Shallow Semantic Annotation of Bulgarian

Kiril Simov and Petya Osenova

Linguistic Modelling Laboratory - IPP, BAS
Acad. G.Bonchev Str. 25A, Sofia, Bulgaria
{kivs, petya}@bultreebank.org

Abstract

The paper discusses shallow semantic annotation of Bulgarian treebank. Our goal is to construct the next layer of linguistic interpretation over the morphological and syntactic layers that have already been encoded in the treebank. The annotation is called shallow because it encodes only the senses for the non-functional words and the relations between the semantic indices connected to them. We do not encode quantifiers and scope information. An ontology is employed as a stock of the concepts and relations that form the word senses. Our lexicon is based on the Generative Lexicon (GL) model (Pustejovsky 1995) as it was implemented in the SIMPLE project (Lenci et. al. 2000). GL defines the way in which the words are connected to the concepts and the relations in the ontology. Also it provides mechanisms for literal sense changes like type-coercion, metonymy, and similar. Some of these phenomena are presented in the annotation.

1. Introduction

Semantic annotation is a relatively new area of research and it still lacks a good methodology and a definition of the range of phenomena it has to cover. In this paper we offer one approach to decide these problems which is based on the Generative Lexicon (GL) model (Pustejovsky 1995) as it was implemented in the SIMPLE project (Lenci et. al. 2000). Under shallow semantic annotation we assume annotation of the senses of nouns, adjectives, adverbs and verbs as well as the relations among semantic indexes within the text. Each sense of a lexical item constitutes of one or more concepts in an ontology (as defined in SIMPLE project) and the arguments of the lexical item. The relations are introduces on the basis of the definition of the senses. They are also defined in the ontology.

Most of the current approaches to semantic annotation (or sematic tagging) exploit the following methodology: (1) select a corpus to be annotated; (2) select a lexicon (dictionary) as a source of senses to be attached to some chunks in the text. For instance, SemCor corpus (Landes et. al. 1998) the corpus is a part of the Brown corpus, the senses are taken from WordNet version 1.6. The corpus contains 700 000 running words annotated with part of speech, 200 000 of them are also annotated with their lemma and corresponding senses. The main problems encountered in the process of annotation, beside the inter-annotator agreement, were the annotation of *idioms*, *metaphors*, and *foreign words*. These were additionally processed by the lexicographers working on WordNet. Another class of words that received a special treatment was the class of proper nouns. They were annotated first with general labels: *person*, *location*, *institution* and *other*, then an appropriate sense from WordNet is used to annotated the proper noun. In our work we extend this general setup as follows. The actual annotation is done on the basis of the Bulgarian HPSG-based treebank — BulTreeBank (Simov and Osenova, 2003). The annotation already presented within the treebank supports the shallow semantic annotation in many respects, e.g. the phrases that are arguments of different relations are already determined and annotated. As a stock of senses and relations we are using the Bulgarian Semantic Dictionary (BSD), which is based on the guidelines of the SIMPLE project and for each word in it represents the ontological class (or classes in case of systematic polysemy), its valency frame and qualia role's relations. We are using ontological classes for sense annotation and also for domain and range constraints over the relations in the qualia structure and in the argument structure. The relations and the argument structure support bridging relation annotation. We build on the co-reference relations in the treebank, which include the following relations: `equality`, `member-of`, `subset-of`.

The structure of the paper is as follows: Section 2 presents in brief the current state of the treebank. In Section 3 the model of the semantic dictionary is discussed. Section 4 focuses on the issues of the sense annotation. Section 5 describes the annotation of the relations. The last section presents authors' conclusions and future work. The work reported here is an extension of (Simov and Osenova 2005).

2. Current State of the Treebank

Currently the treebank comprises 214000 tokens, a little more than 15000 sentences. Each token is annotated with morphosyntactic information. This fact is very important, because from this level we can derive useful semantic information for next levels of annotation. For example, the information that possessives introduce two indices — with respect to possessor and with respect to the entity, which is possessed. Additionally the Named Entities are annotated with ontological classes as `person`, `organization`, `location`, and `other`. Based on HPSG theory the annotation scheme defines a number of phrase types which reflect both - the constituent structure and the head-dependant relation. Thus we have phrase labels with the explication of the dependant types like `VPC` (verbal head complement phrase), `VPS` (verbal head subject phrase), `VPA` (verbal head adjunct phrase), `NPA` (nominal head adjunct phrase) etc. Behind the constituent structures and the head-dependant relations the treebank also represents phenomena like ellipsis, pro-dropness, word order, secondary predication, control. As an important mechanism for dealing with these phenomena we are using co-reference relations.

