be developed in order for the consistency of the annotation to be guaranteed. The evaluation of BulSemCor is performed with respect to both the consistency and completeness of the corpus against the wordnet. The completeness check up has to take into account the following considerations:

There is still a large number of wordnet senses that are not mapped in BulSemCor, thus BulSemCor can be further enlarged with texts that include such words;

We may consider separately single-sense and multiple-sense words (as found in BulNet); this may reflect on the weights given to those categories.

Since the senses encoded in BulNet reflect largely the definition of senses in PWN, we may additionally perform experiments to estimate the number of senses attested in the existing lexicographic works, such as the Bulgarian explanatory dictionaries, that are mapped in BulSemCor.

8. Acquisition of multilingual Sense tagged corpora

The Bulgarian sense tagged corpus underlies an HMM formalism combined with additional operations over the wordnet (for the time being relatively low recall but high precision has been achieved) implemented for word sense disambiguation.

BulSemCor will provide an appropriate WSD foundation for a number of future purposes with a special focus on machine translation which is currently poorly explored for minority languages such as Bulgarian. For this purpose the Bulgarian Sense Tagged Corpus will be translated in English (this is also possible for any other language for which a WordNet is constructed). The resulting English corpus will be lemmatized and sentence aligned with the Bulgarian source corpus. Then, to every lemma from the corpus located in the English WordNet a corresponding identification number can be automatically assigned. This is one possible methodology among others (Bentivogli & Pianta, 2005) for obtaining a parallel sense-tagged corpus for Bulgarian and English (as well as for other language).

It has been noted that the sophistication of the statistical methods used in MT makes use of linguistic information (Hutchins, 1995) at different levels. An indispensable step in the further work is providing a proper basis for enhancement of the system in this direction. This will involve the encoding of different types of metalinguistic information in the corpus as well as the elaboration of approaches towards handling specific classes of words not encoded in dictionaries and units above the word level.

One major class to be considered is that of the named entities. Beside proper names (see section 6.2.5.) named entities subsume also locations (place names), company names, organizations names, etc. This is a heterogeneous group which will require different handling for the purposes of MT - transliteration, translation, etc. The task is even more challenging since named entities may incorporate units that require different type of rendering in another language, e.g. in *Емакар ООД*, the first part (*Емакар*), being the name of the company is transliterated (*Emakar*) while the second part of the name (*ООД*) denotes the type of company and is translated (*Ltd.*).

Another task whose relevance to MT has been acknowledged is "the use of syntactic transformations to bring source structures closer to those of the target language" (Hutchins, 1995). The task actually consists in finding functional equivalents of phrases and constructions and can be used in combination with the example-based approach where models are learned from actual expert translations of the same text.

9. Conclusions

The Bulgarian Sense Tagged Corpus contains 63 440 words, part of them linked to form MWE. Three-fourths of the corpus have been annotated and the results have been employed in the experiments on developing a WSD system. Our immediate goal is to complete the task of the annotation of the presented corpus, as well as to carry on enlarging it with more data. The next selection for annotation from BCB has to take into account not only the frequency, but also the already defined BulNet senses, especially those with more than one sense.

In the longer run, as noted in Section 1, our sense-annotated corpus will be employed as training and test dataset for a bidirectional machine translation system based on HHM.

Along with their immediate applications the MT platforms from and to minority languages will ensure these languages' equality at the international level. The experience gained in the elaboration of BulSemCor will be helpful to any future effort in this field and will further national and international cooperation in the creation of tools and resources for minority languages.

10. References

Banko, M., Brill, E. (2001). *Scaling to Very Very Large Corpora for Natural Language Disambiguation.* In Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics. *ACL*, pp. 26-33.

Bentivogli, L., Pianta, E. (2005). *Exploiting parallel texts in the creation of multilingual semantically annotated resources: the MultiSemCor Corpus.* In Natural Language Engineering, Special Issue on Parallel Texts, Volume 11, Issue 03, September 2005, pp. 247-261.

Budanitsky, A., Hirst, G. (2001). *Semantic Distance in WordNet: An Experimental, Application-oriented Evaluation of Five Measure.* In Proceedings of the Workshop on WordNet and Other Lexical Resources, North American Chapter of the Association for Computational Linguistics. Pittsburgh, pp. 29-34.

Fellbaum, C. (1998). *Towards a representation of idioms in WordNet.* In Proceedings of the Workshop on the Use of WordNet in Natural Language Processing Systems (Coling-ACL 1998). Montreal, pp. 52-57.

Fellbaum et al. (1998). Fellbaum, C., Grabowski, J. and Landes, S. (1998). Performance and confidence in a semantic annotation task. In Fellbaum, C. (ed.), *WordNet: An Electronic Lexical Database.* Cambridge (Mass.): The MIT Press, pp. 217-237.

Hutchins, W. John (1995). Machine translation: a brief history. In E.F.K.Koerner and R.E.Asher (Eds.), *Concise history of the language sciences: from the Sumerians to the cognitivists.* Oxford: Pergamon Press, 1995. pp. 431-445

Koeva et al. (2004). Koeva, S., Tinchev. T., Mihov, S. *Bulgarian WordNet-Structure and Validation.* In Romanian Journal of Information Science and Technology, Volume 7, No. 1-2, 2004, pp. 61-78.

Koeva et al. (2005a). Koeva, S., Rizov, B., Leseva S. *Flexible Framework for Development of Annotated Corpora..*In International Journal Information Theories & Applications, Sofia.. [In press].

Koeva et al. (2005b). Koeva, S., Krstev, C., Obradovic, I., Vitas, D. *Resources for Processing Bulgarian and Serbian.* In Proceedings from the International Workshop Language and Speech Infrastructure for Information Access in the Balkan Countries. Borovets, 2005, pp. 31-39.

Stamou et al. (2002). Stamou S., Oflazer, K., Pala, K., Christoudoulakis, D., Cristea, D., Tufis, D., Koeva, S., Totkov, G., Dutoit, D., Grigoriadou, M. *BALKANET: A Multilingual Semantic Network for the Balkan Languages.* In Proceedings of the International Wordnet Conference, Mysore, India, 21-25 January 2002, pp. 12-14.

Tufis et al. (2004). Tufis, D., Cristea, D., Stamou, S. *BalkaNet: Aims, Methods, Results and Perspectives. A General Overview.* In Romanian Journal of Information Science and Technology, Volume 7, No. 1-2, 2004, pp. 1-32.