Using the linguistic knowledge in BulTreeBank for the selection of the correct parses

Petya Osenova and Kiril Simov
Linguistic Modelling Laboratory
Bulgarian Academy of Sciences
petya@bultreebank.org, kivs@bultreebank.org

Abstract

In this paper, a method is presented for transferring of linguistic knowledge between two treebanks of Bulgarian, constructed within the same linguistic theory, but in its different versions and from different perspectives. BulTreeBank (BTB) follows HPSG94 and the sentences have been analyzed per se. The target Treebank BURGER is constructed by an HPSG grammar for Bulgarian. The linguistic information in BTB and BURGER is presented in the format of lexical categories and dependency relations. A set of transferring rules on the level of the categories (or list of categories) is defined to ensure the compatibility of the representations. Currently our goal is to provide a mechanism for the usage of the linguistic knowledge encoded in BTB as a set of discriminating properties for the selection of the correct analyses produced by BURGER.

1 Introduction

Any annotation effort over some language resource would take into account the usability of the annotated resource. The question of which treebank annotation is better has been discussed in many works – see for example Kübler *et al*. (2008). In the current project, we aim at a treebank which to support a Bulgarian-English HPSG-based statistical machine translation. For this task, a parallel treebank is needed, which to meet at least the following requirements: the analyses in both languages to be comparable; and the size to allow estimation of parameters for correspondences on different levels of linguistic analyses. In our case, the first requirement is ensured by sharing as many categories and principles in the analyses of both languages as possible. The second requirement imposes the usage of automatic methods in the creation of the treebank. Thus, we have to use two parsers – one for English and one for Bulgarian – which produce similar analyses with respect to common HPSG principles. For English we envisage to use the English Resource Grammar (ERG) (Flickinger 2000) and for Bulgarian we are developing a resource grammar based on the same principles.

This paper presents our first experiments on transferring of the linguistic knowledge between two HPSG-oriented resources of Bulgarian with the aim to disambiguate the analyses in one of them. The first resource is the

Bulgarian HPSG-based Treebank – BulTreeBank – (Simov et al. 2004), and the second one is another HPSG-based treebank under construction on the base of the Bulgarian Resource Grammar (BURGER). BulTreeBank (BTB) is based on an annotation schema, designed with respect to HPSG94 (Pollard and Sag 1994). It was semi-automatically constructed by using partial parsers. The full analyses had been completed manually in an XML format. The annotators had at disposal a reporting service, which prompted the places of errors when their decisions did not conform to the specified nodes and attributes in the DTD. The final analyses have been checked by two people. We consider this source reliable with respect to the following annotation levels: morpho-syntactic level (manually annotated), constituent level (partially automatically annotated, manually checked and completed), dependency relations – within each constituent the head-dependent relations have been annotated. Additionally, named entity annotation, co-reference annotation (relations – *equality*, *member-of*, *subset-of*), annotation of ellipses have been provided. BURGER (Osenova 2010) is an HPSG grammar under implementation by customizing the Matrix grammar. A Treebank to be constructed on the base of BURGER would be a treebank in which the correct analyses produced by BURGER are selected and stored. In this we follow the Redwood approach (Oepen et al. 2002a, 2002b). Here we investigate the way in which the two resources can be used in order to construct the BURGER Treebank via transfer of knowledge from BTB.

The result of the construction of BURGER Treebank will be used for the construction of a parallel Bulgarian-English treebank. The main usage of such a treebank is the implementation of machine translation system between the two languages. The steps of the BURGER Treebank annotation include:

- Selection of parallel sentences from a given aligned parallel corpus;
- HPSG analysis of corresponding sentences;
- Establishing of correspondences between the HPSG analyses.

The first step is relatively easy as much as already there are reasonably sized Bulgarian-English parallel corpora. The third step requires the analyses of Bulgarian sentences and the analyses of the English sentences to be comparable. In order to achieve this, the sentence analyses have to be modelled in the same way for both languages (Step 2). Our approach is based on the usage of the same grammar formalism – HPSG as implemented within Matrix grammar, thus, we will have similar grammars – ERG for English and BURGER for Bulgarian, implemented in the same grammar development environment – Linguistic Knowledge Builder (LKB). On the other hand, a construction of a grammar with wide coverage is a long term project which we cannot achieve within our current project. Thus, we need to reuse as much as possible from the already available resources for both languages. In this paper, we report a case study of the possibility to transfer the linguistic knowledge which is already incorporated within BulTreeBank in order support the creation of the BURGER treebank and its related grammar.

