Detecting Errors in Corpus Annotation

Large corpora that are annotated with various types of linguistic annotation are central for computational linguistics and arguably also to theoretical linguistics. They play a crucial role as training and testing data for a wide range of natural language processing algorithms, and they provide access to natural examples relevant for creating and testing linguistic theories. At the same time, the "gold standard" annotations used for these purposes contain a significant number of errors, which have been shown to negatively affect both kinds of uses.

As a step towards addressing this situation, we discuss an automatic method for detecting errors in annotated corpora that is generally applicable to corpora with a wide range of annotation schemes. The approach, developed in collaboration with Markus Dickinson and Adriane Boyd in our DECCA project, is based on the idea that data recurring within a comparable context should be annotated the same way in all occurrences. Variation in the annotation within similar contexts thus is likely to be erroneous. We demonstrate the applicability of this variation n-gram method by illustrating that it can detect errors with high precision for a range of annotation types, including positional (part-of-speech), tree-based syntactic, discontinuous syntactic, and dependency annotation.