The Bulgarian HPSG Treebank: Specialization of the Annotation Scheme*

Petya Osenova, Kiril Simov

BulTreeBank Project
http://www.BulTreeBank.org
Linguistic Modelling Laboratory, Bulgarian Academy of Sciences
Acad. G. Bonchev St. 25A, 1113 Sofia, Bulgaria
petya@bultreebank.org, kivs@bultreebank.org

1 Introduction

The process of building our HPSG-based (for HPSG see [Pollard and Sag, 1994]) treebank involves two main tasks: integration of the pre-processing components and an adequate annotation scheme. The first is required for ensuring the consistency of the next levels and to facilitate annotators' work. The underlying techniques have to be adjusted to each other in such a way that maximum linguistic adequacy is attained at the subsequent stages. The integration goes into three directions:

1. *Looking-forward strategy.* It adjusts the resources and tools with respect to the deeper analysis and can be divided into two sub-mechanisms:

   (a) adaptive mechanism, which is not concerned only with the technical transfer of the information from one format into another. Basically, it is concerned with the adaptability of the stored information to more refined and elaborate language schemes. For example, the morphosyntactic tagset is converted into attribute-value pairs in accordance with the HPSG sort hierarchy. At this level one can specify information

*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.

from the sort hierarchy which is not presented within the morphosyntactic tagset explicitly.

(b) additive mechanism, which repairs the chunker output by adding head-nonhead dependency relations and grammatical functions to the phrase structures. The chunker usually recognizes maximal chunks without representing their internal structure. This task is performed by additional set of regular grammars which are applied internally to the chunks and benefit from the recognized top category.

2. *Looking-backward strategy.* It handles top-down constraint mechanisms like disambiguation within phrases or clauses. Such an automatic strategy has been recently exploited intensively for German (see [Müller and Ule 2002], [Kermes and Evert 2002]). In Bulgarian, determining the recursive structures is not a trivial task. For this reason we combine the strategy of manual identification of maximal phrases (usually clauses) with the automatic detection of the smaller chunk units in them. Compared to the above mentioned techniques, in this strategy the initial grammar is not precise enough and requires manual intervention.

3. *Creation of a golden standard.* The golden standard perspective ensures at least two things: (1) checking procedures over the relevant levels; (2) top-down filtering of incomplete partial analysis. Apart from that, it plays an important role in the construction of a reliable HPSG grammar for Bulgarian, because the existing formal analyses for this language are far from sufficient.

The annotation scheme defines the linguistic relevance of the information within the treebank and facilitates its annotation and validation. The final goal of our work is to construct a treebank containing complete HPSG analyses of Bulgarian sentences. In this case the theory dependency is inevitable. The process of achieving the goal consists of several levels of specialization of the annotation scheme. In [Simov and Osenova, 2003] we have described our first practical annotation scheme, which reflects the basic HPSG assumptions about the analyses of the linguistic structures. The scheme is practical because it abstracts over all details that can be inferred from the information in the treebank and the annotator has to deal with minimum information. Besides the theory relevance, the scheme has the following advantages: (1) it allows for further gradual transition to complete HPSG analyses; (2) it incorporates all the results from the pre-processing steps; (3) it reflects the monostratal nature of HPSG.

In this paper we present how the initial annotation scheme can be tuned to the requirements of some specific language facts and its connection with the pre-processing architecture. The structure of the paper is as follows. In Section 2 the

flexibility of the annotation architecture is described. Section 3 discusses the specialization of the annotation scheme. Section 4 focuses on the relation between the annotation scheme and the annotation process in the context of validation techniques. The last section outlines the conclusion and future work.

2 Towards a flexible architecture of annotation

In this section we present the architecture of the annotation process in two directions: 1. as integrating different preprocessing tools and 2. as mapping into the HPSG-based annotation scheme. In our work we try to ensure a maximal level of consistency via the combination of the preprocessing tools which produce 100 % correct analyses with a subsequent manual annotation. For this task we use constraints defined over the current version of the HPSG-based annotation scheme. In the design of the preprocessing tools we followed state-of-the-art techniques in the area and we tried to make them maximally independent from the current use in the construction of the treebank. On one hand, this imposes some problems in the mapping between the linguistic knowledge represented in these modules and the annotation scheme, but, on the other hand, it ensures wider usage of the produced resources.

