Abstract for Svetoslav Marinov, General Linguistics, University of Gothenburg

The interest in dependency grammar has clearly been on the rise in the last 10--15 years. The parsing community has seen a number of benefits in the use of and parsing with dependency-based representations of syntactic phenomena. On the one hand, dependency trees are much simpler than phrase-structure trees -- they contain exactly the same number of nodes as there are tokens in the sentence. The words in the sentence are linked by binary, asymmetric relations, the so called dependencies. In addition, every link is often labelled with the syntactic or semantic relation which holds between the two words. A tree labelled with such relations encodes the predicate-argument structure of a clause in a much straightforward way than phrase-structure trees. Finally, a number of parsing algorithms have been proposed whose running time is linear with respect to the number of tokens in a sentence.

The availability of syntactically annotated corpora has facilitated the use of supervised machine learning methods in the field of parsing. Data-driven parsers, which rely on corpora for the training stage can be acquired for any language for which such an annotated corpus exists. In addition, they are robust and efficient for the given task. Therefore in the present work we have used data-driven parsing with syntactic dependency representations for Bulgarian.

This thesis deals with three topics, with each part building on the findings of the previous one. First, we have chosen to adopt the framework of dependency grammar and apply it to syntactic phenomena in Bulgarian. We have identified dependency structures and relations which cover a number of phenomena in the language. However, the central issue is the proper encoding of the predicate-argument structure within the chosen framework. No other studies of Bulgarian have dealt with this issue from a dependency grammar point of view.

Secondly, once the crucial structures and relations have been identified, we have used those priciniples in the conversion of a constituency-based treebank of Bulgarian. Again we argue for the need of the proper transfer of the predicate-argument structure from a constituent tree into the dependency graph. This need reflects the nature of the underlying structures in the original treebank, where complex verb-phrases have a flat annotation. In addition, the principles we adopt lead to increase of non-projective structures compared to previous conversions of the treebank.

Finally, the performance of two data-driven parsers have been tested on the acquired treebank. The aim is not only to achieve state-of-the-art results for parsing Bulgarian, but what is more important -- to test the performance with respect to a higher number of non-projective structures and a verb argument structure which is more complex than previously created.

Key words: dependency parsing, Bulgarian, syntactic analysis, data-driven parsing, treebank conversion