CONSTRAINT GRAMMAR AS A FRAMEWORK FOR PARSING RUNNING TEXT

Fred Karlsson

1. Outline
Grammars used in parsers are often directly inspired from autonomous grammar theory and descriptive practice that were not consciously designed. The grammars have been designed for English based on e.g. GB, FG, and LFG. We present a formalism to be used for parsing where the grammar statements are easier to read and more directly address some notorious parsing problems, especially ambiguity.

The descriptive statements, constraints, don't have the ordinary task of defining the notion 'correct sentence in X'. They are less categorical in nature, clearly tied to morphological features, and more directly geared towards the basic task of parsing: to infer structure from a stream of concrete tokens in a basically bottom-up mode.

Our tokens are morphologically analyzed words. The central idea is to maximise the use of morphological information for the purpose of pruning ambiguities and determining syntactic structure. Another central idea is to have all relevant structure assigned directly via lexicon, morphology, and simple mappings from morphology to syntax. The task of the constraints is basically to discard as many alternatives as possible, the optimum being a fully constrained sequence with one syntactic reading only.

A good parsing formalism should satisfy at least the following requirements: the constraints should be declarative rather than procedural, they should be clearly separated from the program code, the formalism should be language-independent, and it should be reasonably easy to implement (optimally as finite state automata).

5. Breaking up the problem of parsing
The problem of parsing running text may be broken up in five subproblems: preprocessing, morphological analysis, morphological disambiguation, context-dependent disambiguation, and assignment of surface-syntactic functions.

Real texts are full of idiosyncrasies in regard to headings, paragraph structure, punctuation, use of upper and lower case, etc. Such phenomena must be properly normalized. We assume that the existence of an adequate preprocessor is here taken for granted. We concentrate on morphological analysis, clause boundary determination, disambiguation, and surface-syntactic function assignment. Viewing the problem of parsing in terms of one or another of these angles clarifies many