The structure of the paper is as follows: in the next section a brief overview on the related works is provided; then a comparison is presented

between the annotation schemas behind BTB and BURGER; in Section 4 the linguistic analyses of both approaches are discussed; Section 5 comments on the transfer of the linguistic knowledge for selecting the good analyses; Section 6 presents the implementation of the transferring rules; the last section concludes the paper and gives some directions for future research.

2 Related Works

There are various approaches to transferring of knowledge from a treebank with respect to a specific task. Some works focused on converting existing constituent-based treebanks into dependency format (Daum et al. 2004), or from one linguistic theory into another theory (Hockenmaier 2006) and many others. The treebanks are used also for extracting lexical types and items for supporting a hand-crafted grammar (Cramer and Zhang 2009). Our task is to use the information in a treebank to select correct analyses produced by a parser. Having started the development of BURGER grammar and the related BURGER Treebank, we will gain from all the components of the developed infrastructure, such as grammar developing workbench (LKB), parsing environment (LKB and PET), profiling software ([incr tsdb()]), etc. The important idea to us is the mechanism behind the development of the Redwoods treebank. This treebank was compiled by coupling ERG and a tree selection module of [incr tsdb()] (Oepen et al. 2002b and Oepen and Callmeier 2000). ERG produces very detailed syntacto-semantic analyses of the input sentence. For many sentences, LKB overgenerates, producing analyses that are not acceptable. From the complete analyses different components can be extracted in order to highlight different views over the analyses: (1) derivation trees composed of identifiers of lexical items and constructions used to build the analysis; (2) phrase structure trees; and (3) underspecified MRS representations. From these types of information the most important with respect to the treebank construction is the first one, because it is good enough to support the reconstruction of the HPSG analysis by a parser. The steps of constructing the Redwood treebank are:

- LKB produces all possible analyses according to the current version of ERG;
- The tree comparison module provides a mechanism for selection of the correct analyses;
- The selection is done via basic properties (called also discriminating properties) which discriminate between the different analyses;
- The set of the selected basic properties are stored in the treebank database for later use in case of treebank update.

In our work, we take all the LKB analyses of the Bulgarian sentences produced by the current version of BURGER. Then we discriminate on the derivation trees because the information there is enough for the full analyses to be determined. Also the derivation trees are used in Redwood treebank setting to define the basic properties. As it was mentioned above, our idea is to use the analyses in BTB to extract the necessary discriminating properties.

The main difference from the manual selection of the correct analyses in our settings is that we can use the total linguistic knowledge, represented in BTB, instead of a predetermined list of discriminating properties. There exist two related tasks: (1) how to extract the discriminating properties from BTB, and (2) how to map them to BURGER analyses. Hence, some ideas have been used from the area of transformation of treebanks and transfer of linguistic knowledge. Our work is based on (Simov 2004), (Chanev et al. 2007) and others. A discriminating property has to be easy to determine within the BURGER analyses and easy to extract from BulTreeBank. In order to facilitate this, we use the ideas from the above mentioned works for transforming the treebank in a new format that would allow a better comparison. Ideally, the transformation has to be done on the knowledge level only, without references to actual implementation formats.

