13

HPSG-based syntactic treebank of Bulgarian (BulTreeBank)

Kiril Simov, Gergana Popova[1]*, Petya Osenova*[2]
The BulTreeBank Project
Linguistic Modelling Laboratory - CLPPI, Bulgarian Academy of Sciences

1. Introduction

This paper is dedicated to the BulTreeBank project[3] which has just started at the Linguistic Modelling Laboratory (LML), Bulgarian Academy of Sciences. Its main objective is to create a high quality set of syntactic structures of Bulgarian sentences within the framework of HPSG. Below we discuss the methodology we employ with emphasis on the new aspects of the adopted approach, as well as its expected results and their applications. The work on the project will be carried out mainly at LML in tight cooperation with researchers at the Seminar für Sprachwissenschaft (SfS), Eberhard-Karls-Universität, Tübingen, Germany.

The following section gives some background information about the project. Next we discuss the methodological assumptions at the core of the project, and in Section 4 we outline what steps we will need to take in order to achieve the aims of the project. In Section 5 we give a short description of the language resources available for Bulgarian and in conclusion we outline the expected results of our work.

2. Background of the project

Recent years have seen a proliferation of various language technology tools and applications requiring a substantial amount of linguistic information. One of the main sources of such information are linguistically processed collections of texts (written and in speech form) called *corpora*. Ideally, each text in a corpus should be expanded with an amount of linguistic information that will enable the user to find its right linguistic interpretation without any additional sources of information. Understandably, corpora are as yet far from such an ideal.

The situation is even worse in Bulgaria where intensive work in this area started a few years ago but there is to date little progress in the construction of corpora for Bulgarian. The available ones include mainly unprocessed texts or texts marked-up with structural information (titles, chapters, paragraphs and similar) and a few collections of texts marked-up with morphosyntactic information of the wordforms.

The process of formal grammar development is also in its inception. The most developed area is the study of morphology where several morphological analyzers and dictionaries were created – the MorphoAssistant system – (Simov et. al. 1990), (Simov et. al. 1992), and the Dictionary of Spelling, Pronunciation and Punctuation of Bulgarian – (Simov and Popov 1996), (Popov, Simov, Vidinska 1998) are two examples. These works were successfully used in several joint international projects for the annotation (semi-automatically) of Bulgarian texts on the level of wordforms. These comprise the European Union *Copernicus* initiative supported project MULTEXT-EAST, and the CLaRK Graduate Programme funded by the Volkswagen Stiftung.

---

[1]PhD student, Department of Language and Linguistics, University of Essex.
[2]Also at the Bulgarian Language Division, Faculty of Slavonic Languages, St. Kl. Ohridsky University, Sofia, Bulgaria
[3]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.

Apart from the morphological grammars and dictionaries there were several initiatives to develop syntactic parsers for Bulgarian. These were important first steps but not yet successful completions of the task and no parser is available to date. The products we are aware of include treatments of a limited number of syntactic phenomena and work with very small lexicons. The development of a formal grammar of Bulgarian with a wide coverage is still a task for the future. One of the prerequisites for progress in this area is an appropriately marked-up corpus of Bulgarian texts that can serve as a source of syntactic information. As we said, a corpus like that is yet to be compiled.

The BulTreeBank project is aiming to fill this gap. Ideally, the tree bank should contain samples of all the syntactic structures of the language or, if this aim proves to be unattainable, of the most frequent ones, covering as a minimum the structure of the Bulgarian simple sentence. These sentences should serve as templates for future corpora development, could become the basis for the development of a more comprehensive test suite for NLP applications, can be used as a source for grammar extraction and for linguistic research. The development of such a tree bank is inextricably linked to the development of a formal grammar and a parser for Bulgarian. We propose to write two grammars for Bulgarian, none of them exhaustive, but in our view a necessary stage on the way to a complete formal grammar of the language: one imposing very general constraints on the syntactic structure and one for partial parsing – see below.

