Cyril Belica, Marc Kupietz, Andreas Witt (Mannheim, Germany), Harald Lüngen (Gießen, Germany)

The Morphosyntactic Annotation of DEREKO: Interpretation, Opportunities, and Pitfalls

Abstract

The paper discusses from various angles the morphosyntactic annotation of DEREKO, the Archive of General Reference Corpora of Contemporary Written German at the Institut für Deutsche Sprache (IDS), Mannheim.

The paper is divided in two parts. The first part covers the practical and technical aspects of this endeavor. We present results from a recent evaluation of tools for the annotation of German text resources that have been applied to DEREKO. These tools include commercial products, especially Xerox' Finite State Tools and the Machinese products developed by the Finish company Connexor Oy, as well as software for which academic licenses are available free of charge for academic institutions, e.g. Thorsten Brants' Trigrams and Trees (TnT) and Helmut Schmid's Tree Tagger.

The second part focuses on the linguistic interpretability of the corpus annotations and more general methodological considerations concerning scientifically sound empirical linguistic research. The main challenge here is that unlike the texts themselves, the morphosyntactic annotations of DEREKO do not have the status of **observed data**, instead they constitute a theory- and implementation-dependent **interpretation**. In addition, because of the enormous size of DEREKO, a systematic manual verification of the automatic annotations is not feasible. In consequence, the expected degree of inaccuracy is very high, particularly wherever linguistically challenging phenomena, such as lexical or grammatical variation, are concerned. Given these facts, a researcher using the annotations blindly will run the risk of not actually studying the language, but rather the annotation tool or the theory behind it.

The paper gives an overview of possible pitfalls and ways to circumvent them, and discusses the opportunities offered by using annotations in corpus-based and corpus-driven grammatical research against the background of a scientifically sound methodology.