Corpus based examination of Hungarian conjunctions

Kata Gábor[1], Enikő Héja[1], Ágnes Mészáros[1]

[1] HAS, Research Institute for Linguistics, Department of Corpus Linguistics
{gkata,eheja,magnes}@corpus.nytud.hu

Keywords: conjunction, clause boundary, syntactic analysis, HNC

The goal of our work is to develop a parser for Hungarian language. The first step of the parsing process is to recognize phrases with a bound word order and to assign labeled tags to them. Then the phrases are provided with relevant syntactic and semantic features. The second phase consists of the identification of the components' syntactic roles. We take the argument structure of the verbal predicate as our starting point which is recognized on the grounds of the available lexical database. We cannot draw on word order while selecting possible argument phrases: we want to mark phrases that meet the subcategorization requirements of the verb if they are in the same *clause* as their possible governor. This means, however, that matching of the complement structure must be preceded by insertion of *clause boundaries.* What we regard as a clause boundary is determined by the chosen frame of analysis. Since our main purpose is the identification of verbal arguments and the domain within which we are looking for them is delimited by clause boundary, it is quite expedient to accept the hypothesis that *there is always a clause boundary in a sentence between two finite verbs.* In our talk we explain how conjunctions can be used to identify the exact location of the clause boundary. Although we cannot suppose to find a conjunction at each clause boundary, it is worth examining whether the presence of a conjunction between two finite verbs implies that there is a clause boundary. If we define conjunction as a POS that always ties components of the same type, then our task is to recognize cases when it has clauses on both sides. If the segmentation of clauses takes place after building the phrases with a bound word order (that may contain conjunctions), we would need to investigate only those conjunctions which are not within a phrase.

We noticed that conjunctions occupy different positions with respect to the clause boundary. Some of them follow it always immediately, others follow the first phrase, and several can occur almost anywhere. After the corpus-based examination of 71 conjunctions, we came to the conclusion that in most cases even the information concerning their potential position in the clause is not sufficient to assign clause boundaries correctly. It is because of the different distribution of conjunctions and homonymy. In our talk we present results concerning distribution of conjunctions based on data provided by the HNC, and we propose a grouping which might be useful for assigning clause boundaries.