2.1 Integrity of the separate components

The interrelation among the different tools and resources, which are used or have been created so far (a morphological tagger, a disambiguator, gazetteers, a machine-readable valence dictionary, partial grammars for named entities, numerical expressions and abbreviations, chunkers, a corpus), is determined in several dimensions.

From **the linguistic point of view** they are as follows:

1. The first one is the *interactive mode*, in which they are combined to support the real annotation task. The order of applications is relatively standard: tokenization, morphological processing, partial grammars, chunking. For instance, the identification of sentence boundaries determines in general the scope of application of the different grammars. Sometimes, however, the strict ordering is impossible, because of the linguistic complexity of the data. Cases of special interest here are: (1) consulting gazetteers yet at tokenization level and (2) the influence of interacting grammars on the annotation scheme.

2. The second dimension is their *extension mode* as relatively independent components. In this mode they become models for the creation of more complete

and elaborate resources, mainly lexical databases. Building the HPSG-based head-complement, head-adjunct and head-subject structures of the treebank, valency frames are extracted for a substantial number of different types of verbs. Additional corpus-derived lexicons are built of parentheticals, introductory units and phrases. The extended lexicons are then used for preidentifying certain linguistic units before the syntactic analysis.

From **implementational point of view** we rely on interactive usage of the tools within the CLaRK system ([Simov et. al., 2001], [Simov et. al., 2003]). The principle of cascadedness ([Abney, 1996]) is fully developed not only for the regular grammars engine, but also for all available tools like constraints, remove operation, insertion, transformations. Thus the output of one processing tool becomes input for another one and enables the gradual annotation of the data. Additionally, control operators can be used in order to direct the processing on the basis of the content of the annotation and the result from the application of the previous tools. This mechanism of processing also allows automatic repair of errors introduced at earlier stages of annotation.

2.2 The HPSG-based scheme

In this subsection we present the general language model, accepted within HPSG. HPSG is a monostratal theory and respects all the levels of language representation. Also it is lexicalist (like LFG) and 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. The data provides language elements with syntactic, semantic and pragmatic weight. In accordance with the theory, we take into account this mixed behaviour of the language elements and respectively, try to encode them appropriately in the CLaRK system as a DTD and a set of constraints over the XML documents. For this reason we distinguish between syntactico-phrasal elements which represent: a hierarchy of phrase types like head-complement, head-adjunct and the corresponding syntactic domains like nominal, verbal phrases (we have elements like `VPA`(djunct), `NPC`(omplement)); functional elements, which help us to represent the multi-dimensional nature of the encoded linguistic knowledge (thus, in order to preserve the original word order we have introduced discontinuous elements like `Disc`(ontinous)`E`(xtracted)), and pragmatic one `Pragmatic`); lexical elements which cover not only the lexical items in the lexicon but also the analytical word forms resulting from the application of some lexical rules (such elements are `N`, `V`, `Prep`).