At the center of the project is an attempt to combine grammar development and corpus compilation in order to create a bank of syntactic trees for Bulgarian annotated with detailed information and equipped with the software tools necessary for its exploration and use. In our work we will use the results of a previous project: the CLaRK Graduate Programme. These include an XML-based system for corpus development /the CLaRK system/, a morphosyntactic disambiguator and a text archive. We will also take into consideration some of the previous formal accounts for fragments of Bulgarian like (Avgustinova 1997).

3. Methodology

The main goal of the BulTreeBank project is to develop a high quality set (TreeBank) of syntactic trees for Bulgarian within the framework of Head-driven Phrase Structure Grammar (HPSG) – see (Pollard and Sag 1994). The term *syntactic tree* is used here for convenience, but in fact the actual syntactic structure for each sentence (or phrase) will be a graph in accordance with the common view in HPSG of linguistic objects. Our hope is to demonstrate the varieties of syntactic patterns in Bulgarian more exhaustively than it has been done so far and within a contemporary linguistic theory - HPSG.

Granularity

The descriptions of the linguistic information in the TreeBank will be very detailed in order to demonstrate the information flow in the syntactic structure of the sentences in the TreeBank.

Theory Dependency

An annotation scheme usually has to be *theory independent* in order to allow different interpretations of the tagged texts in different linguistic frameworks. We think, however, that on a certain level of granularity (as it was mentioned linguistic descriptions in the BulTreeBank will be very detailed in order to demonstrate the information flow in the syntactic structure) we will have to exploit some linguistic descriptions that are theory dependent. We choose HPSG for the following reasons:

- HPSG is one of the major linguistic theories based on rigorous formal grounds;
- HPSG allows for a consistent description of linguistic facts on every linguistic level: phonetic and phonological, morphological, syntactic, even the level of discourse. Thus, it will ensure the easy incorporation of linguistic information which does not be-

long to the level of syntax if such is needed for the correct analysis of a given phenomenon;
- HPSG allows for both integration and modularization of descriptions and will therefore enable different experts to work on different parts or levels of analysis.
- The formal basis of HPSG allows easy translation to other formalisms.
- There are universal HPSG principles that can be used to support the work of the annotators during the development of the TreeBank.

The Logical Formalism

We not only choose HPSG to be the linguistic theory within which we will explicate the syntactic structure of Bulgarian texts, but make a step further and choose the actual formalism that we will use in the annotation process: namely, SRL augmented (in an RSRL style) with some relations common in HPSG. For SRL see (King 1989), (King 1994) and (King 1999), for RSRL see (Richter 1997), (Richter 1999), (Richter, Sailer and Penn 1998), (Richter, Sailer and Penn 1999). For the annotation we will use SRL descriptions called feature graphs, which correspond to clauses in a normal form of a SRL theory. In their complete form feature graphs are morphs in the sense of (King 1994). Such detailed descriptions will be extremely useful in the future exploitation of the TreeBank, but they might be difficult to use in the annotation process. Here we hope to use the (special) inference mechanisms of SRL and some of the HPSG principles (universal and specific to Bulgarian) in order to enable the annotator to provide only part of the needed information with the rest of it being inferred automatically (see (King and Simov 1998) for a special inference for such purposes).

Partial Analyses and Predictions

As was mentioned in the previous paragraph, the detailed level of the envisaged syntactic descriptions will require entering a huge amount of linguistic information for each sentence. In order to minimize the necessary human intervention, we will exploit all possibilities to provide an automatic partial analysis of the input string before the actual annotation starts. We would also use the partial information entered by the annotator in order to predict or constrain the possible analyses in other parts of the whole description of the element. In this way we will exploit all the constraints available from pre-encoded grammars.

