Robust Syntactic Annotation of Corpora and Memory-Based Parsing

Erhard W. Hinrichs

Seminar for Sprachwissenschaft, Abteilung Computerlinguistik

Eberhard-Karls-University Tübingen, Wilhelmstr. 113

D-72074 Tübingen ,Germany

email: eh@sfs.uni-tuebingen.de

Abstract

This talk provides an overview of current work in my research group on the syntactic annotation of the Tübingen corpus of spoken German and of the German Reference Corpus (Deutsches Referenzkorpus: DEREKO) of written texts.

Morpho-syntactic and syntactic annotation as well as annotation of function-argument structure for these corpora is performed automatically by a hybrid architecture that combines robust symbolic parsing with finite-state methods ("chunk parsing" in the sense Abney) with memory-based parsing (in the sense of Daelemans).

The resulting robust annotations can be used by theoretical linguists, who are interested in large-scale, empirical data, and by computational linguists, who are in need of training material for a wide range of language technology applications. To aid retrieval of annotated trees from the treebank, a query tool VIQTORYA with a graphical user interface and a logic-based query language has been developed. VIQTORYA allows users to query the treebanks for linguistic structures at the word level, at the level of individual phrases, and at the clausal level.