BULGARIAN TAGGED CORPORA

Svetla Koeva, Svetlozara Lesseva, Ivelina Stoyanova, Ekaterina Tarpomanova, Maria Todorova

Department of Computational Linguistics, IBL – BAS

52 Shipchenski prohod, Bl. 17, Sofia 1113, Bulgaria

svetla@ibl.bas.bg, zara@ibl.bas.bg, iva@ibl.bas.bg, katja@ibl.bas.bg, maria@ibl.bas.bg

ABSTRACT

The Bulgarian Part-of–Speech (POS) and Word-Sense (WS) Tagged Corpora are derived from the "Brown" Corpus of Bulgarian, automatically annotated respectively with POS and WS tags and manually disambiguated with the annotation application Chooser. The adopted methodology for constructing and preprocessing the source corpora is briefly described. The paper also presents the annotation criteria underlying respectively the POS and WS selection process. At the present stage 217 210 tokens (single words, punctuation marks and numbers) are POS annotated and 50 368 words (single words and multi-word expressions) are WS annotated. The chief intended application of the Bulgarian Tagged Corpora is to serve as a test and / or training dataset for POS and WS disambiguation with the further aim of developing a Bulgarian-English bi-directional machine translation system.

Introduction

The main objective of this paper is to present the Bulgarian Part-of-Speech Tagged Corpus (BulPoSCor) and the Bulgarian Sense Tagged Corpus (BulSemCor) both derived from the "Brown" Corpus of Bulgarian. The first one is annotated with Part-of-Speech (POS) tags from the Bulgarian Grammatical Dictionary, the second one - with Word-Sense (WS) tags available in the Bulgarian WordNet. The chief intended application of the Bulgarian Tagged Corpora is to serve as a test and / or training dataset for POS and WS disambiguation with the results to be further employed in the implementation of a Bulgarian-English bi-directional machine translation system.

1. Annotation tool

The annotation tool Chooser[1] is a multi-purpose multi-functional platform aimed at performing various tasks involving corpora annotation as well as at enabling automatic analysis and manual disambiguation of large volumes of text (Koeva et al., 2005a). So far two applications have been employed in linguistic disambiguation – a POS-tagging and a WS-annotation implementation.

The user interface of the tool is based on the Model-View-Controller paradigm and is divided into three primary functional areas: **Info View**, **Text View** and **Choice View**.

Text View displays the source corpus in a user friendly format. The application's viewer and editor functionalities provide text display management, various strategies for text navigation and editing. An important feature is the possibility for group selection of adjacent as well as distant units (e.g. multi-words expressions). On the selection of an item in the corpus (current item is a default) the chosen linguistic unit is synchronized with the information available for it in **Choice View** and **Info View**.

Choice View takes care of the display of and navigation along the annotation choices, as well as of the annotation proper. The set of choices associated with a particular language unit are retrieved from a database. The selection of a particular option in **Choice View** results in the assignment of this piece of information to the corresponding unit in **Text View**. In the POS implementation the choices provided by the database are labels including part-of-speech and grammatical information associated with it. The WS implementation' **Choice View** contain the glosses of the wordnet senses available for the annotated item.

Info View displays all the linguistic information available in the database for the selected linguistic item. Its function is to facilitate decision making by providing additional linguistic knowledge which helps annotators identify quickly the **Choice View** items. In the POS implementation the meta-information provided by **Info View** is description of the POS annotations, whereas in the WS implementation it is the synset associated with the selected **Choice View** option.

---

[1] The tool was developed by B. Rizov from the Department of Computational Linguistics (DCL).

Info View and **Choice View** are synchronized so that on navigation along the options in **Choice View** the information displayed in **Info View** is dynamically updated.