During the second phase we are extending the co-reference annotation and the semantic (ontological) annotation as it is discussed in the next sections. In this work we rely on the experience gained from our earlier experiments on extraction of semantic lexicons from our treebank. Thus, we can think of the treebank as non-purely-syntactic database even at this moment.

An important notion that we use in the paper is *markable*. It was introduced in order to denote the segments of texts that play important role within an annotation scheme. In the context of our treebank markables denote the arguments of the co-reference and bridging relations. We rely on the syntactic annotation and thus (similarly to others) our markables are among phrases encoded within the treebank. We follow the second principle of (Chiarcos and Krasavina 2005) and consider the maximal constituents. Thus although following HPSG theory in which the discourse indices are borne by the head noun, the modifiers are also included in the markable, because they additionally constrain the interpretation of the semantic value of the noun phrases. Note that according to our annotation scheme the discontinuous elements also can be well part of markables. More on the semantic weight of markables is presented below.

3. Bulgarian Semantic Dictionary

In this section we present the model of the lexical entries which we exploit in the process of the creation of the Bulgarian Semantic Dictionary. This dictionary is used in the process of shallow semantic annotation. The model follows in detail the model of the semantic dictionary within the SIMPLE project (see (Lenci et. al. 2000)). The SIMPLE model is an extension of the model developed within the Generative Lexicon (GL) of James Pustejovsky — (Pustejovsky 1995). According to GL model each lexical item comprises the following elements:

1. Argument list;
2. Event structure;
3. Qualia structure; and
4. Hierarchical inheritance.

This information is encoded via different mechanisms within the SIMPLE model. The qualia structure is presented on two levels: first, within an upper ontology called *Core Ontology* and second, in each lexical entry where the qualia structure stated in the Core Ontology is further specialized by a set of specific relations. The argument structure, the event structure and the specialization of the qualia structure are defined within the semantic templates of SIMPLE. These templates are used for the creation of the actual lexical entries. The hierarchical inheritance in SIMPLE is defined by connection of lexical entries with the concepts within the Core Ontology with the idea this to be extended to a full hierarchical structure. In the creation of the Bulgarian Semantic Dictionary we are encoding the semantic category of the lexical item (presented as a concept in the ontology), the qualia structure and the argument structure.

Qualia Structure, as it was mentioned, in SIMPLE follows the main division of GL model: *Formal*, *Constitutive*, *Telic*, *Agentive*. Each qualia structure determines one or several relations between an instance of the given concept and instances of other concepts in general. Thus, each concept potentially has a complicated structure. This complex structure is available when the corresponding lexical item is used in an utterance and allows different levels of substitution of one concept with another. In this way GL model handles different kinds of type coercion.

The SIMPLE project developed further the GL model by constructing an upper ontology that reflects the four qualia roles of GL model. The main difference from other upper ontologies is that the first level of concepts under the top are in fact the qualia roles: *Constitutive*, *Telic*, *Agentive*, represented as concepts and *Entity* as a fourth very general concept. *Entity* can be thought as representing the qualia role *Formal*. In addition to each concept can have each of the qualia roles *Constitutive*, *Telic*, *Agentive* explicitly specified. In the Core Ontology, for example, the concept *Artifact* obligatorily has values for the *Telic* and *Agentive* roles. The values of the qualia roles are semantic functions or semantic relations. These functions and relations connect the instances of the given concept to instances of other concepts to which the given instance is connected with the corresponding relation or function. The set of semantic relations and semantic functions is called *an extended qualia structure*. Thus, we can consider the SIMPLE Core Ontology as a set of four interconnected ontologies: one ontology consisting of a hierarchy of concepts, additionally connected to other concepts via relations and functions and three ontologies consisting of hierarchies of relations. Each of the relational ontologies corresponds to one of the qualia roles. Having classification of each relation in one of the hierarchies determines its usage in the interpretation of the corresponding lexical item in the text. For example, the values of *Telic* and *Agentive* relations can be used as source of type-coercion in sentences like the following: "John finished the book at the night."

