Review of Master Thesis

Author: Sibel Ciddi
Title: *Processing of Turkic Languages*
Supervisor: RNDr. Daniel Zeman, Ph.D.

The goal of the thesis is to investigate natural language processing problems specific to the agglutinative morphology of Turkic languages, as exemplified by the case of Turkish. The author experiments with TRmorph, a publicly available Turkish morphological analyzer, evaluates its coverage and explores the possibilities of its improvement. A significant portion of the work is devoted to the correct recognition of inflected multi-word expressions (MWE), which is interconnected with tokenization and morphological analysis.

The topic of the thesis is important and there is a potential benefit for the research community, as the resources for Turkish are relatively scarce, especially those that are publicly available. An extended lexicon will have immediate impact on processing Turkish; the (semi-)automatic *methods* of extending the lexicon will be reusable for other Turkic languages.

There are 118 pages, out of which roughly 40 describe the author's own contribution (chapters 5 to 7). The contribution lies in extending the lexicon with tens of thousands of lemmas, semi-automatically acquired by mining Wikipedia, Wictionary and other sources. Various date and time formats are newly recognized by the analyzer, and recognition of multi-word expressions (including multi-word named entities) at the tokenization level has been greatly improved. The author of the original TRmorph, dr. Çağrı Çöltekin, kept going on with the development of the tool while this thesis was being prepared, which made it somewhat difficult to come up with results that will still be novel at the time of the defense. One example is the guesser for unknown words that was not included in the original tool but appeared during the work on the thesis; nevertheless, in Section 5.3, Sibel Ciddi shows that the new guesser of TRmorph is too strong to be useful in practice; and she proposes her own guesser that obeys more constraints and does not overgenerate so much.

The author has demonstrated that she has a good knowledge of related work and literature. The experimentation is designed and described methodically and conscientiously. I especially like the Chapter 4 where a broad and thorough analysis of the performance of the pre-existing analyzer is given.

The thesis is written in grammatically good English with minimum typos. Comprehensibility could be improved by reducing long and complex sentences.

The attached DVD contains the electronic version of the text of the thesis, both the baseline analyzer TRmorph and the new analyzer TRmorph+, an installation and usage guide, and a selection of corpora that are mentioned in the thesis and their license permits redistribution. There is a reference to the documentation in the footnote at page 86 of the thesis; it could be made more visible if it was a chapter heading (or an appendix) rather than just a footnote. Also, the README subfolder in TRMORPH+ is a little misleading because it belongs to the baseline tool (TRmorph).

To summarize, I believe that the present thesis is a nice contribution to computational processing of Turkic languages, and the author has proven that she is able to conduct

independent research and present it, even if the form of the presentation could indeed be improved. I recommend the thesis for the defense.

Specific questions and comments

Page 94: What is "first stem selection"? Is the ordering predefined in which TRmorph returns stems? Or is it de facto random order?

Table 6.7: Could you explain the contents of the table? The table caption does not explain or re-explain what the individual data sets are, thus it is not clear what the numbers say and why one of them is bold (even though it is not the highest one).

Jenštejn, January 26, 2014

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