From Phrase Structure to Dependencies, and Back

Tylman Ule and Sandra Kübler

January 30, 2004

Overview

Describing linguistic information in a corpus requires choosing a certain kind of representation. This representation guides annotators when they add linguistic information to a markup. We highlight cases in the German branch. Tübingen Textbank of Written German (TübText) - Deutsch / Zeitung: TübText D2/2 [Telthau et al., 2003] where enriching 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 constellations
- Phrase structure describes constituents inside of fields
- Grammatical functions relate constituents and verbal parts in clauses

Appositions, Proper Names, and Foreign Material

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

Coordinations

Multi-part Conjunctions

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

Coordinations Divided by the Verbal Bracket

- Coordinate terms are marked explicitly
- Category identity alone is often present to be tax weak

Ellipses

- Ellipses are still encoded as clause conditions - no information with ellided head will be necessary for retrieving dependencies

References

Heike Telthau, Rüdiger W., Hinz, and Sandra Kübler, Stylebank for the German Textbank of Written German (TübText D/2), Seminar für Sprachwissenschaft, Universität Tübingen, Tübingen, December 2003.