COGNITIVE STUDIES | ÉTUDES COGNITIVES, 12
SOW Publishing House, Warsaw 2012

SVETLA KOEVA
Department of Computational Linguistics, Institute of Bulgarian Language,
Bulgarian Academy of Sciences
svetla@dcl.bas.bg

BULGARIAN SENSE-ANNOTATED CORPUS — BETWEEN THE TRADITION AND NOVELTY

Abstract

The Bulgarian Sense-annotated Corpus (BulSemCor) is compiled according to the general methodology established by the SemCor project. It is a subset of the Brown Corpus of Bulgarian semantically annotated with a corresponding synonym set (synset) in the Bulgarian wordnet. Unlike the bulk of sense-annotated corpora where only (sets of) content words are annotated, in BulSemCor each lexical unit has been assigned a sense. The main contributions achieved in the work on BulSemCor are briefly decides in the presented paper: definition of an annotation schema, compilation of an input corpus, development of a sense-annotated corpus, Bulgarian wordnet enlargement.
Keywords: Key words: corpus studies, corpus annotation, annotation principles.

I. Introduction

The Bulgarian Sense-annotated Corpus (BulSemCor) (Koeva 2010b) is compiled according to the general methodology established by the SemCor project (Landes et al. 1998). It is a subset of the Brown Corpus of Bulgarian semantically annotated with a corresponding synonym set (synset) in the Bulgarian wordnet (BulNet) (Koeva 2010a). Unlike the bulk of sense-annotated corpora where only (sets of) content words are annotated (Ng and Lee 1996; Pianta, and Bentivogli 2003; Wu et al. 2006, to mention but a few), in BulSemCor each lexical unit has been assigned a sense. Lexical unit is considered a word (single word, designating a unique and constant concept), as well as a multiword expression (two or more words, not necessarily contiguous, designating a unique and constant concept, in a relationship of equivalents with a single word from the same language or other language) (Koeva 2006).

It is a well known fact, that the most important difficulty for language processing result from its ambiguity and indeterminacy at many levels. The problem becomes even more difficult when such well-known phenomena as homonymy and polysemy become involved. For example, in the collocation *голям град* (a big city

or a heavy hail) it is not clear whether the word *град* means a "large settlement" or "hail", but the homonyms are clearly distinguishable in the sentences *Построиха голям град* (They have built a big city) and *Падна голям град* (There was a big hailstorm). The process of (automatic) disambiguation can be described in most general terms as the choice in a given context of the most accurate meaning (in the wide sense) for a given language expression from the set of acceptable meanings. (Automatic) word sense disambiguation, correspondingly, is an association of a given lexical unit from the text (a single word or a multiword expression) with a sense, distinguishable from amongst a set of senses which are potentially connected with this word. Such definition of word sense disambiguation raises a number of questions: 1) what is the word sense and whether the meaning can be defined outside the context; 2) what are the senses of a given word (and how are they formulated) and 3) what is the correct sense associated with the given word in a given context. The first question is not only linguistic — we accept that the sense of a given lexical unit is expressed by means of the unique set of semantic relations with other words in the language. The problem with defining the appropriate set of senses associated with a given word can be interpreted in the following way — if necessary, new senses are added in the repository resource without the ambition for completeness. Consequently the task which needs to be resolved in the Bulgarian Sense-annotated Corpus is what is the most appropriate sense among the available set of word senses that can be associated with the particular lexical unit in a given context.

Section two presents a brief overview of the similar projects. Next, the compilation of the input corpus is described, followed by a discussion for the basic approaches in meaning representation. The selection of Bulgarian wordnet as a word senses repository is grounded. Section six discusses the principles of annotation accepted, along with the some linguistic assumptions behind the annotation process. At the end the achieved results are sketched as well as the application of the corpus for word sense disambiguation.

II. State of the art

Current practice accepts two basic methods for the creation of sense-annotated corpora — via selection from a set of meanings which are offered in a given language repository — interpretative dictionary, thesaurus, wordnet, or via translation of sense-annotated corpus from one language into another.

In the creation of BulSemCor the primary methodology applied is the one proposed in in the development of English Sense-annotated Corpus (SemCor), created in Princeton University (Landes et al. 1998). The term semantic concordance introduced as a theoretical background refers to the association between corpus and lexicon, in such a way that each word in the text is connected with a suitable sense from the lexicon (Miller et al. 1993: 303–308; Miller 1995: 92–94). SemCor includes part of the Brown Corpus and one additional text — a novella (Landes et at. 1998) which consists of 352 texts and contains about 700 000 words. The words belonging to four parts of speech — nouns, adjectives, verbs and adverbs, are manually annotated with senses from the Princeton wordnet with the help of a purpose built programme (the initial aim of the creation of the annotated corpus

was to support or reject the representation of senses in wordnet). In 186 texts (359 732 words) the words from open classes (192 639) are tagged with the part of speech, lemmatised and sense-annotated with wordnet senses. In the remaining 166 texts (316 814 words) only the verbs are annotated with the relevant lemma and sense (41 497 verb encounters). The following general procedure is used: if the word has only one sense — the sense is verified; if the word has more than one sense — the task is to choose the correct one (the same procedure is used in the annotation of the Bulgarian Sense-annotated Corpus).

Similar methodology is applied in a variety of projects for semantic annotation. For example the most frequently used ambiguous nouns and verbs in English (121 nouns and 70 verbs) are annotated with their correct wordnet senses in 192 800 sentences extracted from the Brown Corpus and the Wall Street Journal (Ng and Lee 1996:44).

In corpora of about 100 000 words for Spanish and Catalan and about 50 000 words from Basque three levels of annotation are offered — syntactic, semantic and pragmatic (Agirre et at. 2006; Navarro et at. 2003). At a semantic level the annotation is a reference to the wordnet synonym sets for the respective languages, wherein only nouns, verbs, adjectives and adverbs are annotated. A particular feature is that it accepts association with more than one synonym set, if such an interpretation is allowed by the context.