The *Formal* qualia role determines the concepts in ontology which defines the core meaning of the lexical item. One important feature of the GL model is that the value of this qualia role can be not just a simple concept, but also a complex type. A complex type is a tuple of simple concepts. Each complex type represent a class of regular polysemy. For some examples see below.

We have extended the ontology to cover some specific domains that are more frequently presented within the treebank. The extension is mainly in the area of concepts for named entities like locations, professions and similar. This information is also used for partial analyses of Bulgarian texts. Each lexical item within the Bulgarian Semantic Dictionary is described by the most specific concept from the ontology and a set of relations for each qualia role appropriate for the lexical item. In fact the creation of the Bulgarian Semantic Dictionary is interconnected to the extension of the ontology.

Pustejovsky's argument list includes all the dependents of the lexical item, classified in four categories: *true arguments*, *default arguments*, *shadow arguments* and *true adjuncts*. In SIMPLE model only the true arguments are presented within the templates. We also encode mainly them,

although in future we envisage also to encode the other types of arguments. For each argument we specified its morphosyntactic features and its semantic restriction. The morphosyntactic features are represented as generalization over the morphosyntactic tagset used within the treebank. The semantic restrictions are concepts from the ontology.
The Bulgarian Semantic Dictionary is under construction. It is based on several machine-readable dictionaries: a Morphological Dictionary of Bulgarian, a Valence dictionary of Bulgarian and an Explanatory Dictionary of Bulgarian. The current size of the dictionary is 7000 lexemes which were selected on the basis of their normalized frequency in a large Bulgarian corpus (72 million running words).

4. Sense Annotation

We annotated each noun, adjectives, adverbs and verbs with senses from BSD. The model of the dictionary provides a good way for the representation of obligatory ontological relations for a given word and the polysemy of the word (if any). The first type of information is encoded as *an extended qualia structure* and the second is represented via *a complex type*. For example, in the first case, the noun *committee* contains information that each committee has *people* as members. Thus, each committee participates as an argument of the relation `member-of` and in a given discourse we can expect some of the members to be expressed by a noun phrase annotated as belonging to the class `person`. An instance of the second case is the noun *book* which can be considered as an artifact and as an information object. Very often the context determines one of the senses: *The second book on the shelf* (an artifact) and *an interesting book* (an information object), but sometimes both senses are presented: *the second book from the series is interesting*. In such cases we annotate the word with the complex type and additionally highlight the distinct senses used in the context. In the last example above both senses for *book* are selected. Thus, each lexical item can be connected to one or more concepts in the ontology depending whether its meaning is represented by complex or simple type. The shallow semantic annotation includes, also, argument annotation (mainly for verbs) and mapping from the arguments to the corresponding constituents.
The annotation consists of the following steps:

1. Sense Assignment.

   Each open class lexical item in the text is looked-up in the dictionary and all semantic classes found are assigned to it. If the argument structure is not empty, it is also attached to the lexical item. For verbs each of the arguments of the verb is connected to some of the verb dependents in the syntactic annotation. This is possible for the subject, the direct object and the indirect object.

2. Disambiguation.

   On the basis of the available semantic information and the information already encoded in the treebank: the constituent structure and all the co-reference relations; automatic disambiguation over the semantic classes and the frames is done along the lines described in (Simov and Osenova 2004).

3. Manual Semantic Annotation.

   The annotator processes manually the unresolved cases, selecting the correct semantic classes and attaching the frame arguments with the appropriate phrases in the treebank.

4. Manual Relation Annotation.

   The annotator inspects all appropriate relations defined in the lexical entries and searches for appropriate arguments. The annotator has the freedom to add new relations, but they are marked as unsupported by the dictionary and thus by the ontology.

5. Validation.

   The annotated relations and arguments are checked automatically for consistency with respect to the semantic classes. The new unsupported relations are checked by an expert and if necessary they are added to the dictionary.

In order to support this architecture of annotation we have to ensure that the mentioned above cases of regular polysemy are adequately encoded. To this we need to introduce the notion of semantic indices. These indices represent the arguments of the relations and the instances of the concepts. We discuss the semantic indices in the next subsection.

