A Language Resources Infrastructure for Bulgarian

Kiril Simov, Petya Osenova,
Sia Kolkovska, Elisaveta Balabanova, Dimitar Doikoff

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

Abstract

This paper describes the infrastructure of a basic language resources set for Bulgarian in the context of BLARK initiative requirements. We focus on the treebanking task as a trigger for basic language resources compilation. Two strategies have been applied in this respect: (1) implementing the main pre-processing modules before the treebank compilation and (2) creating more elaborate types of resources in parallel to the treebank compilation. The description of language resources within BulTreeBank project is divided into two parts: *language technology*, which includes tokenization, morphosyntactic analyzer, morphosyntactic disambiguation, partial grammars, and *language data*, which includes the layers of the BulTreeBank corpus and the variety of lexicons. The advantages of our approach to a less-spoken language (like Bulgarian) are as follows: it triggers the creation of the basic set of language resources which lack for certain languages and it rises the question about the ways of language resources creation.

1. Introduction

One of the central questions within Human Language Technologies discusses ”what is minimally required to guarantee an adequate digital language infrastructure for a language”? (Bimnnenpoorte et al. 2002). Thus the notion of Basic Language Resources Kit (BLARK) was introduced and discussed within NLP community. Its definition and scope have been considered in several European initiatives, see ENABLER Network and Dutch LT Platform among others.

BLARK is defined as a set of three distinct groups: *applications*, *processing modules* and *language data* (Strik et. al. 2002).

This paper aims at localizing the BulTreeBank (a project devoted to the creation of an HPSG-based treebank of Bulgarian) language resources for Bulgarian within the notion of BLARK. Several problematic issues are addressed:

- How close are the language resources (LRs) to the BLARK requirements?
- How can a more advanced resource like a treebank give rise to a number of basic language resources, which lack in this language?
- How can the existent LRs be turned into a solid basis for the development of other LRs?
- Is it always the case that basic language resources are produced first, and the more advanced ones afterwards?

We consider the creation of such a complex language resource an application which tests the availability of other resources and processing modules. During the project we have discovered the ’white spaces’ in the resources for Bulgarian. We have been trying to fill these gaps developing LRs with respect to the actual work on the treebank. However, in this creation we have not restricted ourselves to the needs of the treebank development only, but we also have envisaged wider range of NLP tasks, such as information extraction, grammar checker, parsing.

2. Treebanking as Basic Language Resources Compiler

The creation of a treebank for a “less-spoken” language like Bulgarian imposes certain challenges due to the limited scientific, technological and financial resources. As a central task we considered the organization of the work with respect to the minimization of human intervention and the achievement of the project goal. The greatest problem appeared to be the lack of already available set of language resources, which to serve as a base for the treebank compilation. Thus, on the one hand, we were aware before the start of the project that most of the required resources had to be produced by us. On the other hand, we have used the situation as a possibility to construct a variety of resources to support the creation of the treebank and to be extensively tested within the project. As a result, we have produced a basic set of language resources for Bulgarian, which are easily adaptable for different mono- and multilingual NLP tasks.

Generally, two strategies have been applied:

1. *Before starting the treebank creation,* we have implemented basic processing modules: tokenization, morphological analyzer, disambiguator, named entities recognition modules, partial grammars, the text corpus.
2. *Parallel to the treebank creation,* we have compiled resources, which need more elaborate and high quality information: specific lexicons of verb frames, lists of fixed phrases, specific introductory patterns for newspaper texts, parenthetical expressions.

Note that the time distinction (before and parallel to the treebank creation) is a relative one, because all the primary resources have been further developed and tested.

The creation of the resources is governed by two principles:

Bootstrapping principle: Its aim is to obtain as much information as possible at the very basic processing levels. For instance, in the tokenization case, according to "a general token classification" (Osenova and Simov 2002), the tokens are divided into the potential classes of common words, abbreviations, names etc.

Corpus-driven principle: Several results are simultaneously obtained by using extraction and observation procedures: the gazetteers are compiled, the dictionaries are improved and the tools are tested against unrestricted data.

