Treebank Data as Linguistic Evidence – Coordination in TüBa-D/Z*

Heike Zinsmeister
SFB 441 – University of Tübingen, Germany
heike.zinsmeister@uni-tuebingen.de

1 Motivation

Linguists use treebanks as resource for collecting evidence of phenomena which cannot be easily recovered from data that is annotated at word level only, this includes collecting quantitative data, getting non-categorical information such as heaviness or finding natural sounding counter examples[1] (e.g. Uszkoreit et al. (1998); Arnold et al. (2000); Bresnan et al. (to appear))[2]. Tools such as TIGERSearch allow us easy access to the encoded information.[3] This poster presents work on the *Tübinger Baumbank deutscher Zeitungssprache* (Tüba-D/Z). It describes the encoding of coordination phenomena in the treebank and gives a qualitative and quantitative survey.

2 The TüBa-D/Z Treebank

It is a corpus of newspaper texts which currently comprises about 22 000 sentences (more than 381 000 tokens) taken from the *Wissenschafts-CD* of 'die tageszeitung' (taz). The annotation combines information on inflectional morphology, part of speech, phrase structure (or rather recursive chunking), grammatical dependencies and topological fields. In addition, it includes marking of named entities and annotation of anaphoric and coreference relations (cf. Hinrichs et al. (2004)).

*Slightly revised version, 07.02.2006

[1] Note the objections of Pullum (2003) against corpus examples for illustration or educational means.

[2] The latter two references do not use treebanks as such but reconstruct equivalent syntactic information. See also specialised conferences such as the annual 'Treebanks and Linguistic Theory'.

[3] TIGERSearch (www.ims.uni-stuttgart.de/projekte/TIGER/TIGERSearch/) offers a user-friendly interface and a query option that is based on graphical tree structures which are then translated to the query language.

2.1 Coordination in TüBa-D/Z

To ensure reusability the annotation of the treebank is as theory-neutral as possible and a flat analysis is adopted for coordination. In the default case the conjuncts of a coordination are dominated by a common mother node of the same category. To mark the hierarchical indeterminacy there is no head function specified but all conjuncts are equally related to their mother by means of the functional label *KONJ*(unkt). The conjunction is marked by the part-of-speech label *KON* and is connected to the common mother node by a default 'non-head' edge. In *coordination of unlike categories* the left-most conjunct by definition determines the category of the whole coordination, e.g. (abridged example)

(1) Sie wird [$_{NX}$ [$_{NX}$ Schwimmeisterin] und [$_{ADJX}$ weise]].
'She will become a swimming champion and (she will become) wise.'

TüBa-D/Z adopts a chunk approach for verb phrases: arguments and verbal or sentential modifiers are immediately dominated by nodes of topological fields. Their functions however are encoded in specific edge labels. Coordination below sentence level is modelled as coordination of (groups of) topological fields (marked by means of special node labels *FKONJ* and *FKOORD*). The same strategy is employed in cases of *gapping* and *deletion*. TüBa-D/Z does neither use empty categories (such as the Penn Treebank) nor crossing branches (such as the TIGER Treebank). Discontinuous structures are to be recovered by means of specific functional labels. A dislocated conjunct in a *split-up coordination* for example is explicitly marked, e.g. the second conjunct (ONK) of a split-up coordinate nominative subject (ON):

(2) [$_{(ON)}$ *Krieg und Frieden*] sollte Pflichtlektüre werden [$_{(ONK)}$ oder Goethes *Faust*].
'War and Peace or Goethe's Faust ought to become mandatory reading.'

2.2 Samples of a Quantitative Survey

8 133 sentence (36.7%) involve a *coordination structure* (marked by the edge label KONJ). A query for *split coordination* results in 88 matches some of which require further discussion of their analysis. A search for *asymmetric coordination* of the type *Subject Gap Fronted* construction (see Höhle (1983) and Frank (2002) among others) provides 82 true positives[4], e.g.

(3) Da [stehen die Zuhörer auf] und [applaudieren herzlich].
'Then the audience stood up and applauded warmly.'

In the case of *coordinate subject NPs* it is interesting to investigate the number marking on the subject conjuncts and the finite verb; there is a default pattern: if the conjunction

[4]The examples were manually checked.

is *und* ('and'), coordinated singular noun phrases agree with a plural verb. On the other hand, if the conjunction is *oder* ('or') the coordinate singular nouns agree with a singular verb (cf. Reis (1979)). The interesting cases are those that do not follow the default: 55 matches for *und*-coordination and singular verb were found, 24 out of which turn out to be true positives.[5] Only eight matches are found of *oder*-coordination of singular NPs plus plural verb, three of which are relevant examples.

(4) [Kommunikation$_{SG}$ und Mobilität$_{SG}$] wird$_{SG}$ dadurch extrem schwierig.
'Communication and mobility become extremely difficult due to this.'

(5) [Vulkan$_{SG}$ oder Hurrikan$_{SG}$] sorgen$_{PL}$ für Irritation in der gottgewollten Ordnung.
'Volcanoes and hurricanes create irritation in God's ordering (of the world).'

3 Conclusion

A treebank is a unique resource for linguistic evidence. It allows us to find examples and collect frequency data on complex syntactic phenomena. We presented a qualitative and quantitative survey on coordination phenomena encoded in the TüBa-D/Z Treebank.

References

Arnold, J., T. Wasow, A. Losongco, and R. Ginstrom (2000). Heaviness vs. newness: the effects of complexity and information structure in constituent ordering. *Language*, **76**:28–55.

Bresnan, J., A. Cueni, T. Nikitina, and H. Baayen (to appear). Predicting the dative alternation. In *Royal Netherlands Academy of Science Workshop on Foundations of Interpretation*. Corpus study.

Frank, A. (2002). A (Discourse) Functional Analysis of Asymmetric Coordination. In *Proceedings of the LFG02 Conference*. CSLI Publications, Athens.

Hinrichs, E., S. Kübler, K. Naumann, H. Telljohann, and J. Trushkina (2004). Recent Developments in Linguistic Annotation of the TüBa-D/Z treebank. In *Proceedings of TLT04*.

Höhle, T. (1983). Subjektlücken in Koordinationen. Unpublished MS., Tübingen.

Pullum, G. K. (2003). Corpus fetishism. Language Log, Nov. 16, 2003.

[5]The survey was pursued when about 69% of the sentences included morphological information.

Reis, M. (1979). Ansätze einer realistischen Grammatik. In *Befund und Deutung. Zum Verhältnis von Empirie und Interpretation in Sprach- und Literaturwissenschaft*, pp. 1–21. Niemeyer, Tübingen.

Uszkoreit, H., T. Brants, D. Duchier, B. Krenn, L. Konieczny, S. Oepen, and W. Skut (1998). Studien zur performanzorientierten Linguisitk. Aspekte der Relativsatzextraposition im Deutschen. CLAUS Report 99, Universität des Saarlandes.