4.1. The Semantic Indices

We assume that each concept introduces an index in the interpretation of the corresponding lexical item. This index may be simple or decomposable. For example, the index of the word 'book' is decomposable, because it comprises two indices: one for *Information Object* (under the concept *Representation*) and one for *Semiotic Artifact* (under the concept *Concrete entity*). We aim at distinguishing this polysemy on index level, i.e. we focus on Formal qualia. The remaining qualia which connect the notion with other related entities are suppressed unless stated explicitly. In SIMPLE the concept *Book* as an *Artifact* has *Agentive qualia* `created-by`, has *Telic qualia* `used-for` and *Constitutive* `is-in`. Book as an information object has *Agentive qualia* `result-of` writing and *Telic qualia* exhibits the relation `give-knowledge`. All these relations are not represented in the annotation except when the related entities are not expressed explicitly.
Explication of indices on markables is important also for establishing the referential relations. The more qualia presented, the better the presentation of relations among indexes. However, it is too ambitious to encode all the net-like depth of the qualia. For that reason, we concentrate on gross information only, i.e. on regular polysemy ('school' as Building and Institution, 'box' as Container and Quantity, etc.) One question to be considered is whether it is always necessary or possible to disambiguate between senses and if yes, when. Very often several senses can co-exist or one of them can be suppressed in some context. The number of the polysemous classes is language specific. Due to conversion in English 'fox' means Animal and Material, while in Bulgarian it is only Animal. Material is encoded in an adjective (*lisitsa* (noun) - *lisitchi* (adjective)).

The indices on markables have to be viewed at least with respect to the following criteria: (1) accessibility (complexity), (2) boundness and (3) type-instance distinction. The first criterion deals with the following facts: (i) what index is projected to the mother node in the recursive NPs, (ii) the type of the NP, and (iii) disambiguation of polysemy. For example, in general, unique person, location and organization names are easily accessible in spite of the complexity of the NP. However, very often the main problem is to decide which index is needed in order to establish the appropriate discourse relations. For example, in the sentence: *Predstoi sreshta mezhdu SDS i DPS, saobshti chlenyt na NIS Rositsa Totkova* (A meeting is forthcoming between SDS and DPS, the member of NIS Rositsa Totkova announced.) there are two organizations, which will meet (SDS and DPS). The interesting case is the NP 'the member of NIS Rositsa Totkova'. Here the head is 'member', which means that Rositsa Totkova is a `member-of` NIS. However, NIS would have an index that it is an organization and is a `subset-of` SDS. Here the relation is viewed as transitional and thus Rositsa Totkova happens to be also a member of SDS. The difficulty comes from the fact that within the sentence there is no direct evidence that NIS is part of SDS. We rely on broader context or our world knowledge competence. Thus, it follows that in order to track properly `part-of` and `member-of` relations both indices are necessary - from the syntactic head and from the dependant.

Another example is the following: a text, in which an invention is discussed, namely a mouse, which memorizes the finger prints of the people. Globally, in this text there are two main `part-of` chains. One is 'finger' as `part-of` human and the other is 'mouse' as `part-of` the computer. Thus, we need an extension of domain specific `part-of` groups in order to track these different cases. In this way we will try to achieve better coverage of small domains. 'Green areas' of uncertainty are not annotated. Otherwise, the main relation, of course, remains `is-a` one. In the text the human participant is called: people, user, person. In this chain there is an agent of temporary activity (user) who `is-a` person, who is a `member-of` people. The next important relation is `part-of`. Thus, additionally to the main chains we mark that the mouse has a sensor in its upper part and the computer has a memory.

The second criterion is 'boundness'. Note that it correlates with the third one as well – 'type-instance' distinction. For that reason we will discuss them both here. Usually such cases follow the pattern: quantified NP as an antecedent of anaphoric expressions ((Hinrichs et. al. 2005, p. 14)). For example, in the sentence *Neka vseki da si plashta za tova pravo* (Let everybody pays for this right) we establish coreference relation between 'vseki' (everybody) and the reflexive clitic 'si'. However, we do not make explicit the fact that the referents are distributionally bound. We decided to analyze such examples as binding at least one instance (which makes their surface interpretation closer to referential non-quantificational markables). In this way, we avoid contradictions and at the same time we do not get stuck in the problems with the quantificational scope. With respect to type-instance relation, let us consider the following sentence: *Sistemata e tolkova chustvitelna, che bi identificirala daden chovek, dori ako toj si e poryazal prysta.*(The system is so sensitive that it would recognize a certain person, even if he had cut his finger.) Here 'human' is identifiable, but still not identified and hence 'he' refers to a potential referent. The same holds very often for names. There is difference between: *Te svikaha Veliko narodno sabranie* (They summoned the Grand parliament) and *Te iskaha da svikat Veliko narodno sabranie* (They wanted to summon a Grand parliament). In both Bulgarian examples there is null marker of indefiniteness. However, the first sentence refers to a concrete Parliament while the second refers to a potential one. The temporal and modal relations can be traced from the morphosyntactic annotation. However, the notions receive their type as a safety marker. Additionally, when they are really or potentially singled, they are marked as instances as well.

