Practical Annotation Scheme for an HPSG Treebank of Bulgarian

Kiril Simov and Petya Osenova
BulTreeBank Project
Linguistic Modelling Laboratory - CLPPI, BAS, Sofia, Bulgaria
kivs@bultreebank.org, petya.bultreebank.org

Abstract

The paper presents an HPSG-based annotation scheme for constructing a Bulgarian treebank: *BulTreeBank*. It differs from other grammar-based annotation schemes in having a hybrid status with respect to the partial parsing component and the full parsing module. As the parsing complexity is handled preferably by the pre-processing step, the task of the HPSG module is maximally facilitated and simplified.

1 Introduction

The paper describes an annotation scheme, which mediates between partial analysis of sentences in Bulgarian and their complete linguistic description in HPSG (Head-driven Phrase Structure Grammar) grammar style. This guiding scheme aims at facilitating the initial annotation of a representative set of sentences for the construction of a general HPSG grammar of Bulgarian within the BulTreeBank Project (see (Simov, Popova and Osenova, 2002; Simov et al., 2002)).

The best practices in treebank-design have been facing several problems: the degree of compromise between the linguistic requirements, on one hand, and the implementation possibilities, on the other (Stegmann, Telljohann and Hinrichs 2000), (Brants et al. 2003); the non-homogeneity of linguistic approaches. Some treebank annotation schemes aim at theory-independent interpretation, such as the Spanish Treebank (Torruella and Antonín 2002), the French Treebank (Abeillé et al. 2003). Other rely on specific theoretical background, such as the HPSG-oriented Polish Treebank (Marciniak et al. 2003), the Dependency-based Czech Treebank (Böhmová et al. 2003), the Redwoods HPSG Treebank (Oepen et. al. 2002).

As the BulTreeBank is theory-dependent, our priority is not only an easy transformation to other linguistic theories, but maximal respect of linguistic motivation when treating language phenomena. Thus, we are dealing with the following issues: availability of the lexical resources; theory conformance: especially important is the focus on language specific parameterizations with respect to the sort hierarchy, principles and lexicon; adequate implementation. Our general design criteria are as follows: (1) Maximal consistency between structural mark-up and linguistic theory specifications; (2) Flexibility with respect to information complexity; (3) Optimal linguistic integrity.

The structure of the paper is as follows: the next section presents the goals and the architecture of the annotation. Section 3 focuses on the HPSG language model and processing in BulTreeBank. Section 4 considers the concrete language parameters of the annotation scheme. Section 5 describes its XML representation. The last section outlines the conclusions.

2 Goals and General Architecture of the Annotation

Goals. Our main goal is to create a Bulgarian treebank with detailed syntactic analyses. For that purpose we organized the language data into several levels: ***Text corpus.*** It is envisaged to reach 100 mln. token-size. At the moment it comprises about 70 mln. morphologically processed tokens, out of which 1 mln. tokens are manually disambiguated. The availability of a large corpus is very important, because it is a source for real-text examples. ***Treebank.*** It is envisaged to reach 1 mln. token-size. Analyses are supposed to be deeper that in the corpus. For this level we rely heavily on successfully pre-processed data as an input to the HPSG grammar and then - on semi-automatic post-editing stage. The treebank can serve for developing, testing and evaluating parsers for Bulgarian. ***Core set of sentences.*** It is compiled using two sources: Bulgarian grammar books and corpus. This set is meant to cover the basic linguistic phenomena in Bulgarian. It is manually processed and therefore, it will have a very high degree of reliability.