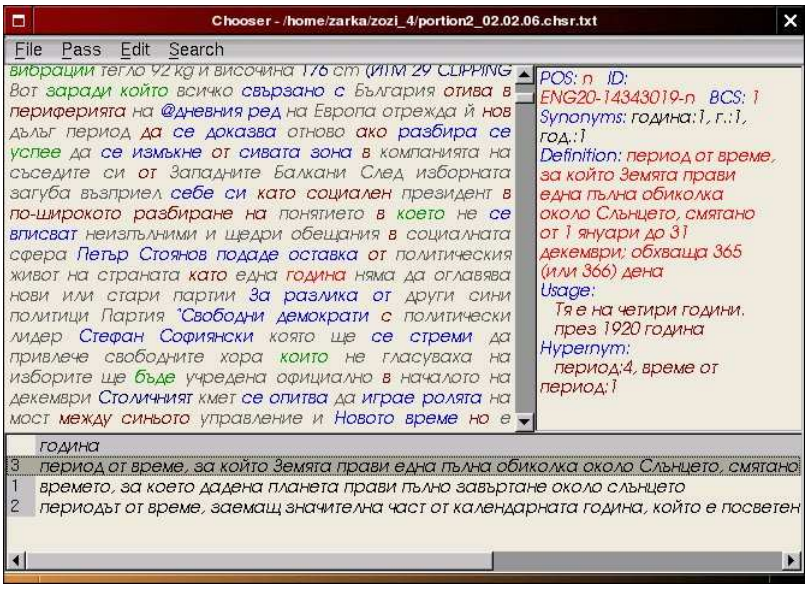


Figure 1: Layout of the annotation tool showing WS-annotation

Chooser is a multiple-user platform that performs dynamic interaction between the local users. A server takes care of a number of activities relating to user communication in two principal directions: 1) Interaction between each of the local users and the linguistic database; 2) Interaction among the local users.

2. Bulgarian POS tagged corpus

2.1. Pre-processing of the source corpus

POS annotation was performed on a subset of the “Brown” Corpus of Bulgarian (BCB). The corpus for annotation was built by selecting portions of 150+ words from each BCB file. Since sentence borders were respected, most of the portions contained more than the specified number of words. Tokenization was performed whereby a tag was attached to each token according to the type of characters the token consisted of: upper case, lower case alphabet characters, numbers, special characters, etc. In the course of work new tokanization rules were suggested, for example rules capturing compound words consisting of numbers, a punctuation mark and alphabet characters, e.g. *10-ite –‘the tenth*’, rules capturing dates, e-mail and web addresses, abbreviations, mathematical expressions, etc. Since punctuation was also subject to annotation, a system of the possible grammatical meanings of each punctuation mark was worked out on the basis of its function.

The automatic grammatical annotation of the corpus employed the Bulgarian Grammar Dictionary (BGD) (Koeva, 1998). BGD contains about 85 thousand words and over 1 million word forms specified with the respective grammatical characteristics (grammatical categories and their values). Semantic differences are not accounted for, except those reflecting on the grammatical meaning, e.g. subcategorization of verbs.

The BGD format represents a kind of DELAS-F dictionary (Courtois & Silberztein, 1990). At the pre-annotation stage tags containing all possible grammatical meanings found in the dictionary were attached to the corpus tokens. As a result of the automatic annotation 51,9% of the tokens were assigned one meaning, 46,7% had multiple meanings and 1,4% of the tokens were left untagged (Figure 2). Untagged tokens included rare words, foreign words or proper names not present in the dictionary.

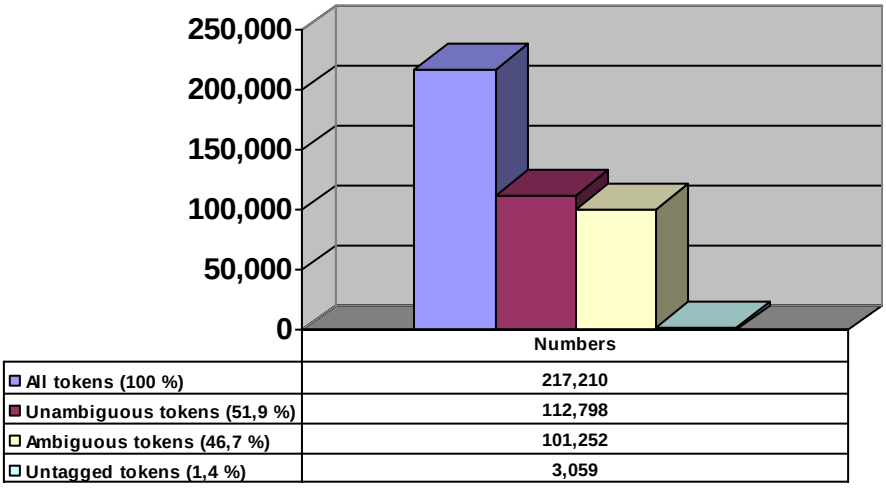


Figure 2: POS Tagged Corpus Result of automatic annotation

The rate of ambiguity (i. e. number of different meanings assigned to ambiguous tokens) varies for different types of tokens. Generally, it is greater for punctuation marks and closed class words because they have various grammatical functions as opposed to open class words, e.g. the comma has 25 different meanings as it has a number of different functions such as marking beginning/end of different types of subordinate clauses, or coordinate phrases and clauses, etc.