5. Annotation of Relations

As we have mentioned above, we have annotated in the treebank the following relations: `equality`, `subset-of` and `member-of`. We capture all main coreferences of the following syntactic representations: subject and object relations, reflexivity, possession, clitic-doubled structures, secondary predicated phrases with the subject or the object. Also we represent co-reference between synonymic expressions, changed referring expressions in direct-indirect speech, nominalizations.

Our goal in the semantic annotation is to extend the current co-referential annotation in two ways. First, we add new relations to cover more complicated cases of co-reference including `part-of` relations and some other relations between entities in the discourse (for instance, some other bridging relations - see (Gardent et. al. 2003)). Second, the relations are ontologically defined. These ontological definitions are done via constraints over the arguments of the relations. Some of these constraints are already defined in the ontology that is a basis of our semantic lexicon (SIMPLE Core Ontology). For example, the relation `member-of(I,M)` (`Is_a_member_of` relation in SIMPLE Core) is constrained in the following way: `I` is a member or element of `M` where `I` is a shaped, countable entity, and `M` is typically a collective entity, i.e. a set of individuals. Another example is the constraint that for the relation `part-of` both arguments need to be of the same ontological category (with the requirement the category not to be the top of the ontology). Thus we can establish the following relations: `part-of(wheel#1, car#3)` and `part-of(instruction#121, manual#86)`, but not the following relation `part-of(instruction#121, car#3)`. These definitions are used to validate the encoded relations and to facilitate the annotation process. In some cases the constraints can be violated by some instance, and then we define more specific relations subsumed by the currently defined relations.

One interesting case of combining the decomposable indices and the relations is given in the following example[1]:

[1] We grateful to the anonymous reviewer who pointed us to this example.

"Take the second book on the shelf. It's very interesting." Here, as we discussed above, to the noun book a decomposable index is assigned — one of its sub-index for book as an artifact and one for book as an information object. But in the anaphoric relation between 'book' and the pronoun 'it' only the second index of book is an argument of the relation.

The inventory of relations that we explicate in the annotation are the relations given in the BSD. Some of the relations are very specific and they are determined by a particular lexical entry like between nouns 'murderer', 'killer' and 'murder'. Having all of these relations encoded explicitly in the dictionary allows us to control the consistency of the annotation because the semantic classes of the arguments of a relation in the text must be compatible with the restrictions given in the lexical entries.

6. Conclusions and Future Work

In this paper we discussed an approach to shallow semantic annotation of Bulgarian texts. The combination of an ontology and a lexicon seems to be fruitful with respect to consistency of the annotation. The chosen lexicon model — Generative Lexicon — proved to be very appropriate for annotation of regular polysemy. In our current and especially in the future work we exploit the elements of the model to cope with some problematic cases. Here we discuss briefly some of them:

6.1. Adding of Time Stamps in the Ontology

Although modeling of time is complicated issues (Hobbs and Pustejovsky 2003), adding simple time stamps to concepts in the ontology is relatively easy. Then some time dependant relations can be encoded in the annotation.

For example, in the case of co-reference relation annotation we also take into consideration some of the critics on some early annotation schemes on co-reference relation (Deemter and Kibble 2000) which in our view has ontological origin. Such an example is the co-reference between an attribute and its value at a given moment. For such cases we are planning to introduce also time stamp as an argument of the relation. Thus, in the sentence *"The stock price fell from $4.02 to $3.85."* the NP 'The stock price' is related to both values, but with different time stamps.

6.2. Metonymy

Generally, metonymy can be defined as a relation between two indices: one which is actually used in the text and second, related to the first and which is actually meant by the speaker (author of the text). For example, "She was wearing stripe." First we annotate 'stripe' as Property and as such it is connected to 'cloth' via Constitutive role and 'cloth' which is annotated as Material and it is connected to 'clothing' again via the Constitutive role. The concept 'clothing' is of the relevant type for the object of the verb 'to wear'. Thus the understanding of the sentence is something like: "She was wearing a clothing made from a textile with a stripe design." Here the index introduced by 'stripe' is connected to the index of 'clothing' via a composition of relations. For more details see (Simov and Osenova 2006).