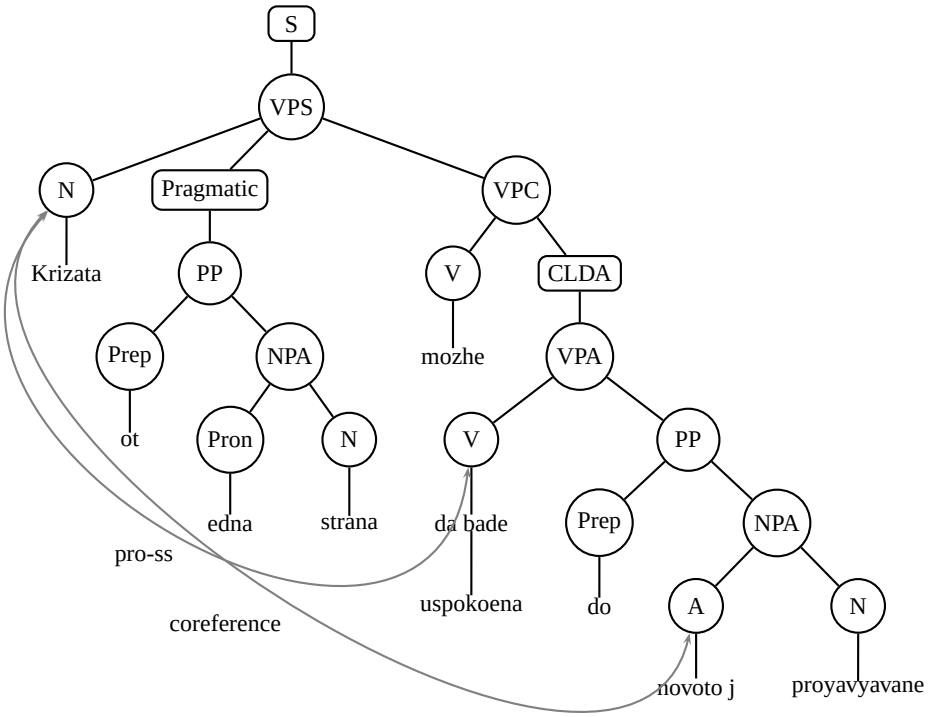

Figure 1: The tree for the sentence: “Krizata ot edna strana mozhe da bade uspokoena do novoto j proyavyavane” (The crisis, on one hand, can be diminished until its new appearance)

Additionally, we take into account the scoping characteristics of some elements and incorporate that information in two ways: lexically (for the negative and interrogative particles) and quasi-phrasally (for the emphasizing words like ’samo’(only)). The structure-sharing mechanism is widely applied for indicating phenomena like: binding, pro-dropness, secondary predication, anaphoric relations, clitic reduplication.

In Fig. 1. we represent an example tree from the BulTreeBank which illustrates the specifications of our annotation scheme. The tree consists of three types of nodes and two types of arcs. The leaves in the tree correspond to the words. The circles correspond to the sign objects in HPSG, the labels inside them determine the subsort of the sign and its constituent structure (lexical (`N`, `V`, `Prep`, `Pron`, `A`), head-complement (`VPC`, `PP`), head-subject (`VPS`), head-adjunct (`NPA`, `VPA`) and the category of the sign. The rectangles correspond to some additional prop-

erties of the signs below them. Here there are three kinds of such properties: the root node of the sentence [S], the da-clause [CLDA] and a pragmatic constituent [Pragmatic]. The immediate dominance relation between the signs is given by the structure of the tree itself. We allow for crossing branches (not presented here). The coreferences among the indices of the signs are given by additional arcs between the nodes of the tree. Here we have two of them: one connects the unexpressed subject of the CLDA clause with the expressed subject of the main verb; the second determines the binding of the possessive clitic and the subject of the main clause. But note that the coreference is a transitive relation and, hence, in this example it holds between the pro-dropped element and the possessive clitic as well. The division of the work between the pre-processing steps and the manual annotation is as follows: all lexical signs are added by the morphological analyser and disambiguator, the complex verb form *'da bade uspokoena'* is determined by the verb chunk grammar, all the NPAs are determined by the NP chunk grammar, the pragmatic PP *'ot edna strana'* is determined by the grammar for pragmatic and fixed expressions, the last PP in the sentence is analyzed by an "opportunistic" PP grammar. The rest of the analysis is done by the annotator.

The specialization of the annotation scheme consists of the following steps: (1) introducing new types of phrases in the HPSG sort hierarchy, including underspecified ones; (2) application of preference rules in the cases where more that one competing analyses can be applied; (3) implementing principles of HPSG for propagating the information along the tree, and checking the consistency of the results.