4. More specific tasks

The previous section outlined the broad methodology of the project. But the development of an HPSG-based bank of syntactic trees will comprise also the following more specific tasks:

- It will be necessary to identify which part of the HPSG sort hierarchy as described in (Pollard and Sag 1994) and others can be successfully applied to Bulgarian and also what modifications and additions will be necessary for a formal grammar of Bulgarian. The Bulgarian-specific part of the sort hierarchy will be subject to change during the development of the TreeBank according to the demands of the Bulgarian data.
  The linguistic knowledge represented in the sort hierarchy will be the backbone of the syntactic descriptions of the Bulgarian data. It will define all possible linguistic structures that will be further constrained by the grammar and/or the information entered by the annotators. The annotation schemata that will be employed during the project will allow for composite tag definitions. It will be possible to decompose each tag in the syntactic structure so that the grammatical information represented by the tag and its elements get distributed to the relevant substructures.
- We would need to provide a representation of the HPSG Universal Grammar principles from (Pollard and Sag 1994). This very general grammar will be used as a top constraint during the annotation of the trees in the TreeBank. For example, each headed phrase in the TreeBank has to satisfy the Head Feature Principle of HPSG. This grammar will be further extended by Bulgarian specific principles during the de-

velopment of the TreeBank. These specific principles can deal with word order in Bulgarian, for instance.

- It would be necessary to adapt the lexical information already contained in the morphological dictionary of Bulgarian (Popov, Simov, Vidinska 1997) to the sort hierarchy and to the HPSG principles. The morphological analyzer available to the project will be used to annotate the words in the sentences. Additionally, the Part-of-Speech disambiguator developed under the CLaRK Programme will be used to reduce ambiguities after the morphological analysis of the sentences. A description of the relevant linguistic resources is given below.

  We will develop a reliable grammar for a partial parsing of the phrases constituting the lower part of each syntactic annotation – a few dominant levels over the lexical categories. Here we will follow ideas in the work of (Abney 1990), (Abney 1991). A preliminary version of such a grammar has already been developed under the CLaRK project with the goal of context analysis in a concordance program. This grammar will be used to analyze the input sentences after the morphological analyses in order to add unambiguously the syntactic structure to some minimal elements of the whole sentence.
- Several well-defined subgrammars are already under development. These subgrammars follow ideas from information extraction and are concerned with closed sets of expressions like numerical expressions, dates, names like "the city of Sofia" or "Prof. Petrov". Each of these subgrammars will assign to each such expression an appropriate description that will allow the consequent incorporation of the expression within the whole syntactic structure of the sentence.

  The result of the application to each sentence of the constraints encoded in the HPSG sort hierarchy, HPSG principles, morphological analyzers and the partial grammar could be considered similar to the supertag ideas in (Joshi, Srinivas 1994). The annotators will have access to all linguistic information relevant to a given sentence and will have to choose the right analyses where the information is ambiguous.
- A core set of sentences demonstrating typical syntactic phenomena in Bulgarian is under development.

  Example sentences will be extracted in the first instance from published Bulgarian grammars which already give some analyses and provide certain coverage of syntactic phenomena in Bulgarian. With respect to syntactic patterns that are not covered by the existing grammars we will proceed as follows: (1) we will first search for the respective patterns in real texts using some of the concordance tools we already have implemented in the CLaRK system; and (2) if we cannot find exemplary sentences we will construct artificial examples.

  In this core set of sentences we will try to cover the following basic syntactic phenomena: sentence and clause types, complementation, agreement, modification, diathesis, modality, tense and aspect, word order, coordination, negation. These universal phenomena will be further detailed with respect to the language-specific characteristics of Bulgarian. Some of these are:

  Bulgarian is a highly inflective, null-subject language. It allows for an object doubling and imposes special word order constraints on both pronominal and verbal clitics. It has a rich analytical verbal complex with discontinuous behavior. One of the peculiarities is a clitic-like definite article.

