Detecting Errors in Corpus Annotation
Detmar Meurers (Universität Tübingen)
3 mars 2011

Large corpora that are annotated with various types of linguistic information are central for computational linguistics and arguably also of relevance 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 developing and testing linguistic theories. Yet, 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, is based on the basic 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.

La présentation aura lieu dans le cadre du **LingLunch Paris Diderot**, organisé chaque jeudi à l'UFR de Linguistique de l'Université Paris Diderot.
Salle 4C92, 12:00 à 13:00; 175 rue du Chevaleret, 75013 Paris, M° Chevaleret.
`http://www.linguist.univ-paris-diderot.fr/linglunch.html`