3 Specialization of the annotation scheme

3.1 Treatment of some special phenomena

In [Simov and Osenova, 2003] we focus on the hierarchy of phrases with respect to both characteristics - constituency and dependency. We also prescribe the order of the dependents realization (obliqueness). The syntactic domains and respective phenomena are distinguished. Phenomena of special interest are those, which combine features from several levels: syntax and pragmatics, syntax and information structure, syntax and semantics, syntax and discourse etc. In this respect we try not to lose information, but incorporate it properly.

For example, we assign to the vocatives the tag Pragmatic taking into account their pragmatic nature, but at the same time we keep track of their syntactic contribution (if any) via the coreference mechanism. It could indicate whether the vocative is structure-shared with the pro-dropped or explicit subject of the sentence, or with the object or a possessive clitic.

Another example of such a phenomenon are the so-called emphasizing words, which act on the level of Information structure. For instance, the adverb 'samo' (only) is not a typical adjunct and has a quantifying semantic nature. For that reason in focused phrases with 'samo', we project the category of the same phrase, not AdvPA. Thus it is indicated that its syntactic contribution is suppressed, but not its scoping characteristics.

There exist some linguistic phenomena, which have always been problematic for the consistency of the annotation schemes. Such as coordination, ellipsis and complement-adjunct distinction. For these phenomena there is not only one linguistically motivated solution. And all of them depend on each other and interfere. For example, if we state that coordination requires elements with the same grammatical role, then it depends on what we consider to be one or another grammatical role, especially in complement-adjunct cases. The same happens when ellipsis is involved.

One solution is underspecification. We identify coordination structures, but we do not try to specify some common category for them. That holds for all the levels. Of course, it works only when we have already decided on the coordination relation. The complement-adjunct distinction, on which our scheme relies, follows the principles of lexical semantics. Although still questionable, we use this dichotomic distinction, thus underspecifying the cases in between (our hierarchy lacks hybrid signs like adjunct-arguments or obliques, because they involve another variety of non-unified criteria).

Nevertheless, in the phenomena listed above, there are still hazy cases, where it is difficult to make distinctions for some elements being pragmatic or adjuncts. When there are several instances of pro-dropness within a smaller context, it is hard to decide whether the elements belong to the same coreference relation or not. Another problem is how to balance between coordination and ellipsis. These tough-nuts are partly repaired by the preference principles, stated in the next subsection.

3.2 Preference Rules

As it has been already pointed out, the basic assumptions in the annotation scheme are not suffice, because: there are always some mixed categories, which in different contexts behave differently, and for some phenomena there are more than one linguistically motivated possibility for an analysis.

Thus some 'preference' rules have to be added to the basic desiderata. Here we list some cases, in which their role is of great importance. Not surprisingly, the 'preference' approach is suitable for well-known problematic phenomena like coordination and ellipsis. For example, we have formulated the following prin-

ciples: *Prefer sentential coordination to pre-coordination* and *Prefer constituent coordination to ellipsis*. We take the first rule, because pre-coordination fits only in cases, when the selectional requirements of the verbs coincide. Our decision is similar to the decision within TIGER treebank [Brants and Hansen, 2002] in the sense that the selectional preferences of the heads are explicitly indicated: if the selectional preferences of two or more heads coincide, then the dependents are pre-coordinated, if not - then the preference rule is triggered. The preference of the sentential coordination guarantees consistency in all contexts. The second rule aims at decreasing the ellipsis in cases, when coordinated elements are of same dependence relation to the head.

Another dimension of preference rules application is the treatment of ellipsis. We distinguish between ellipsis, which can be restored within the sentence and ellipsis, which can be restored in the discourse or from our world knowledge. Sometimes the two interpretations are plausible, but the preference rule says: *If in the sentence there is an anchoring element for the ellipsis restoration, prefer it to the discourse one.*

Preference rules are needed for the treatment of the modal verbs, especially those, which could provide personal and impersonal reading at the same time (depending on the type of epistemic modality). The rule says: *In sentences with two readings possible: personal and impersonal, prefer the personal reading*. Thus subject extraction from CLDA clauses is avoided. For example:

(1) Predpriyatieto mozhe da uchastva v drugi targovski i grazhdanski
Enterprise-the can to participate in other merchant and civil
druzhestva
companies
Non-preferred reading: It is possible that the enterprise will participate in other merchant and civil companies.
Preferred reading: The enterprise can participate in other merchant and civil companies.