General Architecture of the Annotation. Parsing with a symbolic grammar is more or less problematic as to the coverage (Dipper, 2000). Therefore it needs the support of a very effective preprocessing module. Adopting such an approach, we move the parsing weight to the preprocessing stage, with the relevant modifications, and leave for the HPSG grammar preferably tasks of attachment. In this way full parsing task becomes fluent and consistent. The annotation process includes the following steps:

Partial parsing step: This step includes all the processing before the application of the HPSG grammar. *Sentence extraction* - from grammar books and the corpus. *Pre-processing*. It includes the following modules: *Morphosyntactic tagging*: defining a tagset and assigning all possible analyses to each word. The morphological analyzer is based on (Popov, Simov and Vidinska 1998) and it is implemented as regular grammars in the CLaRK System ((Simov et. al. 2001)); *Part-of-speech disambiguator*: for each ambiguous word the most probable part-of-speech is predicted ((Simov and Osenova, 2001)). For the core set of sentences it is performed manually and for the entire corpus - automatically; *Partial grammars*: recognition of names, numerical expressions, dates, abbreviations, special tokens (Osenova and Kolkovska 2002; Ivanova and Dojkoff 2002); sentence boundary determination (Ivanova and Dojkoff 2002). *Chunk parsing*: the non-recursive constituents are identified - NPs (Osenova 2002), verb complex (Slavcheva 2002).

HPSG step: The output from the previous step is encoded into an HPSG compatible representation. Then it is sent to an HPSG grammar tool, which takes the partial sentence analyses as an input and evaluates all the attachment possibilities for them. The output is encoded as feature graphs.

Resolution step: Here the output feature graphs from the previous step are further processed in the following way: (1) their intersection is calculated. The intersection exists because all analyses include the partial parsing from the first step and the HPSG grammar tool can not delete information from it; (2) then, on the basis of the differences, a set of constraints over the intersection is introduced; (3) during the actual annotation step, the annotator's task is to extend the intersection to full analysis by adding the missing information. The constraints determine the appropriate extensions and also propagate the information, added by the annotator, in order to minimise the number of the incoming possibilities.

3 HPSG Language Model and Processing in BulTreeBank

As it is clear from the previous section, the core set of sentences and the treebank itself are analyzed within HPSG-style of grammar. In this section we first present the general language model, accepted within HPSG, and then, how we plan to use it for the actual creation of the treebank.

HPSG is a lexicalist linguistic theory, in which the linguistic objects are represented via feature structures. It includes: a linguistic ontology (sort hierarchy) and grammar principles (constraints over the sort hierarchy). The sort hierarchy represents the main types of linguistic objects and their basic characteristics. The principles impose restrictions on the objects and thus predict the well-formed phrases. A basic mechanism for ensuring the right sharing of information among the various parts of the linguistic objects is the *co-reference*. The main linguistic object in HPSG is of sort *sign* (whose subsorts are *word* and *phrase*).

It is a complex entity that is assigned two features: PHON (string of phonemes) and SYNSEM (syntactic and semantic characteristics). Further within the attribute SYNSEM there are three important attributes: CATEGORY (which encodes the syntactic information), CONTENT (which encodes the semantic information) and CONTEXT (which encodes the pragmatic information). The constituent structure is encoded for each phrase via the attribute DTRS. Assigning different values to this feature, HPSG theory distinguishes between (at least) the following type of phrases – *headed-phrase* and *non-headed-phrase*. The first kind is additionally divided into *head-complement*, *head-subject*, *head-adjunct*, and *head-filler*. The current hierarchy of phrases is presented in the following sort hierarchy:

```
sign
    PHON : phonlist
    SYNSEM : synsem
  word
  phrase
      DTRS : dtrs
    headed-phrase
      head-complement
      head-subject
      head-adjunct
        head-sem-adjunct
        head-pragmatic-adjunct
      head-filler
    non-headed-phrase
```

The distinction between *head-sem-adjunct* and *head-pragmatic-adjunct* is on the basis of whether the given adjunct modifies the semantics of the head or its pragmatic nature only. An example of a pragmatic adjunct are the vocative phrases in Bulgarian (see (Osenova and Simov 2002)). The *head-filler* phrases account for the cases of unbounded dependency. The *non-headed-phrase* is used for dealing with coordination phrases.

The linearization of the constituents in HPSG is separated from the constituent structure and in this way the theory allows for different orders of the same constituent structure and discontinuous realization of the constituents. This separation ensures the representation of the grammatical relations within the constituent structure. The actual realization of the head dependents is governed by a set of immediate dominance schemata. The realization of the dependents follows the sequence: **complements** → **subject** → **adjuncts**. The actual number and kind of dependents is determined by lexical elements within each phrase.