For Italian (Montemagni et al. 2000) in a corpus of about 300 000 words (together with syntactic annotation) the nouns, verbs and adjectives (and the respective multiword expressions) are annotated with the Italian wordnet word senses. A particular feature is that it includes additional lexical-semantic information, for example in cases of non-literal use, and also information about the annotation agreement, for example notes about problematic cases.

For the Swedish corpus of about 1 000 000 words designed along the model of the Brown Corpus more than 250 000 words are annotated with senses from the Gothenburg Lexical Satabase (Järborg et al. 2002: 1494). The initial aim — to verify the hypothesis whether it is possible to describe the lexical senses of words in a way that predicts their senses in actual linguistic usage — developed into the creation of a sufficiently large sense-annotated corpus suitable for natural language processing. The particular feature is that the source of the semantic annotation is not wordnet.

For Chinese, the words from a corpus (Wu et at. 2006) are cross-referenced with senses which are defined specially for the aims of the annotation — the words are linked with the relevant synonym sets from a lexical-semantic network similar to wordnet. As of the moment of writing the article, 813 nouns and 132 verbs had been included in the specially constructed repository and 60 895 encounters in texts from the People's Daily have been semantically annotated. The particular feature is that the semantic annotation is done simultaneously with the creation of the lexical-semantic network — resource for annotation.

The SENSEVAL (Evaluation Exercises for the Semantic Analysis of Text) scientific initiative, conducted periodically, has a collection (mainly for English) of sense-annotated corpora (linked to various versions of wordnet or interpretative dictionaries — less frequently). Those collection is used for the training and evalu-

ation of the word sense disambiguation systems (Kilgarriff and Rosenzweig 2000:18). A considerable effort is made to use the existing wordnets as sense inventories. The particular feature is that in some of annotated corpora only certain words are annotated.

Using the second possible method — via translation — the parallel (English-Italian) semantic corpus MultiSemCor was developed (Pianta and Bentivogli 2003; Bentivogli and Pianta 2005). The hypothesis is that in translation from one language to another the semantic information is preserved to a large extent. Consequently, semantic annotations can be transferred from the source text to the translated un-annotated equivalent by means of an automatic word to word alignment. Such an approach is used in the creation of the sense-annotated corpus for Romanian (Lupu et at. 2005). The main difficulty (and hence disadvantage) in the transfer method is the automatic word alignment in different languages. It is known that there is no strict correspondence in different languages between single words and multiword (phraseological) expressions, between morphological and syntactic constructions, or between levels of lexicalisation, and another matter entirely is the subjective influence of the translator in the selection of the lexical, grammatical and stylistic (pragmatic) means.

In conclusion, the known sense-annotated corpora in different languages can be summed up with the following features:

- The source corpora are small in volume, in many cases this is the Brown Corpus or corpora composed on its model, in other cases these are periodical printed publications.
- The semantic annotation is carried out manually by a person, in the majority of cases experts, who predetermine both the expected high quality, but this also results in the small volume of annotated corpora and the duration of the process (of course, in some cases different techniques are used to optimise the annotation).
- The predominant resource for sense-annotation is wordnet, but there are examples when the annotation is done in relation to another dictionary or thesaurus. Wordnet is preferred in comparison with the other possible sources for sense-annotation because of its structure, size and accessibility.
- One of the main features in the known examples of sense-annotated corpora is that only verbs, nouns, adjectives and adverbs are annotated and in some cases only given words (frequency defined) in these classes are selected.
- In the majority of cases sense-annotation is combined with another type of annotation — usually morphologic (lemmatisation) and morpho-syntactic (part of speech), but in certain cases with syntactic or discourse annotation.
- Usually in the case of sense-annotation only one appropriate sense is chosen, but in certain approaches more than one is permitted, if the senses are defined in an excessively granular manner and are suitable for interpretation in the context, as well as additional notes for the level of annotator's confidence and agreement are provided.

III. The input corpus for sense-annotation

The Bulgarian Sense-annotated Corpus offers a semantic annotation in a representative sample extracted from the Bulgarian Brown corpus[1]. The Bulgarian Brown corpus was created in accordance with the methodology of the well-known Brown Corpus which constitutes a well balanced model for the synchronous condition of a given language. To the best of our knowledge although some attempts to develop a methodology for corpus balance and representativeness have been made (Brysbaert and New 2009), no commonly accepted methodology exists. For this reason the Brown Corpus is more suitable for sense-annotation than a spontaneous collection of texts — for example a series of editions of a given newspaper. The source corpus for semantic annotation preserves the original structure of the Bulgarian Brown corpus and contains an excerpt of approximately 100 words for each of the 500 texts (the samples are expanded to the left and right to the end of the sentence and for this reason the number of words in them is not exactly 100). The source corpus consists of two parts.

The first subset is selected according to the density of highest frequency open-class lemmas with heuristics applied to provide a balance between different parts of speech and a better coverage of lemmas.The aim is to find out those samples that contain the maximum number of the most frequently used open-class lemmas calculated in a very large corpus — the Bulgarian national corpus[2]. The selection criteria for the second subset of the source corpus are: the maximum number of words in the samples should not be encountered in BulNet, and the maximum number of words in the samples which are met in BulNet should appear in one or more synonym set. The aim is to create an input corpus which on the one hand contains words whose senses have not been annotated in the first subset, and on the other — to contain a maximum number of ambiguous words to illustrate the contextual choice of differing senses.