For the tags a standard DELAF structure was adopted consisting of: lemma. GRAMMATICAL CATEGORIES OF THE LEMMA: grammatical features of the word form (if any).

2.2. Development of the BulPoSCor

After the automatic tokenization and annotation, disambiguation proper took place. As a result a POS disambiguated corpus was obtained consisting of 217 210 tokens, including 172 482 single words, 42 058 punctuation marks and 2 670 numbers. Different types of systematic ambiguity were attested which fall into three groups: morphological, lexical and combined (Figure 3).

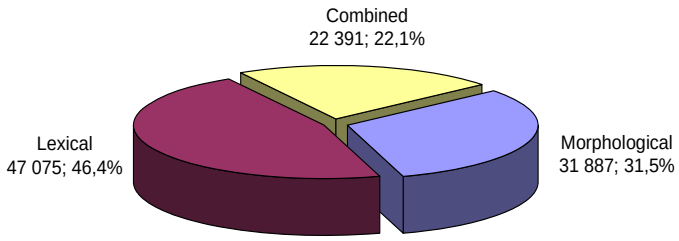


Figure 3: Types of ambiguity occurring in the POS tagged corpus

2.2.1. Morphological ambiguity

Morphological ambiguity occurs when a given lemma has two or more identical distinct word forms, e. g. inanimate masculine nouns such as '*film*' whose singular definite short article – (*sh*) and counted form (*c*) coincide: *filma{film.N+M:sh:c}*. Disambiguation in this case includes identification of the correct form in a given context and selection of the corresponding grammatical tag: *Gledah* ***filma*** *{film.N+M:sh} – 'I saw the film'* vs. *dva* ***filma*** *{film.N+M:c} –' two films'*.

2.2.2. Lexical ambiguity

The identical word forms pertaining to different lemmas (usually with different POS) are defined as lexically ambiguous, e. g. the word '*razhodite*' may either be the plural definite form of the masculine noun *razhod – 'expense',* or the second

person plural form, present tense of the verb *razhodya - 'to take for a walk':* ***razhodite****{razhod.N+M:pd,razhodya.V+F+T:R2p}.* This type of ambiguity requires the identification and selection of the correct citation form in order to disambiguate the corpus item (***Razhodite****{razhod.N+M:pd} sa nalozhitelni – 'The expenses are necessary'* vs. *Trybava da razhodite{razhodya.V+F+T:R2p} kucheto – 'You have to walk the dog.*'

2.2.3. Combined ambiguity

In a number of cases there exists grammatical ambiguity between different word forms of one lexical item and a word form of another item, e.g. *istoricheski* might be an adverb as in ***istoricheski****{istoricheski.ADV} dostoverno tvardenie* – '*historically authentic statement*', or masculine indefinite singular of the adjective, e.g. ***istoricheski****{istoricheski.ADJ:sm0} muzey* – '*history museum*', or indefinite plural form of the adjective, e.g. ***istoricheski****{istoricheski.ADJ:p0} fakti* – '*historical facts*'.

2.3. Annotation Criteria

The POS annotation was performed by human experts[2] whose task was to assign the correct sense out of two or more possible meanings for an ambiguous token. Some of the most frequent types of ambiguity include: coinciding 3rd conjugation verbs' forms; participles and adjectives; adjectives and adverbs; pronouns with various grammatical meanings (se/si); pronouns and particles (se/si); other closed class words. On the basis of the analysis of classes of ambiguous forms a number of annotation principles had been outlined in order to provide a uniform approach to the grammatical annotation.

2.3.1. Linguistic context analysis

In a number of cases the ambiguity of a word-form is resolved in the immediate context according to its morphological or syntactic function or certain lexical semantic properties derivable from the context.