  We can view this part of the TreeBank as a "draft" test-suite (see the TSNLP project (Oepen et. al. 1994) and the Polish HPSG TreeBank (Marciniak et. al. 1999)). Following the guidelines for test-suites, we will strive to give negative constraints and examples within the core set of sentences. These will be used also during the second phase of the annotation of the sentences. Whenever the sentences chosen turn out to be ambiguous, we will rank the ambiguity according to the plausibility of the corresponding

reading. At this stage of the development of the TreeBank the plausibility of the analysis will be judged by the annotators on the basis of their intuitions about the usability of the corresponding sentence or phrase. We hope that such plausibility information will turn out to be useful during the annotation of the base set of sentences as well.

- It will be necessary to prepare guidelines for the annotation of the actual sentences. It is important to make two points here:
  (i) the annotators will specify mainly information about the syntactic configurations over chunks and phrases over chunks (such as PP attachment), and
  (ii) some frequently used descriptions will be abbreviated in an appropriate manner in order to facilitate the process of annotation. These guidelines will allow annotators with different levels of experience to work on the TreeBank.
  For example, an abbreviation for noun phrases is defined as shorthand for a description constraining all noun phrases in (Pollard and Sag 1994):

$$\text{NP[i]} \Leftrightarrow \left[\text{LOC} \left[\begin{array}{l}\text{CAT} \left[\text{HEAD noun}\right] \\ \text{CONTENT: INDEX [i]}\end{array}\right]\right]$$

- We will assemble a base set of annotated sentences taken from real Bulgarian texts. This set of sentences will constitute the main body of the TreeBank. In contrast to the core set developed during the first phase of the project, here the sentences will be chosen in such a way as to demonstrate the most widely spread syntactic phenomena in Bulgarian texts. Special attention will be paid to resolving ambiguity in the context of use. This will be done by including in the TreeBank not only isolated sentences, but whole paragraphs or even a number of consequent paragraphs. When sentences are annotated in context, a description of the disambiguating context will be added to the syntactic annotation. For the moment we envisage an informal description of the context, but we hope to be able to formalize that, time and resources permitting. The actual annotation process will proceed in accordance with the following steps:
  Sentence extraction. Sentences will be chosen from the text database collected during the CLaRK Programme (this database is being constantly updated and enlarged). A description of this text archive is given below. We will use the concordancer developed under the CLaRK project. The concordance software will have access to the morphological information for the words in the texts and it will allow queries similar to, for example: "Find all sentences in which a personal pronoun is followed by a passive participle which is followed by a preposition."
  Automatic pre-processing. Each annotated sentence will be pre-processed to allow the system to add linguistic information that possibly describes its syntactic structure. This pre-processing will include:
  (i) *Morphosyntactic tagging* — each word will be marked up with the appropriate morphological information, namely: part of speech, gender, number, tense, person, transitivity, etc.;
  (ii) *Part-of-speech disambiguator* — for each ambiguous word the most probable part-of-speech will be predicted;
  (iii) *Partial parsing* — where possible minimal constituents will be identified;
  (iv) *Addition of syntactic information* — on the basis of the results from the previous steps and the general information available from the type hierarchy and the universal grammar the possible syntactic information will be given.
  Obviously, the precision of the information added with each step decreases. We can be quite certain of the accuracy of the analyses on the morphological level but the syntactic analyses will be incorrect in a large number of cases or will contain a high degree of ambiguity. Therefore, each consecutive step will require increased human interven-

tion, with most of it needed on the syntactic level. The aim of the pre-processing is to minimize the amount of this human effort.

The results of the automatic pre-processing will be loaded into the CLaRK system in two forms:

(i) *XML mark-up*. This will be the actual annotation of the sentence. The annotator will have to edit this information in order to describe the sentences fully.

(ii) *Constraints over XML documents.* The constraints from the grammars will be encoded as constraints over XML documents and will guide the annotator in the process of annotation but this information will not be part of the actual syntactic description of the sentence.

To avoid the possible combinatorial explosion on the last level we plan to apply special compilation techniques which will distribute syntactic information locally onto the overall structure and will encode part of that information as constraints over the structure.