As a result, the source corpus consists of 811 samples containing 101791 tokens, which on their part constitute 95 119 lexical units, of which 89 341 single words and 5778 multiword expressions. The source corpus is subjected to pre-processing — normalisation (for example, the removal of certain typing errors) and lemmatisation (automatic cross-referencing with the lemma) while the link between the primary condition and the variables remains restorable and can serve for the creation of rules for automatic normalisation.

Although the majority of the known sense-annotated corpora use the Brown corpus or corpora created in accordance with its methodology, to the best of our knowledge they do not use heuristics to determine what parts of the corpora should be annotated — on the contrary (representatives of) open-class words are annotated sequentially in the texts.

IV. Word sense repositories

The (automatic) disambiguation consists in the association of a given lexical unit with a most appropriate sense from amongst the set of appropriate senses which are associated with it. This assumes the restriction that the meaning of a collocation

[1] http://dcl.bas.bg/Corpus/home_en.html
[2] http://search.dcl.bas.bg

or sentence is based only on one of the senses of a given ambiguous word and that there is no interpretation of the meaning dependent on extra-linguistic factors. This brings as well the requirement for a repository in which all the possible senses of a given lexical unit are described completely and consistently. Of course, neither the restriction is valid (for the definition of the correct meaning both context and extra-linguistic factors are important), nor is the requirement satisfactory (there is no fully complete or non-contradictory language resource with a description of the all possible senses), and thus the task is limited to making the most appropriate choice from amongst the existing possibilities (the choice of language repository for word senses to a certain extent predetermines the results).

For the definition of the word senses a number of basic approaches can be identified: traditional interpretative dictionaries, theories based on primitives and relationist approach.

IV.1. Word senses in traditional dictionaries

Traditional interpretative dictionaries present meaning through definitions which interpret the uniquely defined concept. In interpretative dictionaries it is accepted that the word senses can be defined by means of (organised) sets of senses of a given word which should not intersect. Definitions in an ideal case consists of generalisation (the class to which the lexical unit belongs) and differentiation (what properties distinguish this word from the remaining words in the class) — of course the definitions of different parts of speech should have different canonic structure. Different dictionaries include different sets of meanings — a given meaning in one dictionary might be divided into a number of meanings, unified with another meanings or even entirely absent in another dictionary. The set of meanings depends on the designation of the dictionary, the chosen word list (perhaps based on the frequency of the encounters or "importance") and on the accepted level of granularity (Fellbaum et al. 2001). In many cases the decision for a more general or more detailed representation of a given meaning is subjectively motivated (Kilgarriff 1997: 100).

The use of traditional dictionaries for the word sense disambiguation is made difficult by the interpretation of multiword expressions (mainly in the case of nouns) and alternations (mainly in the case of verbs). The presentation of multiword expressions (for example *кръвна банка* (blood bank), *зърнена банка* (grain bank), *банка за стволови клетки* (stem cell bank) can be appended to the meaning of the basic word, in this case "bank as a financial institution" (Hanks 2000: 207), or cross referenced to a more general definition which relates to a group of multiword expressions, for example, "each of the institutions for the storage and conservation of different types of objects" or multiword expressions could be viewed as individual lexical units with their own meanings. Although the most logical approach to multiword expressions is to interpret them as individual lexical units, the practice in traditional dictionaries is to ignore them or to list them in the dictionary entry under the head word (in some cases under the one of the dependent words chosen by some unclear criteria) without interpretation.

Alternations is also included albeit partially and illogically in interpretative dictionaries (similar to systemic derivations) (Koeva 2008). A possible explanation for this is that the alternations can be predicted if the source structure complies

with certain requirements. The systematic nature of alternations, however, does not cancel the fact that diathesis, as the strongest type of alternations, constitutes different meanings which have to be described in a suitable manner.

There are two main trends in lexicography — to differentiate a wider range of meanings, in order to encompass all cases of use, or to search for a unifying and more general meaning for all uses. In the first case it is possible that there is no clear difference between the set of granulated meanings which might overlap to a certain extent, while on the other hand the generalised meanings are usually not sufficiently specific. Definitions per se, even if they comply with the familiar models of generalisation (hypernym) to the concretisation, or with the accepted formal models for representation, do not constitute a sufficiently non-contradictory description, suitable for automatic analysis. Consequently the division of meanings, in the way in which they are traditionally presented in interpretative dictionaries, is not suitable for the natural language processing (Kilgarriff 1997:108). To generalise, dictionaries are usually created to refer to certain senses and their purpose is not semantic annotation — in many cases the dictionary senses do not comply with the senses of words in their real use (Fellbaum et al. 2001).

IV.2. Primitive-based theories for word senses

Primitive-based theories accept that meaning can be defined with a fixed set of primitive elements (Katz 1972), wherein the differences are basically reduced to the set of primitives to which the meanings can be decomposed. In the most general terms semantic primitives can be seen as elements which are used to explain concepts, but for their parts cannot be interpreted with the help of other elements. Thus the aim is to define words by means of a restricted language for representation. The positive sides of this are the systematic presentation of definitions in (possibly) a non-contradictory way and the avoidance of circularity, while the limitations of the approach can be generalised as a lack of a single opinion for the primitives (semantic language) and the impossibility to represent certain semantic differences (Jackendoff 1990, Wierzbicka 1996).

A similar approach is one in which words are defined by means of probabilistic and prototypical semantic components — individual, combinable and viable units which form one or another meaning (Hanks 2000). It is accepted that outside the context of the event in which the meaning occurs, words can be characterised in terms of potential meanings, rather than actual meanings (Hanks 2000: 211). The potential meaning of a given word is seen as a set of probabilistic semantic components and in a particular context only some of them are activated. In each case of the use of the word at least one of these components is manifested, more frequently a combination of them — the probabilistic component approach allows for both ambiguity but also indeterminacy. From the position of word sense disambiguation this can be seen as a competition between different components within the set of possibles — the different usages activate different combinations of the components.