The structure of the linguistic objects in HPSG makes its language model very appropriate for encoding the information in a treebank. In fact, we could consider it as a hybrid approach to representation of syntactic information because it represents the constituent structures and grammatical relations at the same time. This flexibility is at the cost of the complexity of the representation and processing. The complexity in our view stems in the processing of lexical signs. The usual mechanisms for the treatment of different lexical alternations and analytical word form are encoded as lexical rules and/or techniques like argument composition which are very complicated. In our project the problem is even more serious because of the following factors:

- There is no appropriate lexicon which can be used as an HPSG lexicon for Bulgarian with wide coverage.
- The grammar has to cover wide range of texts.
- Although the grammar will overgenerate it has to produce all the linguistically relevant analyses.

An unfortunate fact is that we do not have resources under the project in order to implement such an HPSG grammar and lexicon. Thus we have to minimize the complexity with other means external to the HPSG grammar. Following our annotation architecture presented in the previous section, we rely on the preprocessing steps to deal with the idiosyncratic information about the lexical items and lexicalized phrases like idioms[1]. Thus the input for parsing by the HPSG grammar is not a list of phonemes as in the usual parsing task, but a list of signs corresponding to the chunks from the partial preprocessing. Although the parsing system has only to complete the partial structures to a complete structures, the signs from the

[1] Also the preprocessing step covers some of the completely compositional phrases as it was mentioned above.

partial analyses have to be checked for consistency against the HPSG grammar. This fact causes the following problems:

- Some constructions that can be recognized by the partial grammars do not have appropriate treatment in HPSG. Such constructions are, for example, some kinds of idioms, date expressions, parentheticals.
- The complexity of checking over the compositional phrases is still high.

In order to overcome this problem we encode each sign from the partial analysis of the sentence as a lexical sign (the sort *word*) with the appropriate characteristics based on its structure in the partial analysis. Additionally, from the partial analysis of the sentence we delete all parenthetical expressions, which are treated as pragmatic adjuncts to the whole sentence. Then the input to the HPSG grammar is a list of "words" instead of list of signs of arbitrary complexity. After parsing the modified input, the substituted phrases are incorporated back in the sentence processing. One major problem for this approach is the possible discontinuity within some semi-idiomatic phrases. Such kind of discontinuity is recognizable by the partial grammars, but for the moment it is handled by the annotators.

In the next sections we present an annotation scheme for manual annotation of the core set of sentences. It is designed to require minimum information from the annotators. This annotation scheme reflects the model and processing described in this section.

4 Defining the Linguistic Parameters of the Annotation Scheme

The classification of the linguistic phenomena that we would like to explicate in our treebank is based on several sources - (Marciniak et al. 2003) and the citations there. Also we have added some phenomena and features missing in the mentioned sources and some specific ones for Bulgarian. Additionally, we introduced some changes in the classification scheme based on the HPSG definition of linguistic objects. Generally speaking, we rely on two basic assumptions:

1. We use a 'domain-phenomena' cross-classification, where the main syntactic domains are defined and the phenomena, which occur there, are analyzed.
2. We analyze the data according to the following HPSG-oriented criteria: the type of the sign (word, phrase), headedness (heads or non-heads), the typology of words and phrases, the saturation condition (saturated vs. non-saturated items).

4.1 Core domains of the phenomena realizations

In our scheme we use the standard phrasal categories in a non-standard way. To put it more precisely, the phrasal categories are used as a macros or an interface to more detailed and linguistically different information. They mean more that just constituents. Let us consider them for clarity. The basic domains are:

NP. NPs are classified with respect to different criteria such as the number of its internal elements (heaviness), the tendency of the elements to be closer or not to the head (we call it 'nearness'), specific features (being names, numerical expressions, abbreviations, pragmatic constructions etc), additional properties like recursivity (recursive vs. non-recursive NPs), coordination (coordinated vs. non-coordinated NPs), ellipsis (elliptical vs. non-elliptical NPs), substantivization (substantivized vs. non-substantivized). Note that most of the mentioned criteria interleave. These concern mainly the automatic processing of NPs than their HPSG analysis and thus supply information about the levels of processing.