Manual annotation. The ambiguities remaining after the pre-processing stage will have to be resolved manually by the annotators. The system will be able to support this process by propagating constraints that follow automatically from the information supplied by the annotator. To give a simple example, if the annotator has had to specify one of two daughters as the head daughter, the system will automatically percolate the relevant head-features to the mother and further up the tree.

- We plan to develop further the CLaRK system in order to support the annotation process. During the development of the TreeBank we envisage to extend the set of possible constraints over XML documents that can be encoded in the system. We will also connect the CLaRK system to a feature logic engine in which we can implement the grammars. We will implement new inference mechanisms to support the annotators even further using the full predictive power of the grammar. Very often in the annotation process annotators develop their own *ad hoc* procedures and constraints for some specific cases. We plan to develop a simple macro language in the CLaRK system that will allow their easy encoding. We will also apply some techniques for generalization and learning from the already developed syntactic trees.

5. Available Linguistic Resources for Bulgarian

In this section we give a brief overview of the linguistic resources we will be able to use freely in compiling the tree bank.

1. *Morphological analyzer - Slovnik.* This is a system for morphological analyses and generation based on (Popov, Simov, Vidinska 1997) developed by Ognyan Chernokozhev and Atanas Kiryakov at OntoText Lab. The system recognizes the wordforms of more than 110 000 Bulgarian lexemes and assigns to them the appropriate morphosyntactic characteristics.
2. *Neural Network MorphoSyntactic disambiguator for Bulgarian.* This system was developed under the CLaRK Programme by Stanislava Vlaseva, Petya Osenova, and Kiril Simov. Currently we have a corpus of about 2600 sentences extracted from newspapers, narratives and textbooks which demonstrate some of the most frequent ambiguities on the morphosyntactic level. We have trained a neural network on the basis of 1500 sentences given in different order and with different number of ambiguous words. The resulting network predicted the right part-of-speech for 95.25% of the words in the rest of the sentences in the corpus. Therefore the accuracy on the morphosyntactic level is 93.17%.
3. *Text archive.* We have collected from the Internet a set of Bulgarian texts. These cover about 15 000 000 running words. We are converting them in TEI (see (Text Encoding Initiative 1997)) compatible XML markup on the paragraph level. We intend to mark them up with morphological information using the Slovnik system and the corpora de-

velopment tools implemented under the CLaRK programme. 33% of the texts come from fiction, 60% from newspapers and about 7% from legal texts, government bulletins and other genres.

6. Expected results

At the end of the project we expect to have a set of Bulgarian sentences marked-up with detailed syntactic information. These sentences will be extracted mainly from authentic Bulgarian texts. They will be chosen with two criteria in mind. First, they will have to cover the variety of syntactic structures of Bulgarian. Second, they should reflect the statistical distribution of these phenomena in real texts. A core set of sentences will be extracted to serve as a test-suite for software applications incorporating syntactic processing of Bulgarian texts. The project should result also in a reliable partial grammar for automatic parsing of phrases in Bulgarian. This grammar will be extensively tested and used during the creation of the TreeBank. It will be used as a module separate from the TreeBank in tasks which require only partial parsing of natural language texts such as information retrieval, information extraction, data mining from texts and etc. Work on the treebank will require the creation of software modules for compiling, manipulating and exploring the data. This software will support both the creation of the TreeBank, and its use for different purposes such as automatic extraction of grammars for Bulgarian.

References

Abney St 1990 Syntactic Affixation and Performance Structures. In Bouchard D. Leffel K. (eds), *Views on Phrase Structure*. Kluwer Academic Publishers, Dordrecht.

Abney St 1991 Parsing By Chunks. In: Berwick R., Abney St., Tenny C. (eds), *Principle-Based Parsing*. Kluwer Academic Publishers, Dordrecht.

Avgustinova T 1997 *Word Order and Clitics in Bulgarian*. Saarbrücken Dissertations in Computational Linguistics and Language Technology, Volume 5. University of Saarbrücken.

Joshi AK, Srinivas B 1994 Disambiguation of Super Parts of Speech (or Supertags): Almost Parsing. In: *Proceedings of the 17th International Conference on Computational Linguistics (COLING '94)*. Kyoto, Japan.