3 Annotation schemata: BTB vs. BURGER Treebank

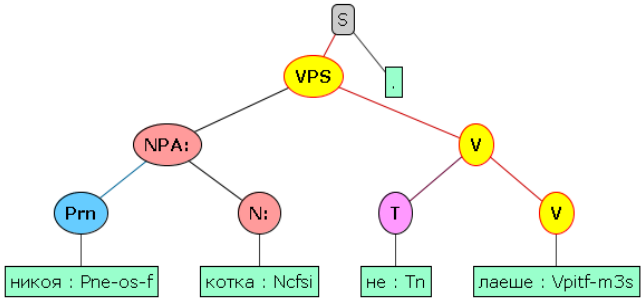
The annotation schema behind BTB (Osenova and Simov 2007) generally follows the HPSG94 linguistic model. It incorporates the universal principles, such as Head Feature Principle, Valence principle, etc. In addition, it follows the hierarchical approach when attaching dependents to their heads. First, the complements are attached, then the subject being an external argument, and finally – the adjuncts. It should be noted that the complements are attached together, by one operation only. Additionally, in BTB the constituent structure is separated from the word order. It means that the topic-focus layer is not distinguished. In such a paradigm, crossing branches are allowed, and three types of discontinuity are envisaged (scrambling, topicalization and mixed). The implementation is in XML, where the XML tree structure is exploited to represent the constituent structure as much as possible with encoding of crossing branches via ID and IDREF attributes. The visualization takes the form of the XML tree and represent it as close as possible to the canonical syntactic trees. The dependency relations are encoded into the syntactic labels. For example, VPC means verbal phrase with a complement.

Apart from the phrase level, another level has been introduced – functional. It handles the various types of clauses (CLR, CLDA, CLQ, and CL), coordination, co-referenced pro-dropness, etc. BTB takes into account the types of named entities (*person*, *organization*, *location* and *other*), various co-references within the sentence as well as the ellipses.

The layers in BTB are modelled separately. Morphological analyses come first. The ambiguous ones have been disambiguated manually. Then chunks have been analyzed, and finally – full analyses with handling the specific attachments, discontinuities and cases of ellipsis. Non-local dependences are handled by the discontinuity markers only.

BTB introduces phrase structures and dependency relations, but lacks feature structures as well as a separate semantic layer of representation. The semantics can be derived as follows: the predicate structure via the dependency labels (arity) and co-references (control, pro-dropness); the

relations – via the functional labels (nominalizations, subordinate clauses among others) and co-references (possession). The scope of quantification is present only in the selected interpretation by the annotator. Additionally, the analysis of names shows the semantically correct analysis with respect to subject and complement selection.

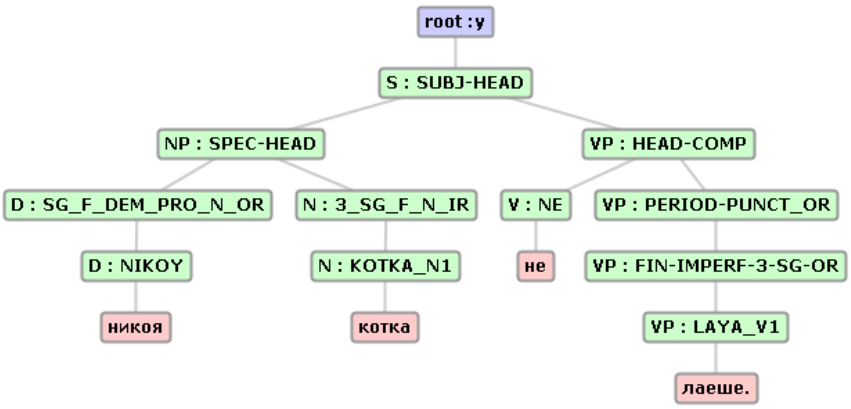


In the above picture the sentence (1) is presented:

(1) Никоя котка не лаеше.
Nobody cat no was-barking.
No cat was barking.

The determiner 'nobody' is viewed as an adjunct within the NPA. The phrase is also a subject to an intransitive verb.

The annotation schema of BURGER Treebank strictly follows the principles behind the BURGER grammar. Therefore, it is in accordance with Matrix grammar and other Matrix-based grammars viewed as best practices. In contrast to BTB, where the annotator had to decide on the correct analysis/analyses according to his/her knowledge using only partial analyses, in BURGER the most appropriate analysis/analyses have to be selected among the all produced by the grammar ones. I.e. the annotator is faced with multiple analyses before his/her selection. BURGER aims at combining all the linguistic levels – morphology, syntax and semantics. At the moment, the morphological module produces analyses which are not disambiguated. The syntactic one produces all the possible structures, including topicalizations where appropriate. The semantic module, which is encoded in MRS, gives information about the various relations, predicate structure, control, etc. Syntactically, BURGER introduces a more relaxed schema. It tolerates various types of attachments since it follows the assumption that all possible syntactic structures are allowed, and later on the best one will be chosen via some appropriate mechanism (statistical one or comparison against a gold standard or another). This presupposed freedom has two dimensions: – spurious-like ambiguity, such as adverb attachment to both – VP and S nodes; and non-fixed attachment of arguments. For example, subject might be attached to the head after the adjunct had been attached; or adjuncts might be attached to the head before the complements. In this way, no crossing branches are allowed. Also, each dependent is attached to its head one by one, irrespectively of being a complement or an adjunct. The next picture presents a tree of the BURGER counterpart of the sentence (1).



