Detecting Errors in Annotated Corpora

This talk will cover two main methods for detecting errors in corpus annotation. The first method is the variation n-gram method from the DECCA project (Detection of Errors and Correction in Corpus Annotation). This method identifies recurring strings in text with varying annotation, in order to identify errors with high precision. I will discuss how it was developed for part-of-speech and syntactic constituency annotation before focusing more on syntactic dependency, semantic role, and alignment annotation types. I will also discuss some techniques that have been employed to increase the recall of the method.

The second method identifies anomalies in syntactic parse structures and has been explored mainly for constituency and dependency annotation. Syntactic rules are compared to the entire corpus of rules, and ones which are the most dissimilar to the other rules are flagged as anomalous. While the method lumps together annotation errors with other types of anomalies (ungrammatical sentences, odd annotation decisions), it has the benefit of being applicable to automatically-parsed data, as well as manually-annotated data. Time permitting, I will also mention initial forays into parse revision and active learning derived from this method.

(Note that Markus said he would complement the talk that Detmar Meurers gave last year, so there won't be much overlap between the two talks.)