Automatic F–structure Annotation of Treebank Trees

Anette Frank

— Abstract —

We describe a method that automatically induces LFG f–structures from treebank tree representations, given a set of f–structure annotation principles that define partial, modular c– to f–structure correspondences in a linguistically informed, principle-based way.

This work extends the approach in van Genabith, Sadler and Way (1999a,b,c) where f–structure annotation of treebanks is driven by manual annotation of treebank-extracted PS rules. In this paper we present a method for automatic f-structure annotation of treebank trees. A rule-based approach is explored, in parallel, by van Genabith, Sadler and Way.

In our method we build on a correspondence-based view of the LFG architecture, where annotation principles define ϕ–correspondences directly in terms of ϕ–projection constraints, relating partial (possibly non-local) c–structure fragments to their corresponding partial f–structures. Application of these annotation principles to treebank trees directly induces the f–structures. Due to the disambiguated tree input, the resulting f–structures require only minimal manual disambiguation.

To define and process f–structure annotation principles, we make use of an existing term rewriting system, originally designed for transfer-based machine translation. The method is fully automated, and inherently robust. It yields partial, unconnected f–structures in the case of missing annotation rules.

We present the results of a first experiment where we apply this method to two sections of the Susanne treebank. The experiment is designed to measure to which extent the partial c– to f–structure correspondences encoded in annotation principles scale up, by applying them to previously unseen tree configurations.

We finally extend the model to *selective* filtering of remaining ambiguities, using lexical subcategorization information, in conjunction with an OT-based constraint ranking mechanism for ambiguity filtering (cf. Frank et al. 1998, 2000).