4 Comparing of the gold linguistic analyses

In order to make the comparison between the two types of annotation possible, a case study was performed. At the moment, BURGER covers the Matrix testset (Bender et al. 2002) with a slight extension. The set comprises 194 grammatical sentences with one or more analyses. First, the correct analysis/analyses was/were selected manually for each sentence. Then the same set was manually annotated with respect to the BTB annotation scheme.

The phenomena that are demonstrated by the testset are as follows: various predicate constructions (intransitive, transitive), control, modification, quantification, illocutionary force (questions, imperatives), clauses (relatives and reduced relatives, if-clauses, that-clauses), modals, negation, copula constructions, hybrid categories (deverbals, gerunds), light constructions, coordination, nominalization, quantification. The typical phenomena in Bulgarian include: clitic doubling, pro-dropness, double negation, some basic verb clusters (da-constructions or future tense without or with clitics), clitics in NPs.

In BURGER, from 654 analyses, 81 analyses are unique. For the rest, 277 have been chosen as good and 27 as possible, but rare. Altogether 348 analyses have been rejected, which makes more than 50 % of the produced ones. This result proves that a mechanism for disambiguation is needed (as expected). In the BTB version there are 207 analyses. From them only 13 cases have 2 analyses. They are mainly cases of topicalization readings. Only 3 cases give attachment varieties.

Concerning the syntactic modeling of the specific phenomena, there are several differing but comparable interpretations in both gold datasets. For example, in BTB phrases like ‘every cat’ or ‘some cat’ are analyzed as NPs, while in BURGER they are analyzed as head-specifier phrases. However, the head is still the noun. BTB distinguishes among pragmatic and other adjuncts. BURGER makes a distinction among intersective and scopal adjuncts. In BTB the subordinators and complementizers are viewed as markers. The projections are therefore functional labels. Then, they either are selected as complements, or they modify phrases. In BURGER both types of

linking words take the introduced clauses as their complements. The coordination in BTB is analyzed in a flat way, i.e. as a non-headed phrase. In BURGER it is analyzed in levels (bottom, middle and top), and is considered a headed phrase. The question polar particles project lexical nodes in BTB, while phrasal ones of the type head-interesective modifier in BURGER. In BTB, the clitics project the lexical label of their heads, while in BURGER they undergo special head-clitic rules. The negative particle in BTB also projects the lexical label of its head. However, in BURGER it is treated as a verb, which takes a complement. To sum up, most of the analyses of both schemas are comparable, but there is also a need of formulating transferring rules, which to ensure the correct mapping.

5 Transfer of Linguistic Knowledge and Disambiguation

Although lacking a semantic layer, BTB analyses have semantically-oriented elements: dependency relations, named entities, co-references, hierarchical constituent structure. The syntactic structure transition seems more trivial as much as the analyses go into the same direction with only slight differences.
The main sources of ambiguity and multiple analyses are as follows: (1) morphological ambiguity, (2) various places of attachment, (3) neutral vs. focused ordering of constituents, (4) proliferation of several competing rules for the same item. They act separately or in various combinations. The more combinations among them, the more analyses appear as results.

Let us comment in more detail on the above sources of ambiguity. The first case produces all morphosyntactic possibilities. For example, in sentence (2) the verb is ambiguous between present and aorist tense of the perfective verb ‘give’. However, present tense is not grammatical:

(2) Абрамс **даде** цигара на Браун.
Abrahms **gives/gave** cigarette to Brown.
Abrahms is **giving/gave** a cigarette to Brown.

BTB analysis in this case would be only one – with the aorist tense. This information is used for 100 % of disambiguation during the transfer of linguistic knowledge to BURGER Treebank.

