HPSG-based syntactic treebank of Bulgarian
(BulTreeBank)

Kiril Simov, Gergana Popova, Petya Osenova
The BulTreeBank Project
Linguistic Modelling Laboratory - CLPPI, Bulgarian Academy of Sciences
Acad. G. Bonchev Str. 25A, 1113 Sofia, Bulgaria
Tel: (+3592) 979 28 25, Fax: (+3592) 70 72 73
kivs@bgcict.acad.bg, gdpopo@essex.ac.uk, osenova@slav.uni-sofia.bg

Our paper will be mostly dedicated to a project about to start at the LML, BAS. Its main objective is to
create a high quality set of syntactic structures of Bulgarian sentences within the framework of HPSG.
We will discuss the methodology of the project with emphasis on the new aspects of the adopted
approach, as well as its expected results and their applications.

Methodology. 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 (and 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: (1)
HPSG is one of the major linguistic theories based on rigorous formal grounds; (2) 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 belong to the level of syntax if such is needed for the correct
analysis of a given phenomenon; (3) HPSG allows for both integration and modularisation of
descriptions and will therefore enable different experts to work on different parts or levels of analysis.
(4) The formal basis of HPSG allows easy translation to other formalisms. We not only choose HPSG
to be the linguistic theory within which we will explicate the syntactic structures, but make a step
further and choose the actual logical formalism that we will use in the annotation process: namely, SRL
for HPSG. For the annotation we will use descriptions called feature graphs. Such detailed descriptions
will be extremely useful in the future exploitation of the Tree-Bank, but they might be difficult to use
in the annotation process. Here we hope to use the (special) inference mechanisms of the logic and
some of the HPSG principles in order to allow the annotator to provide only part of the needed
information with the rest of it being inferred automatically. In order to minimise 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.

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 mainly extracted 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 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.

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