With regard to the definition of the meaning of the word, this is probably an appropriate approach, which when combined with the organisation of words in lexical-semantic networks such as wordnet, would constitute a productive model for natural language processing, including the case of (automatic) word sense dis-

ambiguation. As of the present moment even English does not have a practically functional lexicon, in which the potential meanings of words are defined as sets of prototype probabilistic semantic components. There is a further problem not only with the formulation of prototype components but also with procedures for cross-referencing a given word form with a set of components which are valid for it within a given context.

To a certain extent and from a conceptual point of view the so-called generative lexicon can be ascribed to the prototype probability model. In this approach the semantically connected meanings are not listed but generated by means of rules which describe the dependencies in the formation of meaning within the complex of representational levels (Pustejovsky 1995: 410–413).

IV.3. Word senses in relationist theories

According to relationist theories, meaning is presented by means of explicitly expressed relations between words (and relevant concepts) (Fodor 1975), categorised by means of a set of properties.

The most popular application of the concept that meaning is presented by means of explicitly expressed relations between words (denoting relevant concepts) is wordnet. The Princeton wordnet (Miller 1990; Fellbaum 1998) constitutes a lexical-semantic network, whose nodes are synonym sets (called synsets) which contain words or multiword expressions (called literals) and whose arcs express semantic, morpho-semantic, derivational and extralinguistic relations between the objects located within the nodes. The sense of literals, and the meaning of synonym sets are seen as an instance of linguistically independent concepts. In semantic annotation the sense of a given word is taken to mean the meaning of the respective synonym set. The meaning of a synset in wordnet is expressed by means of the relations to other synsets in the network, on the one hand, and through the properties of the node itself (implicitly through the synonym relation between the literals in the synonym set and explicitly through the gloss, examples of usage and notes to the literals and synsets) on the other hand. The Princeton wordnet includes two types of relations: between literals (called lexical) and between synonym sets (called semantic). Semantic relations refer to all literals in both synonym sets which they connect. Lexical relations connect literals within the framework of one (less frequently) or two synonym sets. The opposition between literal relation and relation between synonym sets is more suitable, since literal relations can also be semantic. Relations between synonym sets are semantic, morpho-semantic and extra-linguistic, while relations between literals are semantic and derivational. The semantic, morpho-semantic, derivational and extra-linguistic relations in wordnet express real relations (between sets of objects or abstract essences) in the real world.

Synonym sets in the Princeton wordnet include literals belonging to four parts of speech: noun, adjective, verb and adverb. The so-called closed classes are partially included (cardinal numbers are classed as adjectives or nouns — less frequently, some pronouns are also included as nouns, adjectives or adverbs).

Wordnet is one of the most complete lexical resources (for the sake of comparison — literals in the Bulgarian wordnet are much greater in number than the wordlist of a standard orthographic dictionary) and synonymous sets from different languages

are connected by means of interlingual relation of equivalence, through which the multilingual lexical-semantic network (the global wordnet) is formed.

Wordnet combines the qualities of the existing language resources: it contains glosses and examples, like typical dictionaries, but also organises synonym sets into a conceptual network by means of the semantic relations which exist between them. It is, therefore, not surprising that the majority of sense-annotated corpora and most systems for the automatic word sense disambiguation use wordnet as a semantic repository. In defining the correct semantic annotation the affiliation of a given lexical unit to a given synonym set in the lexical-semantic network is taken into account, i.e. not only the defined interpretative meaning but all the semantic, morpho-semantic, derivational and extralinguistic relations it belongs to, examples of usage, notes pertaining to the literal or synonym set, in other words — the entire place of the lexical unit in relation to the other lexical units (and concepts) in the language.

In the structure of wordnet words and multiword expressions are included on an equal basis, i.e. it overcomes one of the disadvantages of traditional dictionaries where multiword expressions are not described consistently. The same can also be said to a certain extent about alternations — wordnet is structure in which alternations have their place (Kohl et al. 1998) although as of the present moment they are represented only in a limited and in certain cases contradictory form — only those alternations which form a new lexical unit with unique meaning (diatheses) can be included in wordnet as a separate synonym set.

The structure of wordnet allows for the addition of new relations of a different type, both between synonym sets and between literals. It is this relational structure which is the main advantage of wordnet. As a result of this, the high level of granularity of wordnet not only can be overcome, but can also be used by various approaches for the automatic word sense disambiguation (Ide and Veronis 1999) by means of ascribing weight according to the type of relations — hypernymy, synonymy etc., and the calculating of the length of the path and the number of relations of one type to a given node (the shortest path is looked for between the nodes for the measurement of semantic proximity — the hypothesis is that for a given set of lexical units, which are adjacent in the text, the meanings selected have the maximum possible semantic proximity); by means of measuring the specific nature of the concepts, organised in a hypernymy relation (the more specific a concept is, the more its hyponyms are semantically connected); by means of the formulation of semantic and probability rules in the aims of uniting synonym sets close in meaning in the aims of reducing ambiguity (Mihalcea and Moldovan 2001) etc.

In conclusion, on the one hand, the existing traditional dictionaries do not provide an adequate description of the meaning of the words from the point of view of (automatic) word sense disambiguation. On the other hand, despite the new theories — generative lexicon, prototype probabilistic components — there is no complete and consistent model for the description of lexical meaning, and to even lesser extent a linguistic resource constituted in such a model which might contain a sufficient quantity of lexical units whose lexical meaning is described in an adequate and consistent way. Thus the Bulgarian wordnet is chosen as a source of senses for the annotation of the Bulgarian Sense-annotated Corpus because of its

comprehensiveness, its size, and its entirety and relational structure which allows for various types of generalisation and grouping.