6.3. Formal Treatment

At the moment the ontology we are using is not formally represented. We are planning an encoding of the ontology in some of the standard ontology languages like:

- RDF

  http://www.w3.org/TR/1999/REC-rdf-syntax-19990222/

- RDF(S)

  http://www.w3.org/TR/2000/CR-rdf-schema-20000327/

- OWL

  http://www.w3.org/2001/sw/WebOnt/

Such encoding will facilitated the usability of the annotations. First, it will allow comparison with other upper ontologies and incorporation of the new developments in the area. Second, it will provide support of easy incorporation of domain ontologies and supporting annotation of domain specific documents.

This line of research is under development within the European project Language Technology for eLearning (LT4eL).

7. Acknowledgements

The work reported in the paper is supported by two project: BulTreeBank project which is funded by the Volkswagen Stiftung, Federal Republic of Germany under the Programme *"Cooperation with Natural and Engineering Scientists in Central and Eastern Europe"* contract I/76 887; and the Language Technology for eLearning (LT4eL) project which is supported by the European Community under the Information Society and Media Directorate, Learning and Cultural Heritage Unit.

The authors would like to thank to the three anonymous reviewers for their valuable comments.

8. References

Christian Chiarcos and Olga Krasavina. 2005. *Annotation Guidelines. POCOS - Potsdam Coreference Scheme.* Draft. Available on-line from http://amor.cms.hu-berlin.de/k̃rasavio/annorichtlinien.pdf

Kees van Deemter and Rodger Kibble. 2000. *On Coreferring: Coreference in MUC and related annotation schemes.* Computational Linguistics 26, No. 4.

Claire Gardent, Helene Manuelian and Eric Kow. 2003. *Which bridges for bridging definite descriptions?* In: Proceedings of the 4th International Workshop on Linguistically Interpreted Corpora. Budapest, Hungary.

Erhard Hinrichs, Sandra Kübler and Karin Naumann. 2005. *A Unified Representation for Morphological, Syntactic, Semantic, and Referential Annotations.* In Proceedings of the ACL Workshop on Frontiers in Corpus Annotation II: Pie in the Sky. Ann Arbor, MI.

Jerry Hobbs and James Pustejovsky. 2003. *Annotating and Reasoning about Time and Events.* In: Proceedings of AAAI Spring Symposium on Logical Formalizations of Commonsense Reasoning. Stanford, California, USA.

Shari Landes, Claudia Leacock and Randee I. Tengi. 1998 *Building semantic concordances.* In Fellbaum, C. (ed.) (1998) WordNet: An Electronic Lexical Database. Cambridge (Mass.): The MIT Press. pp 199–216.

Alessandro Lenci, Federica Busa, Nilda Ruimy, Elisabetta Gola, Monica Monachini, Nicoletta Calzolari, Antonio Zampolli, Emilie Guimier, Galle Recourc, Lee Humphreys, Ursula Von Rekovsky, Antoine Ogonowski, Clare McCauley, Wim Peters, Ivonne Peters, Robert Gaizauskas, Marta Villegas. 2000. SIMPLE Work Package 2 - Linguistic Specifications, Deliverable D2.1. ILC-CNR, Pisa, Italy.

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

Kiril Simov and Petya Osenova. 2003. *Practical Annotation Scheme for an HPSG Treebank of Bulgarian*. In: Proc. of the 4th Workshop on Linguistically Interpreteted Corpora, Budapest, Hungary.

Kiril Simov and Petya Osenova. 2004. *A Treebank-Driven Approach to Semantic Lexicons Creation.* In: Proceedings of the Third Workshop on Treebanks and Linguistic Theories 2004, Tuebingen, Germany.

Kiril Simov, Petya Osenova. 2005. *Extending the Annotation of BulTreeBank: Phase 2.* In: Proceedings of the Fourth Workshop on Treebanks and Linguistic Theories 2005, Barcelona, Spain.

Kiril Simov and Petya Osenova. 2006. *Ontology-Based Lexicon and Semantic Annotation.* In: Proceedings of the Workshop on Ontology-based Modelling in The Humanities. Hamburg, Germany.