Case 2 from the above list produces all possible attachments irrespectively to the meaning of the sentence. For sentence (2) there are incorrect analyses, which attach the PP ‘to Brown’ to the noun ‘cigarette’ besides the correct ones, in which the PP ‘to Brown’ is attached to the verb as its second complement. Another example is sentence (3):

(3) Котката е **в** **градината**.
Cat-the is in garden-the
The cat is in the garden.

The PP is attached not only as a complement to the copula (as expected), but also as a modifier to a verb-complement phrase (as rejected in this case). In BTB there is only one analysis, namely the one with the complement. Thus, again – 100 % of such cases are disambiguated.

Source 3 introduces sentences with topicalized constituents. Let us consider sentence (4):

(4) **Онова** **куче** преследваше **Браун**.
That dog was-chasing Brown.
That dog was chasing Brown.

The canonical reading is the one in which the dog is the chaser and Brown is the chased one. In the topicalized version, it is vice versa. In BTB there are these both analyses presented with the preference to the first one.

The fourth case is triggered when the same item can undergo more than one rule. For example, for sentence (2) analyses are generated with the semantically empty preposition 'на' (dative) as well as with the modifying preposition 'на'. The latter analyses have to be rejected in this reading.

6 Implementation

As it was discussed earlier in the paper, our mechanism for transferring of linguistic knowledge from BulTreeBank to BURGER Treebank is based on the ideas of the treebank transformation. We decided to use a common target format to represent the important knowledge from both treebanks. The procedure is as follows:

- The analyses from BURGER are transformed into a new format;
- The corresponding analyses from BulTreeBank are also transformed into the new format;
- The knowledge within the new representations is unified on the basis of correspondences rules;
- The parse selection is done on the basis of comparing both unified representations.

In order to facilitate the comparison, we decided to use as a common format a dependency-like representation as follows. Each sentence is represented as a list of wordforms:

w1 w2 w3 … wn

This representation is necessary in order to keep track of word forms used in the lexical and head-dependent descriptions and their word order.

Lexical elements:

wk:pos list-of-categories

In this part of the representation, all the lexical categories that dominate the wordform in the representation of the corresponding treebank are stored. We assume that this list describes the lexical features of the wordform. Also, the position of the wordform in the sentence is stored.

Head-dependent pair:

<wi:posi, wj:posj> list-of-categories

For any two wordforms in the sentence where one of them is a lexical head of the other, the category of the minimal path between the two wordforms is stored. Sometimes we need to include additional information from the unary branches in order to have all the relevant information represented. Here is an example from the testset:

(5) Кога лаеше кучето
when barked dog-the
When did the dog bark

BulTreeBank case:

(**VPA** (**Adv** (**Pit** кога)) (**VPS** (**V** (**Vpitf-m3s** лаеше)) (**N** (**Ncnsd** кучето))))

Lexical elements:

кога:1 (Adv, Pit)
лаеше:2 (V, Vpitf-m3s)
кучето:3 (N, Ncnsd)

Head-dependent pair:

<кога:1 лаеше:2> (VPA)
<лаеше:2 кучето:3> (VPS)

BURGER case:

There are two analyses, which attach the adverb either before the subject had been attached, or after it has been attached:

(**MOD-INT-OTHER-PHRASE**
(**ADV** кога)
(**HEAD-SUBJ**
(**FINITE-IMPERF-THIRD-SG00476-ORULE** лаеше)
(**THIRD_SG_NEUTER_NOUN_IRULE**,
DEF-THIRD_SG_NEUTER_NOUN_ORULE,
BARE-NP кучето)
)
)

and

(**HEAD-SUBJ**
(**MOD-INT-OTHER-PHRASE**
(**ADV** кога)
(**FINITE-IMPERF-THIRD-SG00476-ORULE** лаеше))
(**THIRD_SG_NEUTER_NOUN_IRULE**,
DEF-THIRD_SG_NEUTER_NOUN_ORULE,
BARE-NP кучето)
)

They have the same representations[1] in our format:

Lexical elements:

кога:1 (ADV)
лаеше:2 (FINITE-IMPERF-THIRD-SG00476-ORULE)
кучето:3 (THIRD_SG_NEUTER_NOUN_IRULE,
DEF-THIRD_SG_NEUTER_NOUN_ORULE)

Head-dependent pair:

<кога:1 лаеше:2> (MOD-INT-OTHER-PHRASE)
<лаеше:2 кучето:3> (HEAD-SUBJ)

Having these two representations, we need to use the rules for mapping of lists of categories between the two treebanks. Our rules are directed from

[1] This fact demonstrates one of the benefits of our representation – namely, that it is an indicator of spurious analyses.