V. Bulgarian wordnet — repository for sense-annotation

The semantic annotation in the Bulgarian Sense-annotated Corpus provided a link to the synonym set of BulNet, including the relevant literal (word or multiword expression), consequently to the definition (including examples and notes to the synset and literal), semantic, morpho-semantic and extra-linguistic relations of the synonym set and semantic and derivational relations of the literal. In other words the semantic annotation is not only annotation to the sense defined in wordnet, but to the entire wordnet structure the relevant literal belongs to.

The definitions in BulNet adhere to the classical structure of generalisation to differentiation which is reflected in their formal structure — different for different parts of speech. The definitions are not borrowed — they are compiled in such a way as to correspond to the English translational equivalent, — and take into account the existing interpretative definitions of Bulgarian, but at the same time they are unique at the moment of their compilation. The synonym set can be further expanded by adding information to show usage — one or more examples some of the taken from the Bulgarian national corpus, some of them — from the internet, and some of them — translated from English or constructed by the experts (illustrative examples for each individual literal is not obligatory), one or more notes about stylistic, morphological or syntactic indications of the synonym set or literal (some of the notes, for example, about verbal aspect are obligatory).

BulNet is expandable in two ways. The first way is connected with the definition of the annotation schema and the accepted principle that each word in the text will receive semantic annotation — this requires the addition in the structure of BulNet of synonym sets for six additional parts of speech (since this is connected with closed classes, this means the addition and classification of all lexical units of the closed classes). The definition of the meanings of the words in closed classes is based on an analysis of the syntactic evidence and the semantic indications which are observed in the corpus, as well as the existing lexicographic and grammatical descriptions. The existing classifications of closed classes, due to the high level of polysemy in the majority of them, are in many cases not sufficiently precise and not based on clear criteria. For some of the words in closed classes there are synonym sets in English and these are taken into account. To the best of our knowledge, this is the only wordnet to cover the function word classes. The following classes have been included — pronouns, prepositions, coordinating and subordinating conjunctions, particles, interjections. Most of the newly-encoded synsets are provided with English translation equivalents in the synset note. Function words are integrated into the wordnet structure through the [category_domain] relation pointing to the synset denoting the relevant category: {preposition}, {coordinating conjunction}, {subordinating conjunction}, {particle}. For pronouns this is the pronoun type {personal pronoun}, {possessive pronoun}, etc.

BulNet is also expandable with cultural specific concepts for which a position in the structure is found and also with language specific concepts — comparative adjectives, diminutive forms etc.

The second type of expansion of BulNet is connected with the need to add new senses, if they are missing, or if those already existing have changed. In the process of annotation monosemic and polysemic words are checked and if there is a suitable synonym set, then it is selected, if there is a synonym set which would be more suitable in the relevant modification and this modification corresponds to the entire structure of wordnet, the change is made, if there is no suitable synonym set, a new one is created.

VI. Annotation schema

There are no unified standards for annotation practice but there are many recommendations and criteria in relation to which the processing of a given annotation schema could be approved. A number of documents contain proscribed recommendations for certain levels of annotation, describing the best practices for what needs to be annotated, the extent to which it needs to annotated and so on (EAGLES 1996a; EAGLES 1996b). The following general recommendations for good practice in textual annotation were adhered to in the creation of the Bulgarian Sense-annotated Corpus (Leach 1997; McEnery and Wilson 2010):

- The original text can be reconstructed easily without the annotation which has been added.
- The annotations can be extracted easily from the annotated text.
- Each annotated text should be accompanied with the suitable documentation including the annotation schema.

The following general recommendations were also taken into account in the development of the annotation schema for semantic annotation of the Bulgarian Sense-annotated Corpus (Wilson and Thomas 1997: 55–57):

- The semantic annotation must be related to the meaning linguistically and cognitively (the Bulgarian Sense-annotated Corpus is annotated with BulNet senses — where the words are organised in synonym sets, interconnected with a range of semantic relations, in such a way as to represent the conceptual structure of the language).
- The annotation must encompass a large part of the words in the corpus and not only some (in the BulSemCor all words are annotated).
- The schema must be flexible in such a way as to allow the description of different details — periods in the development of the language, registers etc. (annotation to a synonym set from BulNet refers also to all literal and synset notes which express different limitations of use).
- A certain level of granularity of meanings must function and the choice of the most suitable level of granularity should remain open (transitivity of hypernymy allows for transitions to a higher and lower level, and according to a more generalised or more granulated sense representation).

- The semantic annotation schema needs to have a hierarchical structure based on augmenting generalisation of relations between senses (the basic wordnet relation is taxonomic — hypernymy relation).
- The annotation must correspond to standards, if such exist (to the moment the preliminary standards for semantic annotations (EAGLES 1999) are known and followed giving the priority of the best practice shown by the SemCor).

The EAGLES (1999) preliminary standards for semantic annotation include certain general criteria which must be taken into account when the annotation system is being developed. They are:

- Adequate coverage. All linguistic properties which are the aim of the annotation have to be reflected in it. The aim of the semantic annotation in BulSemCor is to link the unique sense which is manifested in the usage of a given lexical unit to the most appropriate BulNet synset. Each lexical unit in the corpus receives semantic annotation not limited to the gloss but to the entire synonym set and its properties. The association of a given lexical unit with more than one synonym set is not permitted. Complete morpho-syntactic and syntactic annotation is not envisaged since it is not relevant for the aims of semantic annotation. Single words and multiword expressions are interpreted in an uniform way, both in the sense-annotated corpus and in BulNet.
- Consistency: The annotation schema must be organised around consistent principles which determine what types of objects are tags, what type of objects are attributes and what type of objects are values. In the Bulgarian Sense-annotated Corpus tags are lexical units. Correspondingly attributes are: word form (value is the used form in the corpus), lemma (value is the lemma ascribed to a word or multiword expression), sense (value is the synset identification number from BulNet).
- Restorability: The annotation schema must allow for the restorability of the initial text from its annotated version. This condition is fulfilled by the Bulgarian Sense-annotated Corpus — not only is the text without annotation restorable, but also the text before normalisation.
- Verifiability: The verification of annotation has to be possible, seen as a process which automatically verifies whether the marking of a document follows the accepted standards. Consistency checks are inherent to the functionality of the annotation programme Chooser used for the creation of the Bulgarian Sense-annotated Corpus (Koeva et al. 2008).
- The possibility for extracting information from the annotations: The annotation schema must allow different levels of analysis. The association with the synonym set in the BulNet structure refers not only to the synset content but also to all levels of the hierarchical and non-hierarchical relations in the wordnet structure.