King PJ 1989 *A Logical Formalism for Head-Driven Phrase Structure Grammar*. Doctoral thesis. Department of Mathematics, University of Manchester, Manchester, England.

King PJ 1994 *An expanded logical formalism for Head-driven Phrase Structure Grammar*. Sonderforschungsbereich 340 technical report 59. Sonderforschungsbereich 340, Seminar für Sprachwissenschaft, Eberhard-Karls-Universität, Tübingen, Germany.

King PJ 1999 Towards Truth in HPSG. In: Kordoni V. (ed), *Tübingen Studies in Head-Driven Phrase Structure Grammar*. Volume 2, pages 301-352. Arbeitspapiere des SFB 340, Bericht Nr. 132. SFB 340, Tübingen, Germany.

King PJ, Simov K 1998 The automatic deduction of classificatory systems from linguistic theories. *Grammars*, 1(2): 103-153. Kluwer Academic Publishers, The Netherlands.

Oepen S., Netter K., Klein J 1998 *TSNLP - test suites fir natural language processing*. In: Nerbonne J, (ed) *Linguistic Databases* CSLI Lecture Notes. CSLI Publication, Stanford, USA

Marciniak M, Mykowiecka A, Przepi'orkowski A, Kup's'c A 1999 Construction of an HPSG TreeBank for Polish. In *Proceedings of the ATALA conference*.

Pollard C, Sag I 1994 *Head-Driven Phrase Structure Grammar*. University of Chicago Press, Chicago, Illinois, USA.

Popov D, Simov K, Vidinska S 1998 *A Dictionary of Writing, Pronunciation and Punctuation of Bulgarian Language*. (In Bulgarian). Atlantis LK, Sofia, Bulgaria.

Richter F1997 Die Satzstruktur des Deutschen und die Behandlung langer Abhangigkeiten in einer Linearisierungsgrammatik. Formale Grundlagen und Implementierung in einem HPSG-Fragment. In: Hinrichs E, Meurers D, Richter F, Sailer M, Winhart H (eds*) Ein HPSG-Fragment des Deutschen. Teil 1: Theorie.* SFB Report 95 Universität, Tübingen, Germany.

Richter F 1999 RSRL for HPSG. In: Kordoni V. (ed) *Tübingen Studies in Head-Driven Phrase Structure Grammar.* Arbeitspapiere des SFB 340, Nr. 132, Volume 1. Universität, Tübingen, Germany.

Richter F, Sailer M, Penn G 1998 A Formal Interpretation of Relations and Quantification in HPSG. In: Bouma G, Kruijff G.-J.M., Oehrle R.T. (eds): *Proceedings of the FHCG-98*. Saarbrücken, Germany.

Richter F, Sailer M, Penn G 1999 A Formal Interpretation of Relations and Quantification in HPSG.

In: Bouma G, Hinrichs E, Kruijff G.-J.M., Oehrle R.T. (eds*): Constraints and Resources in Natural Language Syntax and Semantics*. CSLI Publications. Stanford, USA.

Simov K, Popov D 1996 Creating a morphological dictionary of the Bulgarian language. In: *Proceedings of COMPLEX'96 Conference*. Budapest, Hungary.

Simov K, Angelova G, Paskaleva E. 1990 MORPHO-ASSISTANT: The proper treatment of morphological knowledge. In: *Proceedings of the 15th International Conference on Computational Linguistics (COLING '90)*, volume 3, pp 453-457. Helsinki, Finland.

Simov K, Paskaleva E, Damova M, Slavcheva M 1992 MORPHO-ASSISTANT - a knowledge based system for Bulgarian morphology. In: *Proceeding of Demo Descriptions of Third conference on Natural Language Application*. Trento, Italy.

Text Encoding Initiative 1997 *Guidelines for Electronic Text Encoding and Interchange.* Sperberg-McQueen C.M., Burnard L (eds).