2.3.1.1. Morphological analysis

Morphological analysis helps in resolving ambiguity on the basis of knowledge about the function of the forms in question in the composition of analytic forms. For instance, only the perfect past participle participates in the formation of the following analytical tenses: perfect, pluperfect, future perfect and future perfect in the past. Therefore, the identification of the tense e. g. *be pomagal* (had collaborated) disambiguates the participle (perfect vs. imperfect past participle) in the sentence: *Na tova vastanie toy [be* ***pomagal****] s vsichkite si sili - He had* ***collaborated*** *to this rebellion with all his strength.*

2.3.1.2. Syntactic analysis

The disambiguation of some types of grammatical ambiguity requires identification of the syntactic function of the ambiguous forms in the context, obtained by means of syntactic analysis. For instance, the ambiguity existing between adjectives and adverbs is resolved in the syntactic context as follows: adjectives are used as modifiers, e.g. *Firmata predlaga [*$_{NP}$ ***barzo*** *[proektirane [na obekti]]] - The company offers [*$_{NP}$ ***fast*** *[designing [of buildings]]]),* or predicatively: *Proektiraneto e* ***byrzo*** *– Designing is* ***fast****.;* or as adverbs - in adverbial adjunct position: *[*$_{VP}$ *Slyazoha* ***barzo****] ot avtobusa - They [*$_{VP}$ ***quickly*** *[got off]] the bus.*

2.3.1.3. Lexical semantic analysis

Lexical semantic analysis consists in disambiguating forms on the basis of their distinct meaning. It is applied to resolve cases of lexical ambiguity, as for example in the case of coinciding forms of participles and adjectives. Some participles have acquired a different meaning in their usage as adjectives and are encoded as separate adjectival entries in BGD. The two meanings may appear in ambiguous morphological or syntactic context and the above principles are therefore inapplicable. In these cases judgment as to the grammatical meaning in the particular context may be done on the grounds of the adjective possessing meaning interpretable as distinct and independent from the verb's semantics. In the sentence: *Nay-nakraya tya [be* ***ubedena****] da otide.– At last she [was* ***convinced****] to go* the word form *ubedena* is part of the passive voice form of the verb *ubedya* 'convince (someone to do something)', while in *Tya [be napalno* ***ubedena****] v pravotata si - She [was fully* ***convinced****] she was right* the form is the adjective *ubeden* 'convinced (in something)'.

---

[2] The annotators are I. Stoyanova, A. Koprinarova and P. Plachkova from the Department of Computational Linguistics.

2.3.2. Substitution

Various substitution tests had been worked out for the purposes of grammatical disambiguation prior to and in the course of POS annotation. They consist in the replacement of an ambiguous form with an unambiguous one belonging to the same or a different lemma in the same context without rendering the sentence incorrect or changing its interpretation. Substitution tests disambiguate a word form on the morphological, syntactic, semantic, etc. level. For example 3rd conjugation verb aorist and imperfect indicative forms are replaced by the corresponding forms of verbs belonging to 1st or 2nd conjugation which are unambiguous. Thus, the application of a substitution test to the participle *opitali 'tried'* in the sentence *Nyamalo da se primiryat, ako ne* ***opitali -*** *They wouldn't reconcile if they didn't try* with the unambiguous *svarshili 'done'* shows the form to be the perfect past participle: *Nyamalo da se primiryat, ako ne *svarshili (perfect p. p.)/* ***svarsheli*** *(imperfect p. p.) tova. - They wouldn't reconcile if they* ***didn't finish*** *this.*

2.3.3. Criteria for independent or combined application of principles

As the annotation is principally performed by native speakers the principles for disambiguation and the criteria for choosing one or another are largely based on language knowledge. When the context is indicative of the form occurring in the context, analysis is performed as it takes less time and effort. When the information provided by the context is not sufficient to identify the meaning, further analyses of the lexical semantic, syntactic and other properties of the words are carried out. The substitution tests are largely used when the context is not indicative of the form occurring in it, and as additional checks in some questionable cases. As can be seen from the example in 2.3.1.3. where semantic analysis involves also analysis of the syntactic properties, often a combination of disambiguation principles rather than an independent one is applied to obtain a more efficient disambiguation of certain ambiguity classes.