- The ability to process: The presentation of annotation must correspond to the requirements of text processing. The corpus is represented in an xml format.

- Expandability: The structure of the annotation schema must allow for the possibility of expansion. Minimum restrictions are imposed on the extension of the specified file format, so that it permits addition of flat and/or hierarchical annotation schemata without affecting the current one, thus enabling other levels of annotation.

- Compactness: This requirement is connected with the limitation of the number of symbols which are introduced into the text for annotation. With the technology development this requirement to a large degree has become irrelevant, but has still been adhered to since the entire information associated with the lexical unit is reduced to the identification number.

- Readability: The annotated text must be understandable (which means either that it is not encoded or that it can easily be transformed into raw text). This requirement has also been adhered to by means of the second option provided.

The process of working on the Bulgarian Sense-annotated Corpus can be described thus: the formulation of an annotation schema, the development of a programme to assist the annotation, compilation of a source corpus and preprocessing, the development of an initial convention for annotation, and sense-annotation of the corpus.

The first stage in the annotation process is automatic lemmatization followed by manual post-editing. Through lemmatization a two-fold purpose is accomplished: association of word forms with a canonical form is ensured (i.e. morphological annotation); each lexical unit in BulSemCor is mapped to all the synsets in BulNet that feature a literal (synset member) with an identical lemma.

The annotation process involves two major tasks: defining the boundaries of the lexical units in the corpus; and choosing the most appropriate sense for a lexical unit from a list of candidates. Multiword expressions pose a number of challenges with respect to: (i) delimitation — the boundaries of a multiword expression are not necessarily straightforward, consider contextual ellipsis, reduced and expanded variants, etc.; (ii) lemma definition — the form of the individual elements in the multiword expressions may differ from their canonical forms as single words; (iii) morpho-syntactic value — the part of speech of a multiword expression may not be the same as the part pf speech of the head; (iv) semantic value — the concept expressed by a multiword expression may not be the sum total of the concepts expressed by its elements.

The appropriateness of candidate senses is assessed according to: interchangeability of the lexical unit in the corpus with the rest of the synonyms in the synset; appropriateness of the definition; and the position of the synset in the wordnet structure. In cases where no appropriate candidate is found among the list of BulNet synsets one checks whether a synset denoting a relevant sense exists in

Princeton wordnet, and, if so, it is encoded in BulNet; otherwise, a new BulNet-unique synset is created, as with most closed-class words and language and culture specific words.

VII. Levels of annotation in the Bulgarian Sense-annotated Corpus

Different levels of linguistic annotation can be distinguished (Leach 1997: 8–15), for example: morphologic, morpho-syntactic, syntactic, semantic and discourse (EAGLES 1996b:3), wherein annotated corpora are associated with more than one level (as in the case with the Sense-annotated corpus of Bulgarian). The accurate definition of the word sense depends on the lower levels of annotation, for example, the restriction of the possible senses of the word *син* (either blue, an adjective, or son, a noun) depend on the definition of the lemma, according as an adjective or noun. Certain words are also encountered with certain grammatical characteristics only as part of a multiword expression or with a different sense (for example: *английска сол* (magnesium sulphate), *минерални соли* (sodium chloride), *солна киселина* (Hydrogen chloride), *солена вода* (salt water), etc.).

In the Bulgarian Sense-annotated Corpus by means of automatic lemmatisation each lexical unit (notwithstanding the form in which it is used) is associated with the suitable synonym sets from BulNet, where the literals are also cross-referenced to their lemma, for example <LITERAL>*продумвам*<LEMMA>*продумвам*</LEM MA></LITERAL> (to utter). The BulSemCor also comprises morpho-syntactic annotation — each lexical unit is assigned the part of speech tag of the synset it is annotated with, for example ENG20-01543395-n. Partial information for the dependencies between compounds' elements is available through syntactic head marking of compounds in BulNet, for example the analogous information in BulNet <SY NONYM><LITERAL> *министър* /*на*/ /*правосъдието*/<SENSE>1</SENSE> <LEMMA>*министър на правосъдието*</LEMMA></LITERAL><LITERAL> /*правосъден*/ *министър*<SENSE>1</SENSE><LEMMA>правосъден мини-стър</ LEMMA></LITERAL></SYNONYM> (minister of justice). In semantic annotation two types can be distinguished (McEnery and Wilson 2001: 61–62) — representation of semantic relations between the words in the text (for example annotation of the semantic roles of arguments, the valence of lexical units, realisation of semantic frames) or the semantic properties of words (for example annotation of the meaning of words). Semantic annotation proper consists in the association of a lexical unit in BulSemCor with the most appropriate synset in BulNet. The annotated item is associated with all the linguistic information in the synset, including synonyms, gloss, part of speech value, the semantic, morpho-semantic and extralinguistic relations pertaining to the synset, the semantic and derivational relations pertaining to the literal, etc. The BulNet synsets are associated with their equivalents in Princeton wordnet through unique identifiers. In such a way, the annotated lexical units are mapped to their translation equivalents in English and, with Princeton wordnet serving as a hub, to all the other wordnets.

