Over the past two decades, there has been increasing use of linguistically annotated sentence collections, such as the Penn Treebank (PTB), for constructing statistically-based parsers across a variety of languages. Such Treebank-based parsers are trained on and exploit syntactic regularities in (manually-annotated) phrase structures - statistically interpolating where data is missing - in order to synthesize a most-likely parse when presented with novel (and pre-existing) sentences. Within this framework, one must rely on the promissory note that all necessary grammatical knowledge is encapsulated and statistically extractable from the Treebank corpus. In this talk we discuss the practical limits of this approach, cognitive implications for the problem of language acquisition, and finally, on-going Treebank work for a bilingual English/Arabic roboceptionist (Hala), a joint project with Carnegie Mellon University (Qatar and Pittsburgh).

Co-sponsored by the School of Modern Languages, Ivan Allen College Dean’s Office and the School of Interactive Computing

Georgia Institute of Technology
http://www.modlangs.gatech.edu/languages/linguistics