One of the most complicated cases of ambiguity is that of *se* and *si,* defined as pronouns or particles[3], whose disambiguation respectively involves combinations of different types of context analysis, as well as combinations of context analysis and substitution tests. *Se and si* may *occ*ur as lexical particles with verbs such as '*usmihvam se*' *(smile).* Semantic analysis possibly combined with other principles is needed where these verbs are paired by the same verbs without a lexical particle having different meaning, e. g. *namiram (find) – namiram* ***se*** *(be situated), predstavyam (present) – predstavyam* ***si*** *(imagine).* A set of substitution tests had been worked out to identify different syntactic meanings of *se* and *si*. E. g. if *se* may be replaced with the full reflexive pronoun *sebe si*, it is definitely a reflexive pronoun as is in the following example: *Toy* ***se*** *zashtiti ot napadkite – Toy zashtiti* ***sebe si*** *ot napadkite (He defended* ***himself*** *from the attacks)*. The reciprocal pronoun ***si*** is established through substitution by the reciprocal phrase *edin* [preposition] *drug 'each other'*, e.g.: *Kupuvame* ***si*** *podaraci – Kupuvame podaraci* ***edin na drug*** *(We buy gifts* ***to each other****)*.

As may be judged from the empirical evidence grammatical disambiguation combines different types of linguistic knowledge and respectively requires the joint application of more than one principles, as well as the further invention of techniques for newly encountered ambiguity classes.

2.4. Validation of the BulPoSCor and expanding the knowledge base – BGD

The first annotation pass of the corpus was aimed at disambiguating the ambiguous tokens, but it also served for detecting errors in the BGD dataset: (i) missing forms in the paradigm of some words; (ii) missing words; (iii) wrong forms; and (iv) errors in the grammatical description of some words. These errors had been corrected and new words had been added to the dictionary. As the first pass was focused on the ambiguous tokens, a second pass was performed to check the unambiguous tokens and to correct errors in their tags, or to add new grammatical meanings. Some unrecognized widespread words were also added to the dictionary. The rest of the untagged words were assigned grammatical description manually. At the final stage of POS disambiguation each annotator performed checks over the other annotators' parts of the corpus.

---

[3] The following senses of the forms *se/si* has been identified: possessive, reflexive or reciprocal meaning of the pronoun ***si***, reflexive or reciprocal meaning of the pronoun ***se***, reflexive, reciprocal or dative ethic meaning of the lexical particle ***si***, reflexive, reciprocal, third personal reflexive, third personal dative reflexive; impersonal reflexive or impersonal dative reflexive meaning of the lexical particle ***se***, and optative, impersonal optative, passive, impersonal passive particle ***se*** (Koeva & Leseva, 2006)

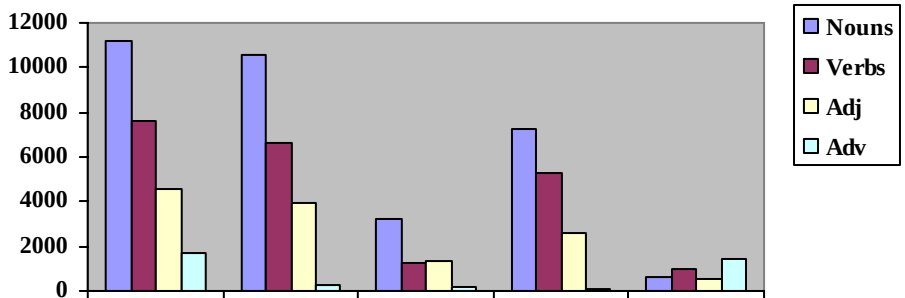
3. Bulgarian Sense Tagged Corpus

3.1. Pre-processing of the source corpus

The source annotation corpus (Koeva et al., 2006) consists of 500 excerpts (clippings) of approximately 100 words each, selected according to a criterion for a well-balanced distribution of highest frequency Bulgarian open-class lemmas located in BCB. To this end, relative weights were assigned to the lemmas featuring on the frequency lists; the weights were further calculated proportionally to the frequency of the lemmas' occurrence both in BCB and in BulNet. Greater lexical diversity was ensured by assigning larger weights to less frequent words. The clippings to be included in the corpus for annotation were selected among the best rating candidates with respect to lexical coverage. The latter was estimated in terms of the greatest number of different lemmas in combination with the greatest number of corresponding wordnet senses. 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.

In the development of BulSemCor the methodology adopted for the English semantically annotated corpus – SemCor, created at the Princeton University (Fellbaum et al., 1998) has generally been followed. The linguistic database which has served as a source for introducing and resolving ambiguity is the Bulgarian WordNet - BulNet (Koeva, 2004a).