We can assume that bare Bulgarian NPs are always functionally complete. Hence, the noun on its own can be considered a phrasal projection, which does not require any SPRs (specifiers) on its Valency list[2]. Thus it is considered lexical and marked with the lexical category N. Lexical are considered substantivized elements as well as nouns, modified by possessive clitics . It is in accordance with the general specification that clitics do not change the lexical nature of the sign. Other NPs are divided into the

[2] See also in (Butt et.al. 1999, p.102)

following dependency structures: head-adjunct (NPA), head-complement (NPC) and non-headed (CoordP). Head-complement structure is reserved for NN groups of type ‘container-content’ and ‘type of assembling-entities’. Head-adjunct relation covers all the other NPs and non-headed type handles the coordinated ones.

AP. Adjective on its own or combined with a possessive clitic is a lexical sign (A). In the latter case the head is definite and ’glues’ the clitic. We assume that some adjectives have complements and the participles inherit the argument structure of the verb. Thus we allow head-complement (APC) and head-adjunct structures (APA). The APA structures cover both - adjunct and subject dependents of the underlying verb.

AdvP. If the adverb is non-modified, then it is marked lexically (Adv). If there is a modifier, the whole structure is considered a head-adjunct type (AdvPA).

VP. VPs can be either lexical, or phrasal. The lexical one is marked V and includes bare verbs, verbs with clitics, da-constructions (inheritors of the infinitive), analytical verb forms, elliptical verbs. The phrasal category is recursive: first, the verb with its full-fledged complement(s) forms a head-complement structure (VPC). Then, the head-complement VP takes the subject and forms a head-subject phrase (VPS). If there are adjuncts, they are attached last and form VPA projections. When the extracted element does not have structural parent, we assign to it a head-filler phrase (VPF).

Verb valency frames are automatically generated from the morphological dictionary and thus some of them remain underspecified. As a filter we use the machine readable valency dictionary of Bulgarian, which specifies the relevant frame for the verb. It covers 1000 verbs. We do not view complementation as one-to-one relation to transitivity, thus allowing VPC nodes over intransitive verbs as well.

CL. When saturated, VP equals a sentence (S) or a clause (CL). The sentences can be simple, complex, coordinated. With respect to the illocutionary force they are: declarative, interrogative, imperative, optative. Different subtypes according to the marker and/or matrix verb frame within CL are distinguished. For example, a clause subtype for relatives (CLR), a clause subtype for da-constructions (CLDA) etc.

PP. It forms a head-complement structure with the following phrase being NP, AP or AdvP.

CoordP. It handles all types of coordinates like NPs, PPs, APs, AdvPs, VPs. Thus, on one hand, the exact type is recoverable compositionally from the elements and, on the other hand, the underspecificaton solves some problematic cases, where the so called ‘non-constutuent’ coordination appears.

XP. It covers all the cases that are not covered by other phrases.

One important question that arises here is the connection of the core domains with the word order and inevitably - with the concept of discontinuity. As it was mentioned in the introduction, we do not accept the idea of the canonical word order and therefore we do not change it. At this annotation level we accept crossing branches in order not to destroy the dominance relation. It means that we first indicate the maximal constituent and then - the place of the non-belonging elements. In our view there exist three kinds of structural discontinuity. All of them are determined by the head:

Head dependants permutation. In this case all dependants of the head are hierarchically ordered in several levels of complexity and it is accepted that each level in the hierarchy is realized continuously in the linear order. Discontinuity appears when a constituent from an upper level of the hierarchy is realized between constituents of a lower level. Is is labelled DiscA. A typical example is when the subject is realized between the verb and the complements, or the adjunct is realized between the verb and the complements.