3.3 Data impact

Every treebank consists of sentences from certain domains (like in PDB: newspapers texts (see [Böhmová et al. 2003]); Verbmobil: spontaneous speech (see [Stegmann, Telljohann and Hinrichs 2000])) and this fact inevitably influences the decisions behind the annotation schemes. Our sentence bank consists of two components: grammar-derived examples (1 500 sentences) and corpus-derived ones (4 000) from newspapers, government documents, prose. Here we are not going to discuss cross-genre differences, but rather the annotation specificities accord-

ing to the sources: *Bulgarian grammars* and *electronic corpus*. Needless to say, both sub-banks of sentences represent the same linguistic phenomena, but from a slightly different perspective and with different distribution and variation.

We definitely disagree with the statement that the inclusion of isolated, grammar-derived sentences is methodologically unsound. They represent the so-called 'core set of sentences' within the treebank and thus can be used as a test-suite for Bulgarian. We consider our 'core set' similar to the Polish test suite, described in [Marciniak et al 2003], which includes non-corpus sentences aiming at covering most the linguistic phenomena, including rare ones, and consists of non-grammatical examples as well. At the moment our task is to ensure satisfying level of phenomena coverage than to concentrate on negative examples. Our grammar-derived sentences are of two kinds: they have already been extracted from Bulgarian literature to illustrate a specific phenomenon, or were constructed by the author for the same purpose. In this respect they are 'intended' sentences. Thus the annotation follows the general pre-classification as far as it is compatible with our formalized scheme. The co-reference relations are limited within the sentence, not higher.

Annotating corpus-derived sentences already requires additional strategies for an adequate interpretation. For example, when sentences of whole paragraphs or even divisions are annotated, then more elaborate co-reference mechanism is needed across sentence boundaries.

The other thing is that the texts supply a bigger variety of syntactic relations, typical for the connected text: a high frequency of introductory words (phrases), more complex word order, attachment ambiguities, dialogue patterns, nonstandard punctuation decoding. There are cases of head-complement dependency between two sentences. In comparison with the isolated sentences, the context-bound ones are more complex for syntactic annotation, but at the same time - easier for reference and ellipsis resolution.

In the phase of annotating such kinds of sentences, we are gradually extending the annotation scheme principles from basic to preference mode and some of the mechanisms on a suprasentential level.

4 Annotation scheme vs. annotation process: validation

The development of the annotation scheme is inevitably influenced by the process of annotation. This interaction follows generally the well-known cycle: *annotation and comparison - discrepancy between annotation scheme and data - changes in annotation scheme, tests for operationalization* [Brants and Hansen, 2002]. Below we present the concrete steps within our architecture:

1. There is a DTD, which encodes the domain specifications of the annotation scheme. It plays the role of a mediating pre-defined set of constraints over the annotation. It has the advantage of being flexible and easily adjustable to the requirements of the data. Additionally some of the non-local relations are encoded as constraints over the XML document. For instance, if a clause is defined to be `CLDA` clause, then its lexical head has to be 'da'-form of the main verb.

   The constraints regulate different linguistic properties. For example, they specify the dominance relation (`VPC` cannot dominate `VPA` or `VPS`; the sequence of a noun and a possessive clitic cannot be dominated by `NPA`, but only by the lexical `N`; `V-Elip` can dominate `pro-ss` etc.). Additionally, they describe the relevant attributes for an element (the nominals can have the attribute 'sort', which encodes whether they are person names, locations, organizations or common noun; `V-Elip` has the attribute 'type', whose values can be equality, variation or negation to the elliptic form etc.)

2. The annotators validate their interpretations against the current DTD and then report regularly lapses or problematic cases. The annotators are guided by the following prompts: 1. in the XML tree area the elements that have not been treated properly or have stayed unattached, are highlighted as problematic, thus requiring some repairing and 2. in an additional window there is a list of all the cases, in which the interpretation disagrees with the DTD or the set of constraints. The annotator can navigate through the list of the problematic places. When some error is repaired, the relevant error message disappears.