The creation of our resources has always been in close connection to the overall annotation process of data. It comprises the following steps:

2.1. Sentence/text Extraction from the Corpus

The source of the sentence extraction is the BulTreeBank text corpus (72 mln. running words at the moment). We aimed at sentences with different lengths and from different genres. Sentence extraction was combined with text extraction, which means that whole newspaper articles or book chapters have been selected for annotation. As supporting modules, the CLaRK concordancer and grammar engine have been relied upon.

2.2. Automatic Pre-processing

Each sentence needs first to be pre-processed at all the levels, that precede deeper syntactic annotation. These include: (1) Morphosyntactic tagging; (2) Named entity recognition; (3) Morphosyntactic disambiguation; (4) Partial parsing (chunking). We aim at a result of a 100 % accurate partial parse of a sentence. The accuracy is checked and validated by a human annotator with the assistance of the CLaRK System (Simov et. al. 2001).

2.3. HPSG Step

The result from the previous step is encoded into an HPSG compatible representation. Then HPSG parsing takes place. The output is encoded as a parse forest.

2.4. Resolution Step

The parse selection is performed by supplying partial information and navigation in the parse forest. However, relevant for this paper is the first step, because the preprocessing module comprises most of the basic LRs. The result is manually checked, the lexicons are extended to cover the tokens within the corpus, the phenomena with bigger impact over the corpus are considered.

3. The BulTreeBank LRs in the Context of BLARK

Language technology is supposed to include the following modules: tokenization and named entities recognition, morphological analyzer and disambiguator, syntactic and semantic analyzer. The data is supposed to consist of: a mono-lingual lexicon, annotated corpus of texts (a treebank with syntactic, morphological and semantic structures) and benchmarks for evaluation. Within this BLARK notion a priority list can be proposed depending on what exists and what needs development in a certain language. In our case the priority was to create all the complementary resources which would support the treebank creation.

3.1. BulTreeBank Language Technology

Tokenization

There is a hierarchy of tokenizers within the CLaRK system, which tokenize the texts in an appropriate way. Additionally, one can decide what the category of the token is and to assign it.

The Morphosyntactic analyzer

It assigns all possible analyses to the word tokens. The lexicon is too large to be loaded as one grammar in CLaRK and this is why we have divided it into several grammars which are applied in a group. The separation of the lexicon is on the basis of the frequencies of the word forms within the corpus. In this way the application has been speeded up. As it was mentioned above, together with the morphosyntactic analyzer we use the gazetteers. They are also implemented within the CLaRK system. In the places where competing analyses arise between a common word and a name or an abbreviation, we try to use the token classification strategy and the prompts of the context. If there is no clear preference, we leave the decision to the human annotator.

MorphoSyntactic Disambiguation

We have already implemented a preliminary version of a rule-based morpho-syntactic disambiguator, encoded as a set of constraints within the CLaRK system. This rule-based disambiguator exploits context information like *agreement between an adjective and a noun in a noun phrase*, specific positions like *a noun after a preposition*, but it also deals with some fixed phrases. The disambiguator does not try to solve unsure cases, but leaves them for further processing. Its coverage is about 80 %. For the purposes of the treebank we have manually disambiguated the rest 20 %. For automatic disambiguation we have developed a neural-network-based disambiguator (see (Simov and Osenova, 2001)). It achieves accuracy of 95.25% for part-of-speech and 93.17% for complete morpho-syntactic disambiguation. We plan to train several taggers and then test them over the manually disambiguated data and manually to check the places where there is no agreement between the taggers and the disambiguated data. As a result, a satisfying validation procedure will be achieved.

Partial Grammars

We have constructed such grammars for:

1. **Sentence splitting.** At the moment it is fully automated and reliable only for the basic and clear cases. For solving complex and ambiguous cases this grammar is combined with supporting modules for abbreviation detection.

2. **Named-entity recognition.** Identifying numerical expressions, names, abbreviations, special symbols (see (Ivanova and Dojkoff 2002), (Osenova and Kolkovska 2002)). They are designed to work in cooperation with

the morphosyntactic analyzer. If necessary, the grammars can overwrite the analysis of the morphosyntactic analyzer.