BulTreeBank to BURGER Treebank, since our goal is to select the correct analyses in BURGER treebank. Thus, the rules have the form:

<list-of-BulTreeBank-categories> = <list-of-BURGER-categories>

Here are some examples

(Adv, Pit) = (ADV)
(V, Vpitf-m3s) = (FINITE-IMPERF-THIRD-SG00476-ORULE)
(N, Ncnsd) = (THIRD_SG_NEUTER_NOUN_IRULE,
DEF-THIRD_SG_NEUTER_NOUN_ORULE)
(VPA) = (MOD-INT-OTHER-PHRASE)
(**VPS**) = (HEAD-SUBJ)
(**VPS**) = (SUBJ-HEAD)

The result from the application of these rules is a new representation of the linguistic knowledge extracted from BTB, which is unified with the representation of the BURGER analyses. It can be used to select the correct BURGER analyses by comparing the two sets of descriptions. Note that the last two rules demonstrate the mapping with respect to word order which is explicitly encoded in the labels used by the BURGER grammar. Generally, such rules overgenerate over the BTB representations. As a result, we have more than one unified representation for one BTB analysis. In order to select a correct BURGER parse, we require an equality of the sets. Also, a procedure to go below the VPS label is needed in order to determine the word order between the head and the subject. This step is trivial in BTB, since the head is determined in most of the cases by its label[2].

The case study has shown that our idea of using BTB as a discriminator of the analyses is justified. 86 % of the correct analyses produced by BURGER were successfully selected by the discrimination properties extracted from BTB. The problematic cases refer to two linguistic presentations: coordination and complementation in NP. The first one needs a more elaborate set of rules, which to relax the BTB language model, since it rejects otherwise acceptable analyses. For example, BTB would accept the analysis in (6a) where there is a co-reference between the subject of the first conjunct and the pro-drop subject of the second, but would reject the analysis in (6b) where the subject is viewed as common to both predicates. The reason lies in the strong hierarchical mechanism of subcategorization:

(6a) [Кучето пристигна] и [залая].
[Dog-the came] and [started-to-bark].
[The dog came] and [started to bark].

(6b) Кучето [пристигна и залая].
Dog-the [came and started-to-bark].
The dog [came and started to bark].

The second problem arises from the fact that in BulTreeBank we accepted that all the dependents within and NP will be viewed as modifiers (see for the same decision in Butt et. al 1999: 46). However, in BURGER a hybrid approach has been taken – the relational nouns as well as the subject

[2] Only some phrases of type NP NP could need manual determination of the head.

counterpart in the frame of a deverbal noun are analyzed as complements. Thus, BulTreeBank contains analyses compatible with BURGER analyses which should be rejected by the ideology behind BURGER. This problem needs also smoothing of the BulTreeBank Schema for this particular task.

7 Conclusion and Future Work

In this paper, a method was presented for transferring of linguistic knowledge between two treebanks of Bulgarian, constructed within the same linguistic theory, but in its different versions and from different perspectives. BulTreeBank follows HPSG94 and the sentences have been analyzed per se. The target Treebank BURGER is more or less conformant to (Sag, Wasow and Bender 2003). It is being produced by an LKB-based Matrix grammar for Bulgarian, and has to discriminate among the overgenerated analyses. The linguistic information in BTB and BURGER is presented in the format of lexical categories and dependency relations. The actual categories and relations are HPSG generated on the basis of constituent labels in the corresponding analyses. A set of transferring rules on the level of the categories (or list of categories) is defined to ensure the compatibility of the representations. Currently our goal is to provide a mechanism for the usage of the linguistic knowledge encoded in BulTreeBank as a set of discriminating properties for the selection of the correct analyses produced by BURGER. Our current experiment proves the feasibility of this approach. We plan to extend the linguistic transfer with respect to higher coverage of data.

