# From Phrase Structure to Dependencies, and Back

Tylman Ule and Sandra Kübler
Seminar für Sprachwissenschaft + SFB 441
Universität Tübingen, Germany
{ule,kuebler}@sfs.uni-tuebingen.de

Transforming constituent-based annotation into dependency-based annotation has been shown to work for different treebanks and annotation schemes (e.g. Lin (1995) has transformed the Penn treebank, and Kübler and Telljohann (2002) the *Tübinger Baumbank des Deutschen* (*TüBa-D/Z*)). These ventures are usually triggered by the conflict between theory-neutral annotation, that targets most needs of a wider audience, and theory-specific annotation, that provides more fine-grained information for a smaller audience. As a compromise, it has been pointed out that treebanks can be designed to support more than one theory from the start (Nivre, 2003). We argue that information can also be added to an existing annotation scheme so that it supports additional theory-specific annotations. We also argue that such a transformation is useful for improving and extending the original annotation scheme with respect to both ambiguous annotation and annotation errors. We show this by analysing problems that arise when generating dependency information from the constituent-based *TüBa-D/Z*.

*TüBa-D/Z* is a treebank of German newspaper texts from 'die tageszeitung' using the Topological Field (TopF) model for the syntactic descriptions of sentences (Höhle, 1986), combined with the annotation of phrase structure and grammatical functions. The conversion into dependencies is based on the head/non-head information of the grammatical functions (for more details cf. Kübler and Telljohann, 2002). The phrase structure annotation shown in figure 1 is transformed accordingly into the dependency representation in figure 2.

The conversion of *TüBa-D/Z* into dependency structure revealed that some constituent-based annotations carried less information than intended. One such area is the distinction between coordination and elliptical structures: both phenomena used to be represented in *TüBa-D/Z* by nodes that only have children with empty edge labels. This annotation made it difficult and sometimes impossible to distinguish coordinations from elliptical structures. In the tree in figure 1, it was very hard to tell whether the last noun phrase (NX) in the middle field (MF) is an elliptical noun phrase or a split-up conjunct. To overcome these problems, a new edge label (KONJ) has been introduced representing conjuncts in all coordinations. Another area where the introduction of

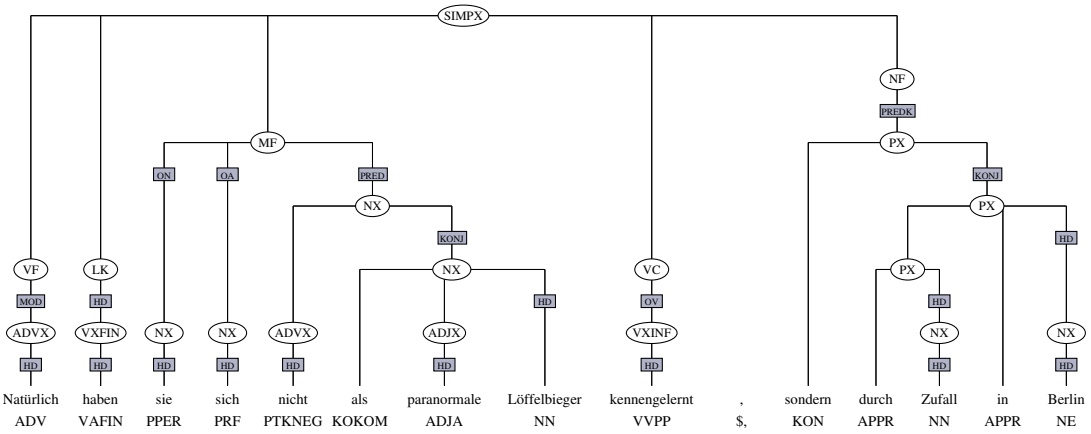


Figure 1: Split up conjuncts resemble ellipsis without edge label KONJ

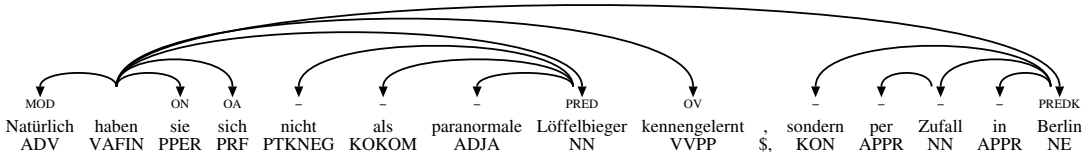


Figure 2: The dependency representation of the tree shown in figure 1.

this label helped to provide sufficient information is complex coordinations, where it turned out to be problematic to distinguish conjuncts from multi-word conjunctions.

In conclusion, we have shown that systematically converting *TüBa-D/Z* into a dependency representation is not only feasible but can be helpful to improve the original annotation.

# References

Höhle, T. (1986). Der Begriff 'Mittelfeld', Anmerkungen über die Theorie der topologischen Felder. In *Akten des Siebten Internationalen Germanistenkongresses 1985*. Göttingen, Germany.

Kübler, S. and H. Telljohann (2002). Towards a dependency-oriented evaluation for partial parsing. In *Proceedings of Beyond PARSEVAL*. Las Palmas, Gran Canaria.

Lin, D. (1995). A dependency-based method for evaluating broad-coverage parsers. In *Proceedings of IJCAI-95, Montreal, Quebec, Canada*.

Nivre, J. (2003). Theory-supporting treebanks. In J. Nivre and E. Hinrichs, eds., *Proceedings of Treebanks and Linguistic Theories*. Växjö Univ. Press, Växjö, Sweden.