VIII. Annotation methods and agreement

Two methods of annotation can be distinguished (Kilgarriff 1998): linear (textual) method in which the words are annotated sequentially without simultaneously ex-

amining lexical units with identical senses or a set of senses with which the given lexical unit is associated. The second method is lexical (intersecting) in which all the encounters of a sense of a given lexical unit in the corpus are annotated, then progressing to the next lexical unit etc. — all senses of a given word are analysed only once. In the annotation of the BulSemCor the two basic methods are applied simultaneously in such a way as to use their advantages and to neutralise the disadvantages.

The subjectivity of decision taking is closely related to agreement between the annotators about the validity of a given annotation. Therefore, in SemCor the annotators note the level of confidence with which a given sense is chosen (Fellbaum et at. 1998). The work on the Bulgarian Sense-annotated Corpus is validated by a second, sometimes third annotator who checks decisions about the annotations made. The degrees of annotation agreement are not marked so far.

IX. Conclusion

The main results achieved in the work on BulSemCor include: definition of an annotation schema, implementation of an annotation tool, elaboration of a methodology and annotation conventions, compilation of an input corpus, development of a sense-annotated corpus, BulNet enlargement (for an overview cf. Koeva 2010b: 7–42).

Some of the quantitative parameters of BulSemCor are presented below — the overall number of tokens and annotated lexical units, along with their distribution into single words and multiwor expressions (MWEs) (Table 1), as well as the distribution of annotated lexical units according to part of speech (POS) (Table 2):

Table 1. Overall numbers of tokens and annotated units

Total number of tokens	Annotated words	Single unique tokens	Annotated single words	Annotated MWEs	Unique tokens in MWEs
101062	99480	88196	86842	5797	12866

Table 2. POS distribution of annotated LUs

POS	Nouns	Verbs	Adj	Adv	Preps	Conj	Pron	Part	Interj
Number	31058	17041	12012	7935	14772	7265	6810	2570	17

The basic function of the Bulgarian Sense-annotated Corpus (and the reason for its creation) is to serve as a training and test corpus for an automatic word sense disambiguation (Fellbaum et al 2001) which is applicable in many areas of the natural language processing.

Last but not least the Bulgarian Sense-annotated Corpus provides information for (linguistic) research, since the sense of a given lexical unit is explicitly connected with its usage.

References

Agirre et al. 2006: Agirre E., Aldezabal I., Etxeberria J., Izagirre E., Mendizabal K., Quintian. M., and Pociello E. A methodology for the joint development of the Basque wordnet and Semcor. In: Proceedings of the 5th International Conference on Language Resources and Evaluations (LREC), Genoa (Italia).

Bentivogli and Pianta 2005: Luisa Bentivogli and Emanuele Pianta 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, 247–261.

Brysbaert and New 2009: Brysbaert, M. New, B. Moving beyond Kucera and Francis: A Critical Evaluation of Current Word Frequency Norms and the Introduction of a New and Improved Word Frequency Measure for American English. Behavior Research Methods, 41 (4), 2009, 977–990.

Jackendoff 1990: Jackendoff, R. Semantic Structures. MIT Press, Cambridge, USA.

EAGLES 1996a: EAGLES: Expert Advisory Group for Language Engineering Standards Preliminary recommendations on corpus typology. EAG–TCWG–CTYP/ P. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale.

EAGLES 1996b: EAGLES: Recommendations for the Morphosyntactic Annotation of Corpora. EAGLES Document EAG--TCWG—MAC/ R. Pisa: Consiglio Nazionale delle Ricerche. Istituto di Linguistica Computazionale.

EAGLES 1999: EAGLES LE3-4244: Preliminary Recommendations on Semantic Encoding, Final Report. http://www.ilc.pi.cnr.it/EAGLES/ EAGLESLE.PDF

Ide and Veronis 1998: Ide, N., Veronis, J., Word sense disambiguation: the state of the art. Computational Linguistics 24 (1), 1–40.

Fellbaum 1998: Fellbaum, C. (ed.). Wordnet: An Electronic Lexical Database. Cambridge, Mass.: MIT Press.

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

Fellbaum et al. 2001: Fellbaum, C., Palmer, M., Dang H., Delfs L. & Wolff, S. Manual and Automatic Semantic Annotation with Wordnet. In NAACL-2001 Workshop on Wordnet and Other Lexical Resources. Pittsburgh, Philadelphia.SIGLEX.

Fellbaum et al. 2007: Christiane Fellbaum, Anne Osherson, Peter E. Clark. Putting Semantics into WordNet's „Morphosemantic" Links. In: Proceedings from 3rd Language and Technology Conference: HLT as a Challenge for Computer Science and Linguistics, Poznan.

Fodor 1975: Fodor, Jerry. The Language of Thought. Cambridge, MA: Harvard .University Press.

Hanks 2000: Patrick Hanks: Do word meanings exist? In: Computers and the Humanities, 34(1–2):205–215.

Järborg et al. 2002: Järborg, J., Kokkinakis, D. and Toporowska Gronostaj, M. Lexical and textual resources for sense recognition and description. In: Proceedings LREC 2002. Las Palmas, 1492–1497.

Katz 1972: Katz Jerrold J. Semantic Theory. New York: Harper and Row.

Kilgarriff 1997: Adam Kilgarriff. I don't believe in word senses. Computers and the Humanities 31(2):91–113.

Kilgarriff 1998: Kilgarriff A.Gold Standard Datasets for Evaluating Word Sense Disambiguation Programs. In: Computer Speech and Language. Special Use on Evaluation 12(4), 453–472.