Mixture of two saturated constituents. This is the case when the constituents of two saturated phrases are mixed with each other. For example, the constituents of two NPs, or two clauses. The elements are labelled DiscM. In this case the points of insertion of the outside constituents are determined by the head of each saturated phrase on the base of *the uniqueness of the interpretation*. The uniqueness of the interpretation means that the outside constituents can not be interpreted as a part of the surrounding phrase and thus can be easily recognized as such.

External realization of an inner constituent. This is the case referred to generally as *extraction*. The element is labelled DiscE. Here we have in mind phenomena like *topicalization*.

4.2 Core phenomena

In our opinion the core phenomena are not just a list, but they can be classified into several groups according to some specific features. Of course, we are aware of the fact, that strict boundaries cannot be set between the groups.

Unexpressed elements. The members of this group depend more or less on world or grammar knowledge, on discourse-based information.

Pro-dropness. Syntactically it is referential and non-referential (dummy, expletive), but pragmatically the types increase (Osenova (to appear)). The pro-dropness is explicitly marked only when it is coreferentially bound within the sentence.

Ellipsis. It is preferably a context-bound phenomenon. We handle the locally recoverable ellipsis, being either head or dependant.

Frame alternation. It includes: passivization and non-overt-realization of the arguments.

Co-referential relations. Here we include language phenomena treated by the structure-sharing (co-reference) mechanisms in HPSG.

Agreement. NP-internal agreement, Subject-Verb agreement, doubling categories as agreement markers. Idiosyncracies are handled as well (Osenova (to appear)).

Binding. It is parameterized with respect to subject oriented binding.

Anaphora resolution of different kinds (different from binding). When analyzing unrestricted texts, it is very important to have information whether the referents within the sentence are specific enough, or they need resolution. The relevant status of the referent is encoded within the feature `CONTEXT|INDICES`. We have two additional features: incoming-indices and outcoming indices. When the referent is specific enough (name), it contributes to other sentences by outgoing-indices. When the referent is not specifics (pronouns) and they are not co-referred within the sentence, they are connected with the incoming-indices.

Definiteness. On one hand, it is a morphological feature of the nominals in the lexicon. On the other hand, it is a phrasal feature and morphologically can be realized only once within the phrase.

Control. Raising and equi verbs. It is important to indicate the control verbs and patterns for a pro-drop language like Bulgarian.

Relative clauses. They are either considered adjuncts within NPs or interpreted with respect to the whole predicate.

Secondary predication. We consider it a separate kind of adjunct with co-reference specification with the subject or object.

Type-shifting. It covers basically two phenomena, in which there is type-shifting from nominal dependants to heads, or from non-nominal elements to nominals.

Substantivization. It is preferably lexicon-based and shifts the usual nominal dependants to heads.

Nominalization. It is syntactically based and shfts predicates, intrejection etc. to nominals.

5 XML Implementation of the Annotation Scheme

We have implemented the annotation scheme as an XML DTD and a set of constraints over XML documents in CLaRK System. The DTD defines the general structures of the documents representing the sentence analysis; the constraints are used in two modes (according to the general use of constraints in CLaRK System): (1) to support the annotators during their work; (2) to validate the result of the annotators' work. There are two basic principles accepted during the DTD design: (1) the XML tree model is used to represent as much as possible from the structure of the sentence analysis; (2) the order of lexical elements corresponds to the word order in the sentence and no empty elements are inserted in the structure. Following these principles, in the DTD we defined the following elements, which correspond to the linguistic domains as described above: `NP`, `VP`, `AP`, and etc (for phrases) and `w` (for words). We call each of these elements – *structural element*. Each phrase element can have as children structural elements which define the immediate dominance relation in HPSG. Additionally, the HPSG features and their values for each structural element are represented in two more specific ways: (1) Separation between lexical (N, V, A etc.) and phrasal ones (NPA, VPS, APA etc.) If there is discontinuity, the corresponding elements are marked-up as external elements (DiscA, DiscM or DiscE) and they are additionally assigned some value of the `idref` attribute at the element. This value coincides with the `idref` value of the non-immediate-dominance element(`nid`) under the dominating phrase. Then the structure-sharing relation is stored at a suprasentential level within `CoIndex` tag; (2) As a set of XML attributes (the grammatical characteristics of the phrase, for example).