3. **Chunking.** Two basic modules have been developed: an NP chunker (see (Osenova 2002), (Osenova and Kolkovska 2002)) and a VP chunker (Slavcheva 2002). Generally speaking, the chunking process conforms to the following requirements: it deals with non-recursive constituents; relies on a clear-indicator strategy; delays the attachment decisions; ignores the semantic information; aims at accuracy, not coverage. Additionally, there are chunk grammars for APs, AdvPs, PPs and some non-problematic clauses.

3.2. BulTreeBank Language Data

3.2.1. BulTreeBank Corpus

The text archive

It is intended to yield the size of a national corpus, that is, 100 million running words. Since the data are gradually annotated, its status at the moment is approximately as follows:

1. Nearly 90 million running words are collected from different sources in HTML and RTF formats. In order to compile a representative and balanced corpus of Bulgarian texts, we tried to gather a variety of different genres: 15% fiction, 78% newspapers and 7% legal texts, government bulletins and others.

2. About 72 million running words are converted into XML documents, marked up in conformance with the TEI guidelines. This conversion is automatic: for each source of text we developed a separate tool for extraction of the relevant information like the text itself, but also the author information, genre classification (where it is available), and other meta-information. The tools are implemented in Prolog and the CLaRK system.

3. 10 million running words are morphologically analyzed. This part of the text archive was used to select data for manual disambiguation and in future it will be substituted by an automatically disambiguated version of the full text archive.

4. Over 1 000 000 running words are morphosyntactically disambiguated by hand. This part of the text archive is used in two ways within the project: (1) as a source of sentences and articles which to be annotated syntactically and included in the treebank, and (2) as training and testing data for POS disambiguation of Bulgarian texts.

The Treebank

The Treebank (200 000 words) is a part of the BulTreeBank corpus. It is meant to be syntactically processed and consists of two layers:

1. Core set of sentences (1 500) - these are sentences, extracted mainly from Bulgarian grammars. They will serve as a test suite and gold standard for Bulgarian, because they are considered to represent the variety of the linguistic phenomena in our language. All of them are processed manually and therefore the analyses are of the highest quality.

2. Treebank (up to now 10 000) - these are sentences, extracted mainly from the electronic archive. First, they are pre-processed automatically, then the attachment operations are performed by the annotators. Note that the annotators are restricted by the software device and thus the analyses are consistent at this level. Finally, the sentences are post-edited and corrected.

3.2.2. Lexicons

The Morphological Dictionary

The dictionary is an electronic version of (Popov, Simov and Vidinska 1998) extended with new words from the corpus. It covers the grammatical information of about 100 000 lexemes (1 600 000 word forms) and serves as a basis for the morphological analyzer.

The Gazetteers

Two basic lists with items, missing in the morphological dictionary, have been compiled with respect to their frequency:

1. Gazetteers of names. These consist of 15 000 items and include Bulgarian as well as foreign person names, international and national locations, organizations. The most frequent names are additionally classified according to three criteria: (1) grammatical (gender and number); (2) semantic - with respect to an extended SIMPLE core ontology (names for different types of locations, organizations, artifacts, persons' social roles etc.) and (3) ontological - some person names were connected with specific individuals in the world and thus some encyclopedic information was provided in addition to the semantic classification. All this information can be used for practical applications like Information Extraction or Retrieval, Data Mining, etc. In the process of the construction of the treebank we envisage to use it for agreement specification and semantic selection. Special attention is paid to the names of mountains and artifacts (books, films, broadcasts), because their internal agreement does not always coincide with the external one, which is needed for the sentence analysis.

2. Gazetteers of the most frequent abbreviations. They consist of 1500 acronyms and graphical abbreviations. The acronyms' extensions were mapped against the names (mostly organizations) and therefore, assigned the same semantic and grammatical label. In cases of idiosyncratic grammatical behaviour, the relevant patterns have been added as well.

3. Gazetteers of the most frequent introductory expressions and parentheticals. This is considered to be a step towards a basic list of collocations. They were classified according to their morphological type or behavior: verbal, adverbial, linking (for conjunctions), nominal (vocatives), idiomatic etc. We use them as an extended supplementary lexicon during the phase of the syntactic annotation.