Kilgarriff and Rosenzweig 2000: Kilgarriff Adam and Joseph Rosenzweig. Framework and Results for English SENSEVAL. Computers and the Humanities 34(1–2) (Special Issue on SENSEVAL) 15–48.

Koeva 2006: Koeva, Sv. Inflection Morphology of Bulgarian Multiword Expressions. In: Computer Applications in Slavic Studies, Boyan Penev Publishing Center, Sofia, 201–216.

Koeva 2008: Koeva, Sv. Semantic Nature of Diatheses, In: Studia kognitywe = Etudes cognitives, Warszawa, Vol. 8, 71–95.

Koeva et al. 2006: Koeva, Sv., Sv. Leseva, M. Todorova, Bulgarian Sense Tagged Corpus. In: Proceedings of the 5th SALTMIL Confernece on Minority Languages: Strategies for Developing Machine Translation for Minority Languages, Genoa, 79–87.

Koeva et al. 2008: Koeva, Sv., B. Rizov, S. Leseva. Chooser — A Multi-task Annotation Tool, In: Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08), Marrakech, European Language Resources Association (ELRA) electronic publication, pp. 728–734, 2008.

Koeva, S. 2010a: Bulgarian Wordnet — Current State, Applications and Prospects. In Bulgarian-American Dialogues (pp. 120–132). Sofia: Prof. Marin Drinov Academic Publishing House.

Koeva, S. 2010b: Balgarskiyat semantichno anotiran korpus — teoretichni postanovki. In S. Koeva (ed.), Balgarskiyat semantichno anotiran korpus (pp. 7–42). Sofia: Institute for Bulgarian Language.

Kohl et al. 1998: Kohl K. T., Douglas A. Jones,. Robert C. Berwick, and Naoyuki Nomura Representing Verb Alternations in WordNet In: C. Fellbaum (Ed.), WordNet: An Electronic Lexical Database. Cambridge: MIT Press, 153–178.

Landers et al. 1998: Landers Shari, Claudia Leacock and Randee Tengi. Building Semantic Concordances. In: Word-Net: An Electronic Lexical Database and Some of its Applications. Ed. Christiane Fellbaum, Cambridge, Mass.: MIT Press,199–216.

Leech 1977: Leech, G. Introducing corpus annotation. In Garside R., Leech, G., McEnery, A. M. (eds.) Corpus Annotation: Linguistic Information from Computer Text Corpora. London: Longman.

Lupu et al. 2005: Lupu, M., Trandabăt[327?], D., Husarciuc M. A Romanian SemCor Aligned to the English and Italian MultiSemCor, 1st ROMANCE FrameNet Workshop, International Workshop held at EUROLAN 2005 Summer School, July, 26–28, 2005, University Babes-Bolyai, Cluj-Napoca, Romania, 20–27.

McEnery and Wilson 2001: McEnery A. M., Wilson A. Corpus Linguistics: An Introduction. Edinburgh: Edinburgh University.

Miller 1990: Miller G. Five Papers on Wordnet. Special Issue in International Journal of Lexicography, vol.3, no.4.

Miller et al. 1993: Miller G. A., Leacock C., Randee T., and Bunker R. A Semantic Concordance. Proceedings of the 3rd DARPA Workshop on Human Language Technology. Plainsboro. New Jersey, 303–308.

Miller 1995: George A. Miller. Building Semantic Concordances: Dissambiguation vs. Annotation. In: AAAI Technical Report SS-95-01, 92–94.

Mihalcea and Moldovan 2001: Mihalcea R. and D.I. Moldovan. Automatic generation of a coarse grained Wordnet. In: Proceedings of the SIGLEX workshop on „Wordnet and Other Lexical Resources: Applications, Extensions and Customizations", NAACL. Pittsburgh, USA.

Montemagni et al. 2000: Montemagni S., Barsotti F., Battista M., Calzolari N., Corazzari O., Zampolli A., Fanciulli F., Massetani M., Raffaelli R., Basili R., Pazienza M.T.,

Saracino D., Zanzotto F., Mana N., Pianesi F., Delmonte R. The Italian Syntactic-Semantic Treebank: Architecture, Annotation, Tools and Evaluation. Proceedings of the COLING Workshop on „Linguistically Interpreted Corpora (LINC-2000)", Luxembourg, 6 August 2000, 18–27.

Navarro et al. 2003: Borja Navarro, Montserrat Civit, Antonia Martı, Raquel Marcos, Beĺen Fernańdez. Syntactic, Semantic and Pragmatic Annotation in Cast3LB. Corpus Linguistics 2003 Workshop on Shallow Procesing of Large Corpora, Lancaster, UK.

Ng and Lee 1996: Hwee Tou Ng and Hian Beng Lee.Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In Proceedings of the 34th Annual Meeting of the Association for Computational Linguistics (ACL), 40–47.

Pianta and Bentivogli 2003: Emanuele Pianta and Luisa Bentivogli. Translation as Annotation. In: Proceedings of the AIIA 2003 Workshop „Topics and Perspectives of Natural Language Processing in Italy", Pisa, Italy, September 2003, 40–48.

Pustejovsky 1995: Pustejovsky, J. The Generative Lexicon, MIT Press, Cambridge, MA.

Wierzbicka 1996: Wierzbicka, A. Semantics: Primes and Universals. Oxford University Press, Oxford, UK.

Wilson and Thomas 1997: Wilson,A., Thomas J. Semantic Annotation. In R. Garside, G. Leech & A. M. McEnery, (eds.) Corpus Annotation: Linguistic Information from Computer Text Corpora. London: Longman.

Wu et al. 2006: Yunfang Wu, Peng Jin, Yangsen Zhang, Shiwen Yu: A Chinese Corpus with Word Sense Annotation. ICCPOL 2006: 414–421.