The most important information encoded for each structural element is the type of phrase (lexical for words). Following the HPSG sort hierarchy we considered the following levels of description:

- **Lexical.** On this level the Morpho-Syntactic characteristics of the individual words are represented together with their combinational potential and semantic properties.

- **Phrase.** The elements represented here correspond to the truly compositional elements of syntax, semantics and pragmatics of individual utterances in the language. They obey the set of principles of the grammar. The hierarchy of the different kinds of phrases is given above.

Following the practical goals of the description in the treebank and some of the latest developments in HPSG, we additionally envisage the following levels:

- **Multi Lexical.** The level is an extension of the lexicon into two main directions. First, here we present all the multiword lexical forms produced by lexical rules. These include mainly analytical verb forms. The second direction is towards representation of idiosyncratic expressions of different kinds – idioms, collocations. For each kind of multiword expression a different marker for the structural elements is defined in the DTD.

- **Discourse.** This is the level of the linguistic elements bigger that a single utterance. It is supported by the above levels. In our work we envisage this level to be used for a representation of the multisentential co-reference and anaphora resolution.

In our treebank not all of these levels will be generated by an HPSG grammar because such a grammar does not exist. Thus some of the levels are treated completely by the preprocessing steps from the above annotation architecture or as post-processing.

Additionally to the definitions in the DTD, we implemented a set of constraints which reflects the possible combinations of values defined in the DTD. For example, we encode that a structural element marked-up as a head-complement VP (VPC) cannot immediately dominate a structural element marked-up as a head-adjunct non-saturated VP (VPA).

The manual annotation is done within the CLaRK System. The annotator has access to two different views of the XML document – a tree view and a textual view. Additional textual views can be defined if necessary. In each view the annotator has the possibility to define which parts of the XML document to be presented and in which colour. When entering the information, the annotator is supported by the constraints, stated in the system and the DTD.

6 Conclusion

The presented annotation scheme is designed to support the building of the core set of sentences, which will constitute not only the golden standard for the HPSG grammar and the chunk grammars, but will serve as a guideline for the annotators. Because of the intended multifunctionallity of this level, the annotation scheme combines pure linguistic features with metafeatures like recursivity, heaviness, specificity (important for chunking). The scheme is robust and annotator-friendly, because: (1) via XML mechanisms it restricts the annotator in making decisions, thus reducing human-driven errors, and (2) it minimizes the cases for manual intervention.

Acknowledgements

The work reported here is done within the BulTreeBank project. The project 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.

References

Anne Abejllé, Lionel Clément and Francois Toussenel. 2003. *Building a Treebank for French.*. In: Anne Abeillé (editor). *Treebanks. Building and Using Parsed Corpora.* Kluwer Academic Publishers. pp 165-187.

Alena Böhmová, Jan Hajič, Eva Hajičová and Barbora Hladká. 2003. *The Prague Dependency Treebank.* In: Anne Abeillé (editor). *Treebanks. Building and Using Parsed Corpora.* Kluwer Academic Publishers. pp 103-127.

Thorsten Brants, Wojciech Skut and Hans Uszkoreit. 2003. *Syntactic Annotation of A German Newspaper Corpus*. In: Anne Abeillé (editor). *Treebanks. Building and Using Parsed Corpora.* Kluwer Academic Publishers. pp 73-87.

Miriam Butt, Tracy Holloway King, María-Eugenia Nino and Fréderique Segond. 1999. *A Grammar Writer’s Cookbook.* CSLI Publications.

Stefanie Dipper. 2000. *Grammar-based Corpus Annotation*. In *Proceedings of the Workshop on Linguistically Interpreted Corpora*, Luxembourg.

Krassimira Ivanova and Dimitar Doikoff. *Cascaded Regular Grammars and Constraints over Morphologically Annotated Data for Ambiguity Resolution.* In: *Proc. of The Workshop on Treebanks and Linguistic Theories (TLT2002)*. Sozopol, Bulgaria. pp 96-113.

