A reusable corpus needs syntactic annotations: the Prague Dependency Treebank

Eva Hajičová and Petr Sgall
Center for Computational Linguistics
Faculty of Mathematics and Physics
Charles University, Prague

e-mail: {hajicova,sgall}@ufal.mff.cuni.cz

The Prague Dependency Treebank (PDT, i.e. an annotated part of the Czech National Corpus) is conceived as a three-layer system of tags; the individual layers can be characterized as follows: (i) morphemic tagging capturing relatively disambiguated values of morphemic categories based on a full morphemic analysis of Czech; (ii) syntactic tags at the so-called analytical level, capturing the functions of individual word forms; in the analytical tree structures (ATSs), every word token and punctuation mark has a corresponding node and is analyzed as for its POS and morphemic value, as well as for the main syntactic functions ('analytical functors', 'afuns'); among the afuns, Subj, Obj, Adv are not classified in a more subtle way; (iii) syntactic tags at the tectogrammatical level (TGTSs) rendering the underlying (tectogrammatical) structure of the sentence, i.e., its syntactic structure proper (with a detailed classification of underlying syntactic functions).

In the sequel we focus on a brief characterization of the TGTSs and on issues that are specific for the PDT scenario and are crucial, especially from the linguistic point of view. These issues concern (i) the transition from ATSs to TGTSs, (ii) the assignment of the features of the information structure of the sentence (topic-focus articulation), and (iii) a tentative treatment of coreference relations. The TGTSs are based on dependency syntax; the tagging at this level is guided by the following principles: (a) a node of a TGTS represents an autosemantic (lexical) word; the correlates of synsemantic (functional, auxiliary) words are attached to the autosemantic words to which they belong; (b) in the cases of deletion in the surface shape of the sentence, further nodes are supplied into the TGTS to 'recover' a deleted word; (c) no non-projective structures are admitted in the TGTSs (they are supposed to be solved by movement rules between the ATS and the TGTS); (d) not only the direction of the dependence on the governing node (dependence to the left, dependence to the right) is taken into account, but also sister nodes are ordered (from left to right).