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 one sense, 3 704 lemmas did not match a sense in BulNet at all, and 15 343 lemmas received more than one sense. Figure 4 below shows the distribution of open class lemmas in the resulting corpus across parts 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 have been subject to validation, and new senses have been encoded in case no sense encoded in BulNet has matched that of the word to be disambiguated. For lemmas not having corresponding entries in BulNet new synsets denoting the appropriate sense have been encoded (or the senses of already existing synsets have been revised) and assigned to the words in a second pass; for multiple sense lemmas the particular sense used in the context has had 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 4: Distribution of content-word lemmas across POS and coverage in BulNet

3.2. 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 was performed[4].

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

[4] S. Lesseva, E. Tarpomanova, M. Todorova and H. Kukova from the Department of Computational Linguistics and K. Alahverdzhieva and N. Radnev from the Master program in Computational Linguistics have worked in different capacity as annotators.

- 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 *Momicheto obitava malkata kashta - The girl lives in the small house* will be annotated as follows:

Momicheto{momiche#ENG20-09478614-n} (girl: a young woman)

obitava{obitavam#ENG20-01941830-v} (live: make one's home or live in)

malkata{malka#ENG20-01343705-a} (small: limited or below average in number or quantity or magnitude or extent)

kashta{kashta#ENG20-03141215-n} (house: housing that someone is living in)

Sense Tagged Corpus	
New senses added in BulNet	7645
Annotated units	50 368
Annotated single words	44 766
Annotated MWE	2 355
Words left for annotation	13 072

Table 1: Current state of BulSemCor

3.3. Annotation criteria for sense selection

Semantic annotation in the adopted methodology consists in the association of word occurrences in the corpus – single words and multiword expressions - with the appropriate senses in BulNet. Coverage has been ensured through the evaluation of the encoded data against the empirical evidence available in the corpus and the revision and enlargement of BulNet with new senses.

The following consistency criteria involved in the annotators' choice of a given sense from among the available candidates in BulNet have been identified in the course of the annotation:

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

The relation of equivalence defined between the members of a synset is the first criterion to be considered in selecting the most appropriate among the candidate senses. Substitution tests are applied to identify semantic equivalents of words found in the corpus. If such equivalence is established between a corpus item and another word, the correct sense is most likely the one encoded in a synset where the two items appear as co-synonyms. Of course, cross-check with other criteria is performed even in this case, to avoid possible errors due to incompleteness or errors in the database.

3.3.2. Consistency with 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 an important clue in choosing between senses.

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

Semantic relatedness between pairs of synsets facilitates decision making in cases where a word has a number of closely related meanings. The annotators consider the relatedness of the candidate BulNet senses to their hyperonyms, hyponyms and other semantically related synsets, on the one hand, and the context of the corpus item, on the other, in order to find out the best match for the latter.

3.3.4. Consistency with the usage examples

Usage examples help to infer the meaning of a synset by illustrating it, and provide quick browsing through different senses of a word as well as of potential candidates for encoding.

Most often the decision is a result of the interaction of the above criteria. For example, in the sentence *Tya vklyuchi **svetlinite** - She switched on the **lights*** the related synsets are *{svetlina:2, osvetlenie:1} (corresponding to PWN 2. 0. {light:2})* with the definition *'any device serving as a source of illumination', and* the synset *{svetlina:1} (corresponding to PWN 2.0. {light:10})* with the gloss 'visual impression for having abundant light or illumination' and *{light:1, visible light:1, visible radiation:1}* - 'electromagnetic radiation that can produce a visual sensation'. The synonym *osvetlenie* readily replaces *svetlinite* without change in the interpretation or rendering the sentence incorrect. Further cross check with the hyperonyms *{source of illumination:1}* - 'any device serving as a source of visible electromagnetic radiation' and *{illumination:2}*- 'the degree of visibility of your environment' and *{actinic radiation:1, actinic ray:1}* - 'electromagnetic radiation that can produce photochemical reactions', respectively, as well as with the pertaining usage examples corroborate the judgment in favour of the first synset.

