Review of the Master Thesis

Author: Pranava Swaroop Madhyastha
Title: Exploring Higher Order Dependency Parsers
Charles University Supervisor: RNDr. Daniel Zeman, Ph.D.

The goal of the thesis is exploring the possibilities of using semantic features in dependency parsing, in particular higher-order (second and third order) features. The thesis consists of 8 sections. Sections 1 – 3 are introductory; they provide the necessary theoretical background and describe related work. Section 4 describes the data and tools used. The experiments are described in Section 5 and some more discussion is provided in Sections 6 and 7. Section 8 concludes the work. There is extensive bibliography as well as the lists of tables and figures. The attached CD contains the author's script and the freely redistributable parser software used in the experiments.

The thesis is well structured. It is written in passable English, though it would benefit from some more proofreading by a native speaker. For most part it is quite clear what has been done but there are exceptions. **Question:** Section 4.3 describes fine-grained word sense extraction. Is there a set of possible word senses for each language known in advance? Where does it come from? Or is this a sort of clustering the word senses based solely on context?

For the sake of reproducibility of the results it would be better if the author gave a more precise description of the data sets used (which section of the Penn Treebank / Prague Dependency Treebank). **Question:** Page 8 says that only projective structures are used. Is this to be read so that only projective sentences were selected from the respective treebanks? What is the expected behavior of the system if analogous experiments are done with non-projective data?

It would be also interesting to see how the parsers would perform if the data contained only machine-assigned feature values (tags), which is a more realistic setting than the gold-standard tags used by the author. Especially with regard to the Czech SEMPOS feature I am not aware of an existing tagger that could provide the values of this feature without having the surface dependency tree available.

A minor note on page 34: POSSGender is not the same as Gender in Czech.

Conclusion
The thesis provides a good survey on existing work in higher-order dependency parsing and augments it by author's own experimentation. It shows that the author has acquired a compendium of knowledge in the field. The experiments conducted would deserve a deeper analysis of their results; nevertheless, I believe that even in the current form it can be defended successfully, provided the author is able to answer the raised questions satisfactorily.

Jenštejn, August 30, 2011

RNDr. Daniel Zeman, Ph.D.
Institute of Formal and Applied Linguistics
Charles University in Prague