8 Acknowledgements

The work reported in the paper is supported by EuroMatrixPlus Project (http://www.euromatrixplus.eu/). We would like to thank Dan Flickinger for his support during the implementation of BURGER and to the reviewers for their comments on the paper.

References

Emily Bender, Dan Flickinger and Stephan Oepen. 2002. *The Grammar Matrix: An Open-Source Starter-Kit for the Rapid Development of Cross-Linguistically Consistent Broad-Coverage Precision Grammars*. Procedings of the Workshop on Grammar Engineering and Evaluation. 8-14.

Miriam Butt, Tracy H. King, M.-E. Nino & F. Segond 1999: *A Grammar Writer's Cookbook.* CSLI.

Atanas Chanev, Kiril Simov, Petya Osenova and Svetoslav Marinov. 2009. *The BulTreeBank: Parsing and Conversion*. In: Nicolov, Angelova, and Mitkov (Eds.). Recent Advances in Natural Language Processing V: Selected papers from RANLP 2007. Vol. 309 in the series "Current Issues in Linguistic Theory", John Benjamins Publ., Amsterdam. 321-330.

Bart Cramer and Yi Zhang. 2009. *Construction of a German HPSG grammar from a detailed Treebank.* Proceedings of ACL-IJCNLP, pp. 37-45.

Michael Daum, Kilian Foth, and Wolfgang Menzel. 2004. *Automatic transformation of phrase treebanks to dependency trees*. In Proceedings of the 4th Int. Conf. on Language Resources and Evaluation, LREC-2004, Lisbon, Portugal. 99-106.

Dan Flickinger. 2000. *On building a more efficient grammar by exploiting types*. Natural Language Engineering, 6 (1) (Special Issue on Efficient Processing with HPSG), 15 – 28.

Julia Hockenmaier. 2006. *Creating a CCGbank and a Wide-Coverage CCG Lexicon for German*. Proceedings of ACL2006. pp. 505–512.

Sandra Kübler, Wolfgang Maier, Ines Rehbein, Yannick Versley. 2008. *How to Compare Treebanks*. Proceedings of LREC 2008.

Stephan Oepen and Ulrich Callmeier. 2000. *Measure for measure: Parser cross-fertilization. Towards increased component comparability and exchange*. In Proceedings of the 6th International Workshop on Parsing Technologies, Trento, Italy. pp. 183 – 194.

Stephan Oepen, Kristina Toutanova, Stuart Shieber, Christopher Manning, Dan Flickinger, and Thorsten Brants. 2002a. *The LinGO Redwoods Treebank: Motivation and Preliminary Applications*. Proceedings of the 19th International Conference on Computational Linguistics (COLING 2002).

Stephan Oepen, Dan Flickinger, Kristina Toutanova, and Christoper D. Manning. 2002b. *LinGO Redwoods. A Rich and Dynamic Treebank for HPSG*. In Proceedings of The First Workshop on Treebanks and Linguistic Theories (TLT 2002), Sozopol, Bulgaria.

Petya Osenova and Kiril Simov. 2007. *Formal Grammar of Bulgarian*. IPP, BAS. Bulgaria. In Bulgarian.

Petya Osenova. 2010. *BURGER – Bulgarian Resource Grammar – Efficient and Robust.* Technical Report.

Carl Pollard and Ivan Sag. 1994. *Head-Driven Phrase Structure Grammar*. Chicago University Press and CSLI Publications.

Ivan A. Sag, Thomas Wasow, Emily Bender. 2003. *Syntactic Theory: a formal introduction*. Second Edition. Chicago: University of Chicago Press.

Kiril Simov, Petya Osenova, Alexander Simov, and Milen Kouylekov. 2004. *Design and implementation of the Bulgarian HPSG-based treebank*. In: Journal of Research on Language and Computation, Vol. 2, Num. 4.

Kiril Simov. *HPSG-based annotation scheme for corpora development and parsing evaluation*. In: Nicolov, Botcheva, Angelova and Mitkov (eds), Recent Advances in Natural Language Processing III: Selected Papers from RANLP 2003, John Benjamins, Amsterdam/Philadelphia, Current Issues in Linguistic Theory (CILT), volume 260, 2004. 327-336.