The thesis submitted by Miloš Ercegovčević aims at proposing a method for machine learning of phenomena centered around verbs at the syntax-semantics boundary. The bold goal would be to automatically learn complex structures like FrameNet or PropBank frames at an appropriate and automatically established level of detail and automatically annotate a given corpus with them. In supervised setting, this task is often called semantic role labeling (SRL).

After an introduction, Chapter 2 presents FrameNet and PropBank, two mainstream approaches to lexical semantics. Chapter 3 introduces a supervised and an unsupervised approach to SRL. The formal means that the author will use for SRL in the presented work, namely Latent Probabilistic Context-Free Grammar, is very briefly described in Chapter 4. Chapter 5 provides a broad discussion on supervised and unsupervised methods for various tasks such as finding verb classes, modifier roles, word classes or word sense. The proposed model is also introduced there. Experimental results are in Chapter 6, followed by the discussion and conclusion.

Unfortunately, there are rather serious problems with the text. On many places, the text is too succinct and too vague to be understandable. The mix of broad discussions and bold plans on one side and tersely commented formulas with unexplained symbols on the other side does not provide a clear picture of what is the task of the thesis and how the task is achieved. For instance: Does the author assume that he is given sentences with the verb and the set of modifiers already marked up and the task is just to assign roles to the modifiers? Or is the input text completely unannotated, so even the morphological analysis is part of the task? On several places, the text is simply unintelligible, e.g.: „Further one cannot prove or disprove our beliefs by themselves as they are to be learned with latent variables which can automatically adapt to the good or the properties of the same.“

The evaluation part is also critically insufficient. While, in my opinion, master theses should be self-contained, neither the input data, reference answers, nor the evaluation measure are described. All we are given is the reference to CONLL-2009 shared task. Moreover, the author uses a system by Pierre Nugues as a sub-procedure and replaces its outputs on verb arguments. The final Table 6.2 then provides the performance of the author's system but it also includes a better-performing system by Pierre Nugues. It is hard to avoid the impression that the author's proposal just damages the performance of the used system, but the truth is that the merit of the work simply cannot be judged with the given limited explanation.

To conclude, the presented thesis is unfortunately below the standard of M.Sc theses at Charles University. I recommend the thesis to be rejected and re-submitted after a significant improvement of the explanations throughout the text, especially in the description of the goal, the method chosen and also the experimental and discussion sections. Formal aspects of the text (grammar, typesetting) are of an acceptable quality.

Prague, August 22, 2012.

RNDr. Ondřej Bojar, Ph.D.
Charles University in Prague, ÚFAL