Review of Habilitation Thesis

Creating and Exploiting Annotated Corpora

by

Barbora Vidová Hladká

The thesis primarily collects conference papers dealing with three topics related to building and using annotated corpora for Czech language exploration. The text is supplied with an introductory part that motivates the work, stresses the author's contributions in the mentioned areas, and overviews her plans and the perspective of future research and development. Barbora Vidová Hladká clearly defines challenges tackled by her work and describes approaches realised solutions followed. The fundamental character of the corpora Barbora co-created and described in co-authored papers demonstrates the high level of knowledge and experience she has gained in recent years.

Although there are some significant contributions brought by the study of "up-translation" of the annotations used in the Czech Academic Corpus into the scheme of Prague Dependency Treebank (PDT), the major assets of Barbora's work in the domain of the academic corpus annotation can be seen in the creation of the corpus-based exercise book for Czech and the Czech legal text treebank. The selection and transformation of the PDT annotations into the system and diagrams used in Czech schools led to a unique resource that has an enormous potential to bridge the gap between the academic research and everyday pedagogical practice related to the Czech grammar teaching. It is a pity that this direction has not (yet) been further explored towards a nation-wide movement providing an alternative approach to teaching Czech grammar at primary schools (cf. the alternative math teaching approach by prof. Hejný). The creation of the morphologically and syntactically annotated Czech Legal Text Treebank then represented a pioneering work in the legal domain and laid the groundwork for future research.

The merits of Barbora's work in the field of the alternative annotation of corpus resources consist mainly in her contribution to the PlayCoref game, aiming at enlarging the gold data for co-reference resolution. It is highly appreciated that she helped to design novel game features aiming at the high quality of co-reference annotations. On the other hand, it would be beneficial to focus also on the annotation of difficult cases where simple mechanisms, such as Hobb's naïve algorithm for pronominal anaphora resolution (adapted for Czech), fail. This could help to identify sentences corresponding to the Winograd Schema in Czech. Unfortunately, it is not clear whether the planned future research would follow this direction.

Another strong aspect of the presented research lies in Barbora's pioneering work in the field of information extraction from Czech legal texts. The RExtractor system, which extracts domain-specific entities and some basic relations from Czech acts, paves the way towards more complex systems analysing Czech juridical and other legal documents. As in the other two areas discussed above, I also appreciate the clarity in delimiting the author's contribution to each particular research direction and publication specified in the thesis.

From the pedagogical point of view, it is also highly valuable that Barbora Vidová Hladká engaged and cooperated with a high number of students that achieved interesting results related to the resources created by her.

To summarize the above-mentioned points, I can ascertain that the reviewed habilitation thesis proves the author's excellent research track in the building and using annotated corpora and clearly demonstrates her potential for future scientific work in the area. I propose to accept it for the habilitation at Charles University in Prague.

Brno, January 5, 2020

Pavel Smrž