3.3.5. 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, e.g. ***Britantsite** sa velika natsia - The **Brits** are a great nation* where the sense to be assigned to ***britantsite*** corresponds to the synset: *{British:1, British people:1, the British:1, Brits:1},* defined as: '*the people of Great Britain'* whereas ***britantsi*** in *Dvama **britantsi** byaha spaseni - Two Brits were rescued,* is semantically equivalent to*: {Britisher:1, Briton:1, Brit:1},* defined as: '*a native or inhabitant of Great Britain'.*

3.3.6. Appropriateness with respect to the available senses encoded in Princeton Wordnet (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, on exploring PWN for the sense of ***priroda*** (*nature)* in the sentence ***Prirodata** se e grizhila za nas prez vekovete i vse oshte otkrivame neynite chudesa - **Nature** has taken care of us for centuries and we are still discovering her many wonders,* one can see that the sense *{nature:2} `a causal agent creating and controlling things in the universe`* corresponds best to the meaning of the word in the sentence. It is therefore encoded in BulNet and associated with the corpus occurrence in the next pass.

4. Validation of BulSemCor and expanding the knowledge base – BulNet

The evaluation of the Bulgarian Sense Tagged Corpus has so far been performed by a second annotator. Further strategies of evaluation (i.e. different treatment of single-sense and multiple-sense words; comparison with the senses attested in the existing lexicographic works; enlargement with wordnet senses that are not mapped in BulSemCor) have been developed in order for a higher level of consistency to be guaranteed in the semantic annotation.

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

- culture specific concepts, for example **bogomilstvo**: an orthodox heretic sect founded by the Bulgarian priest Bogomil; the language-specific concepts shared among Balkan languages were linked via a BILI (BalkaNet ILI) index (Tufis et al., 2004);

- language specific instances of lexicalization, such as ingressive verbs, classifying adjectives derived from nouns, feminine gender nouns, diminutives, etc. (Koeva at al. , 2005b);

- missing English senses and unaccounted systematic differences between senses, for example: moderniziram: ‘change something in a positive direction by implementing technological achievements’;

- closed word classes – BulNet has been artificially expanded to incorporate in a systematic way the classes of prepositions, conjunctions, pronouns, particles, modal verbs, etc.;

- proper nouns denoting unique entities – person names, geographical names, names of institutions, companies, etc.;

- multi-word expressions that denote a unique and constant concept, e.g. otbivam nomera 'do something negligently without caring much about the result';

- domain relations – in the pre-annotation stage all adverbs in the BulNet database were linked to a synset corresponding to their semantic domain (such as time, location, manner, quantity, degree, frequency, etc.) through an extralinguistic relation: Category domain; grammatical peculiarities and syntactic function of certain items such as intensifiers, quantifiers, etc. were accounted through linking these items to the relevant domain synset by means of the relation Usage domain.

5. Conclusions

The Bulgarian POS Tagged Corpus is fully disambiguated and had successfully been used for the purposes of POS tagging (Doychinova and Mihov 2004). The Bulgarian Sense Tagged Corpus contains 63 440 words including single words and MWE. Eighty percent of the corpus has been annotated and the results have been employed in the experiments on developing a Hidden Markov Model formalism (for the time being relatively low recall but high precision has been achieved) underlying a WS disambiguation system. BulPoSCor and BulSemCor provide an essential foundation for a number of future purposes with a special emphasis on machine translation.

References

Courtois B. and M. Silberztein. 1990. Dictionnaires électroniques du français. In B. Courtois and M. Silberztein, eds., *Langue française 87*. Larousse: Paris.

Doychinova V. & S. Mihov. 2004. High Performance Part-of-Speech Tagging of Bulgarian. In *Proceedings of Eleventh International Conference on Artificial Intelligence: Methodology, Systems, Applications (AIMSA-2004).* LNAI #3192, pp. 246-255.

Fellbaum et al. 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.

Koeva, S. 1998. Bulgarian Grammatical Dictionary. Description of the linguistic data organization concept - Bulgarian language, 6, 49-58.

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. Flexible Framework for Development of Annotated Corpora.In *International Journal Information Theories & Applications, Sofia* [In press].

Koeva et al. 2005b. Resources for Processing Bulgarian and Serbian. In *Proceedings from the International Workshop Language and Speech Infrastructure for Information Access in the Balkan Countries. Borovets, pp. 31-39.*

Koeva et al. 2006. Bulgarian Sense tagged Corpus. In *LREC2006 Proceedings, Genoa, 23-25 May 2006, ELRA, Genoa-Paris 2006*.

Koeva S. and S. Leseva. 2006. "Word sense Disambiguation for Purposes of Machine Translation – The Nature of Bulgarian Pronominal Clitics studied by means of INTEX". In *Proceedings of the 6th and 7th INTEX Workshop* [In press].

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