The Valence Dictionary

It consists of 1000 most frequent verbs and their valence frames and it is based on a paper dictionary (see (Balabanova and Ivanova, 2002)). Each frame defines the number and the kind of the arguments and imposes morphosyntactic and semantic restrictions over them. The semantic restrictions over the arguments are extracted and matched against the SIMPLE core ontology. The frames of the most frequent verbs are compared to the corpus data and repaired if necessary (new frames are added, some of the existing frames are deleted or fine-grained). We envisage to enlarge the coverage of the dictionary with the help of some derivational means, such as the verb prefixes.

The Semantic Dictionary

Semantic information plays a crucial role in the process of parse discrimination on which the construction of our treebank depends. Thus, in order to support the selectional restrictions imposed by the valence dictionary and to facilitate its usage, we decided to compile a semantic dictionary along the guidelines of SIMPLE project. It is worth mentioning that we follow an extended variant of the SIMPLE core ontology. At the moment we are classifying the most frequent nouns with respect to the ontological hierarchy without specifying the synonymic relations between them. Up to now we have classified about 3 000 nouns. Recall that the named entities also have been classified with respect to the same ontology.

The main strategy we have adhered to in our work is the preparation of a minimum set of resources with substantial impact over the text archive.

According to the mentioned scope of BLARK our resources have the following gaps: (full) syntactic and semantic parsers, completed semantic annotations. At the same time we have created other resources which are obviously considered as suitable for a next step to BLARK: machine readable valency dictionary (see above), discourse patterns. This fact might slightly change the view on the limits between BLARK components and the components of more advanced LRs in the following sense: sometimes the creation of more advanced LRs can precede the compilation of more basic ones.

4. Conclusion and outlook

The intensive work within a project for creation of a treebank for a less-spoken language pays off in several ways:

4.1. Practical

It triggers the creation of LRs which still lack for the certain language. Consequently, the created set of LRs minimizes the work during the actual annotation of sentences within the treebank and ensures a high quality result.

Another advantage is that the developed resources and processing modules have a natural environment for intensive testing and improvement. This guarantees their appropriateness and adaptability for other NLP applications.

Also these LRs are a reliable basis for further development of the resources and processing modules included in BLARK.

4.2. Theoretical

It rises the question about the ways of LRs creation. It turns out that there are two ways: (1) starting from basic tasks and after their completion pursuing next-level tasks of complexity or (2) having in mind some more complex task and compiling all the other basic resources in order to adequately face it. We believe that the latter is an appropriate way for a less-spoken and less-processed language to come up with the state-of-the-art LRs in the natural languages.

The worked out methodology for the creation of basic language resources is implemented in the CLaRK system as reusable modules and can be parameterized to other languages as well.

5. References

Strik, Daelemans, Bimnnenpoorte, Sturm , De Vriend, Cucchiarini. 2002. Dutch HLT Resources: From BLARK to Priority lists. ICSLP-2002.

Elisaveta Balabanova and Krassimira Ivanova. 2002. *Creating a machine-readable version of Bulgarian valence dictionary: (A case study of CLaRK system application).* In: *Proc. of The First Workshop on Treebanks and Linguistic Theories.* Sozopol, Bulgaria.

Bimnnenpoorte, Cucchiarini, D'Halleweyn, Sturm and De Vriend. 2002. Towards a roadmap for Human Language Technologies: Dutch-Flemish experience, LREC-2002

Krassimira Ivanova and Dimitar Doikoff. 2002. *Cascaded Regular Grammars and Constraints over Morphologically Annotated Data for Ambiguity Resolution.* In: *Proc. of The Workshop on Treebanks and Linguistic Theories*. Sozopol, Bulgaria.

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

Petya Osenova and Sia Kolkovska. 2002. *Combining the named-entity recognition task and NP chunking strategy for robust pre-processing.* In: *Proc. of The Workshop on Treebanks and Linguistic Theories*. Sozopol, Bulgaria.

Petya Osenova and Kiril Simov. 2002. *Learning a token classification from a large corpus. (A case study in abbreviations).* In: *Proc. of the ESSLLI Workshop on Machine Learning Approaches in Computational Linguistics*, Trento, Italy.

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

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. pp 558–560.

Kiril Simov and Petya Osenova. 2001. *A Hybrid System for MorphoSyntactic Disambiguation in Bulgarian.* In: Proc. of the RANLP 2001, Tzigov chark, Bulgaria.

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