Margorzata Marciniak, Agnieszka Mykowiecka, Adam Przepiórkowski, Anna Kupść. 2003. *An HPSG-Annotated Test Suite for Polish*. In: Anne Abeillé (editor). *Treebanks. Building and Using Parsed Corpora.* Kluwer Academic Publishers. pp 129-146.

Stephan Oepen, Ezra Callahan, Dan Flickinger and Christopher D. Manning. 2002. *LinGO Redwoods. A Rich and Dynamic Treebank for HPSG,* In: *Proc. of The Workshop Beyond PARSEVAL. The Third LREC Conference*. Las Palmas, Spain.

Petya Osenova. *On Subject-Verb agreement in Bulgarian (An HPSG-based account).* In: *Proc. of the Fourth Formal Description of Slavic Languages Conference 2001*. Potsdam, Germany.

Petya Osenova. 2002. *Bulgarian Nominal Chunks and Mapping Strategies for Deeper Syntactic Analyses.* In: *Proc. of The Workshop on Treebanks and Linguistic Theories (TLT2002)*. Sozopol, Bulgaria.

Petya Osenova and Sia Kolkovska. *Combining the named-entity recognition task and NP chunking strategy for robust pre-processing.* In: *Proc. of The Workshop on Treebanks and Linguistic Theories (TLT2002)*. Sozopol, Bulgaria. pp 167-182.

Petya Osenova and Kiril Simov. 2002. *Bulgarian Vocative within HPSG framework.* In: *Proc. of the 9th International Conference on Head-Driven Phrase Structure Grammar (HPSG)*. Kyung Hee University, Seoul, South Korea. August 8-9. pp 94-100.

Dimitar Popov, Kiril Simov and Svetlomira Vidinska. 1998. *A Dictionary of Writing, Pronunciation and Punctuation of Bulgarian Language*. Atlantis LK, Sofia, Bulgaria. In Bulgarian.

Kiril Simov, Gergana Popova and Petya Osenova. 2002. *HPSG-based syntactic treebank of Bulgarian (BulTreeBank).* In: *“A Rainbow of Corpora: Corpus Linguistics and the Languages of the World”*, edited by Andrew Wilson, Paul Rayson, and Tony McEnery; Lincom-Europa, Munich, pp. 135-142.

Kiril Simov, Petya Osenova, Milena Slavcheva, Sia Kolkovska, Elisaveta Balabanova, Dimitar Doikoff, Krasimira Ivanova, Alexander Simov, Milen Kouylekov. 2002. *Building a Linguistically Interpreted Corpus of Bulgarian: the BulTreeBank.* In: Proceedings from the LREC conference, Canary Islands, Spain.

Kiril Simov and Petya Osenova. 2001. *A Hybrid System for MorphoSyntactic Disambiguation in Bulgarian.* In: Proc. of the RANLP 2001 Conference, Tzigov chark, Bulgaria, 5-7 Sept., pp. 288-290.

Kiril Simov, Zdravko Peev, Milen Kouylekov, Alexander Simov, Marin Dimitrov and Atanas Kiryakov. 2001. *CLaRK - an XML-based System for Corpora Development.* In: Proc. of the Corpus Linguistics 2001 Conference. Lancaster, England. pp: 558-560.

Milena Slavcheva. 2002. *Segmentation Layers in the Group of the Predicate: a Case Study of Bulgarian within the BulTreeBank Framework.* In: *Proc. of The Workshop on Treebanks and Linguistic Theories (TLT2002)*. Sozopol, Bulgaria. pp 199-210.

Rosmary Stegmann, Heike Telljohann and Erhard Hinrichs. 2000. *Stylebook for the German Treebank in VERBMOBIL.* Report 239. SfS, Universitat Tubingen.

Monserat Civit Torruella and M. Antonín Martí Antonín. 2002. *Design Principles for a Spanish Treebank.* In: *Proc. of The Workshop on Treebanks and Linguistic Theories (TLT2002)*. Sozopol, Bulgaria. pp 61-77.