From Phrase Structure to Dependencies, and Back

Tylman Ule and Sandra Kübler
January 30, 2004

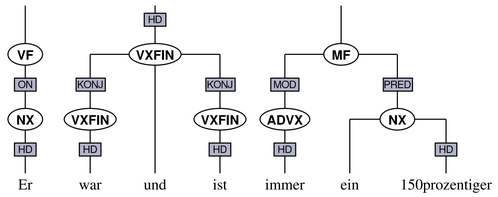
Deutsche Forschungsgemeinschaft
DFG
Sonderforschungsbereich 441
Linguistische Datenstrukturen

Eberhard Karls Universität Tübingen
Seminar für Sprachwissenschaft
Abteilung Computerlinguistik

Overview

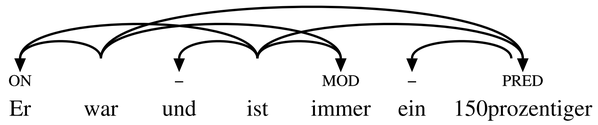
Describing linguistic information in a corpus requires choosing a certain kind of representation. This representation guides annotators when they add linguistic interpretation to a treebank. We highlight cases in the German treebank *Tübingen Treebank of Written German* (*Tübinger Baumbank – Deutsch / Zeitung*: TüBa-D/Z [Telljohann et al., 2003]) where switching the representation of linguistic annotation proved useful not only for obtaining a new target representation, but also for detecting weak spots in the original annotation scheme.

Topological Fields, Phrase Structure, and Dependencies



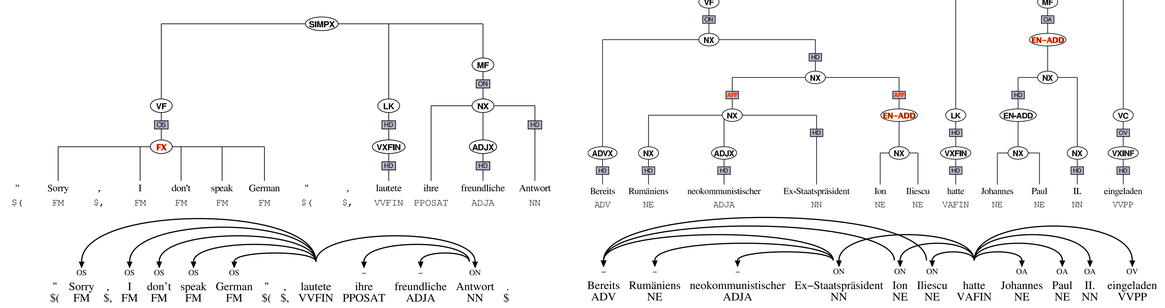
- Topological fields provide a distributional clustering of constituents in a clause
- Phrase structure describes constituents inside of fields
- Grammatical functions relate constituents and verbal parts in clauses

- Heads need to be determined for assigning dependency relations [Lin, 1998]
- Different kinds of phenomena without heads that are difficult to spot used to be encoded similarly in TüBa-D/Z
- Coordinations may result in multi-headed constructions
- Conjunctions are assigned to following (preceding) head



Appositions, Proper Names, and Foreign Material

- Sister appositions share grammatical functions
- No heads can normally be determined in proper names and in foreign material



References

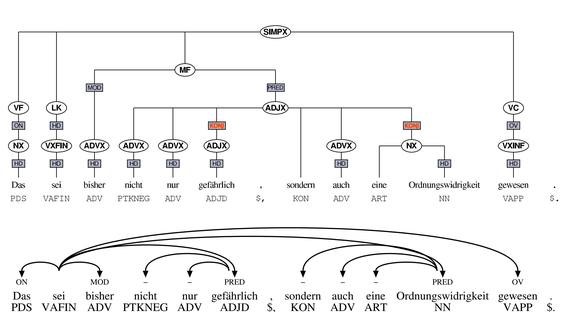
Dekang Lin. A dependency-based method for evaluating broad-coverage parsers. *Natural Language Engineering*, 4(2):97–114, 1998.

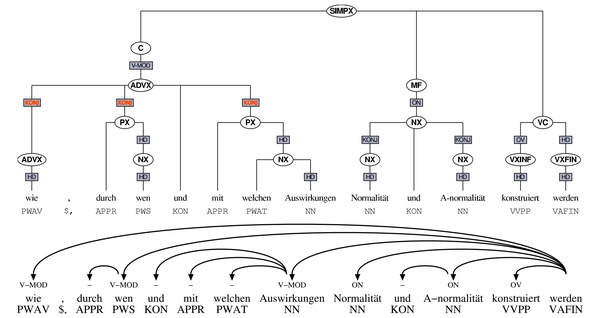
Heike Telljohann, Erhard W. Hinrichs, and Sandra Kübler. *Stylebook for the German Treebank of Written German (TüBa-D/Z)*. Seminar für Sprachwissenschaft, Universität Tübingen, Tübingen, December 2003.

Coordinations

Multipart Conjunctions

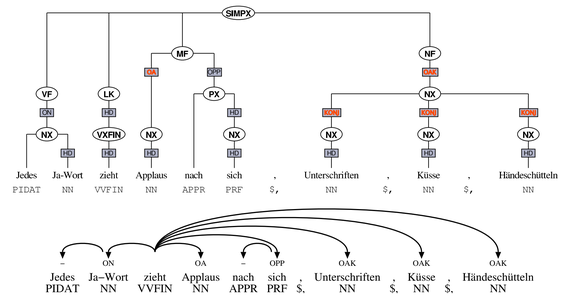
- Conjuncts are often coordinated by more than one word
- Frequently, different types of conjuncts appear in coordinations

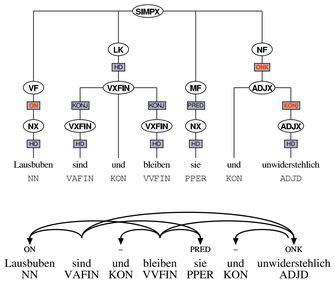




Coordinations Divided by the Verbal Bracket

- Coordinate terms are marked explicitly
- Category identity alone often proves to be too weak





Ellipses

- Ellipses are still encoded as HD-less constituents — co-indexation with elided head will be necessary for retrieving dependencies