Universal Description of Morphology and Syntax of Natural Languages

Daniel Zeman

Institute of Formal and Applied Linguistics, Faculty of Mathematics and Physics, Charles University in Prague

Abstract: I will present Universal Dependencies, an international community project and a collection of morphosyntactically annotated data sets ("treebanks") for more than 100 languages. The collection is an invaluable resource for various linguistic studies, ranging from grammatical constructions within one language to language typology, documentation of endangered languages, and historical evolution of language. From the engineering perspective, UD treebanks serve as training data for automatic parsers that can be subsequently used to analyze previously unseen text. The parsed output is an intermediate representation between the input text and its underlying meaning. It is helpful in foreign language learning as well as for automatic extraction of semantic relations (answers to "who did what to whom"). I will thus discuss these semantic aspects in the last part of my talk; in particular, I will look at extensions of UD that have been proposed and that focus more on deep-syntactic and semantic relations expressed in natural language.