3. Then the reports are discussed and the relevant changes are made - first in the DTD, and then in the annotation guidelines. Note that this ordering of registering the changes is not arbitrary. The DTD regulates the annotation process in a more direct way that the annotation guidelines. Hence, the annotators' disagreement is more restricted and well controlled.

4. The annotations are checked by two people for linguistic consistency. The checks are made in two ways: 1. over the sentences in their linear sequence and 2. over the extractions from different domains (`CL`, `NPA`, `Pragmatic`, `VPC` etc.)

   This step is needed, because it is very difficult to set such a DTD, which neither overgenerates, nor undergenerates. Hence, there could be annotations, which do not violate the DTD, but at the same time they are linguistic-

ally non-justified, or annotations that do not fit DTD, even if being linguistically motivated.

At the same time, the process of syntactic annotation helps to discover and repair some morpho-syntactic errors or misconceptions. Such revisions are not an exception. See [Müller and Ule 2002] among others.

5 Conclusion and future work

The paper focuses on the process of tuning the annotation scheme of the Bulgarian HPSG treebank with respect to the interrelation between the HPSG theory and the specificity of the language data, and with respect to the connection between the pre-processor and the manual annotation. Thus we aim at producing a treebank, which is linguistically well interpreted and at the same time, close to the real texts.

We plan to produce a two-level grammar for Bulgarian, which will reflect the two-layer treebank-design. We will have a general grammar, based on the basic principles and specification in the annotation scheme and we will have a more elaborate grammar, based on the complex interaction of more or less discourse-connected phenomena.

References

[Abney, 1996] Steve Abney. 1996. *Partial Parsing via Finite-State Cascades.* In: *Proceedings of the ESSLLI'96 Robust Parsing Workshop.* Prague, Czech Republic.

[Böhmová et al. 2003] 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.

[Brants and Hansen, 2002] Sabine Brants and Silvia Hansen. 2002 *Developments in the TIGER annotation scheme and their realization in the corpus.* In: *Proceedings of the Third International Conference on Language Resources and Evaluation (LREC-2002).* Las Palmas: 1643-1649.

[Kermes and Evert 2002] Hannah Kermes and Stefan Evert. 2002.*YAC – A Recursive Chunker for Unrestricted German Text*. In: Proc. of the 3rd International Conference on Language Resources and Evaluation vol. V pp. 1805-1812.

[Marciniak et al 2003] Malgorzata Marciniak, Agnieszka Mykowiecka, Adam Przepiorkowski and Anna Kupsc.2003. *An HPSG-annotated test suite for Polish.Treebanks. Building and Using Parsed Corpora*. Kluwer Academic Publishers. pp 129-146.

[Müller and Ule 2002] Frank H. Müller and Tylman Ule. 2002. *Annotating topological fields and chunks – and revising POS tags at the same time.* In Proceedings of the Nineteenth International Conference on Computational Linguistics (COLING 2002), Taipei, Taiwan, August 2002.

[Pollard and Sag, 1994] Carl J. Pollard and Ivan A. Sag. 1994. *Head-Driven Phrase Structure Grammar*. University of Chicago Press, Chicago, Illinois, USA.

[Simov and Osenova, 2003] 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.

[Simov et. al., 2001] Kiril Simov, Zdravko Peev, Milen Kouylekov, Alexander Simov, Marin Dimitrov, Atanas Kiryakov. 2001. *CLaRK - an XML-based System for Corpora Development.* In: Proc. of the Corpus Linguistics 2001 Conference, pages: 558-560.

[Simov et. al., 2003] Kiril Simov, Alexander Simov, Milen Kouylekov, Krasimira Ivanova, Ilko Grigorov, Hristo Ganev. 2003. *Development of Corpora within the CLaRK System: The BulTreeBank Project Experience.* In: Proc. of the Demo Sessions of the 10th Conference of the European Chapter of the Association for Computational Linguistics (EACL'03), Budapest, Hungary.

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