NP annotation using the CLaRK system

Tamás Váradi

Linguistics Institute,
Hungarian Academy of Sciences
1068 Budapest, Benczúr u 33
varadi@nytud.hu

Keywords: partial parsing, NP annotation, cascaded regular grammars

The paper presents interim results of work in-progress to develop a robust NP annotation system based on finite state technology. The grammar uses the notions of cascaded regular grammar proposed by Abney (1995). The input text selected for the analysis was taken from *Heti Világgazdaság* on account of its extremely elaborate style containing numerous very complex NP's. The text was processed with the same technology developed for the Hungarian National Corpus as a result of which each word has its lemma and morphosyntactic description stored with it. The disambiguation process developed by Oravecz and Dienes (2002) was around 98 %.

Hungarian has some peruliarities which bars the straightforward adaptation of parsing techniques developed for other languages. Its word order, better to say, order of constituents is relatively free, while word order within constituents is bound. The dificulties making the automatic recognition of NP boundaries include the possible replacement of its head with its modifiers and the left recursive insertion of participles (progressive and perfect), which can bring in an unspecifiable and open-ended list of their modifiers.

The grammar was developed with the CLaRK system (Simov 2001), an XML based corpus processing software tool containing a finite state grammar compiler and a variety of other technologies, which altogether make this environment highly suitable for text annotation. The NP annotation rules rely heavily on the cascaded use of regular expressions defining increasingly complex NP's in two main stages. First, base NP structures containing noun heads are created. In the second stage more complex NP's produced through coordination and merge of possessive NP's are defined including those that have heads other than nouns.

The results of the analysis are tested against a hand-compiled corpus of a hundred sentences. Precision and recall figures are given both in terms of number of NP's and number of tokens involved. The numerical per-token recall and precision measure is quite acceptable and an intuitive evaluation of the parse output gives a better impression, considering the extremely elaborate NP structures that are successfully analyzed.