Recent Developments in Linguistic Annotations of the TüBa-D/Z Treebank*

Erhard Hinrichs, Sandra Kübler, Karin Naumann, Heike Zinsmeister *{eh,kuebler,knaumann,zinsmeis}@sfs.uni-tuebingen.de*
Seminar für Sprachwissenschaft, Universität Tübingen
Heike Telljohann *Dr.Heike.Telljohann@t-online.de*
Julia Trushkina *20215770@puknet.puk.ac.za*

Corpus

The data is taken from daily issues of the German newspaper 'die Tageszeitung' (taz) currently ranging from May 3rd to May 7th 1999 as well as April 30th 1999. It is semi-automatically annotated using the *annotate* tool (Brants & Plaehn 2000). Release 2 (May 2005) comprises 22.087 sentences with 381.565 tokens. The creation of the TüBa-D/Z treebank was originally funded by the Kompetenzzentrum für Text- und Informationstechnologie KIT, Stuttgart & Tübingen. The recent developments originated in the context of the A1 project of the Sonderforschungsbereich 441, Tübingen.

Annotation scheme

The scheme was adapted from the annotation of the VERBMOBIL treebank for spoken German (Stegmann et al. 2000) to accommodate the characteristics of written text (Telljohann et al. 2003). It is based on a context-free backbone. The annotation originally comprised information on:

- grammatical functions
- syntactic constituency
  - lexical level (part of speech tags)
  - phrasal level (partial parsing analysis)
  - level of topological fields
  - clausal level

Additions to previous layers of annotation

- lexical level: inflectional morphology
- phrasal level: (complex) named entities
- semantic level: anaphoric and coreference relations

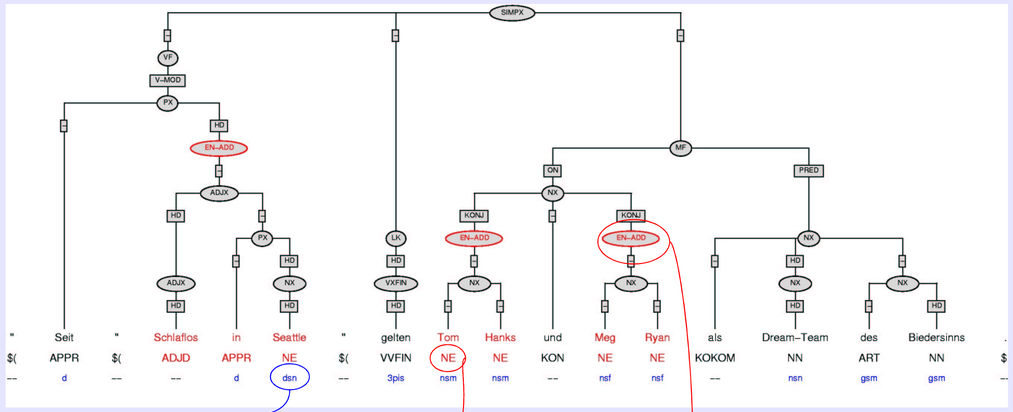
Morphological annotation

An initial set of possible analyses is provided by the morphological analyzer for German developed by the Xerox Research Centre Europe (XRCE). It results in an overall ambiguity rate of 5.8 analyses per token. An automatic rule-based disambiguation (Trushkina 2004) reduces the ambiguity rate to 1.91 analyses per token and provides full disambiguation for 70 % of all tokens. The remaining ambiguous analyses are manually disambiguated. Finally the morphological information is incorporated into TüBa-D/Z.

Lexical tokens	Feature combination
nouns, adjectives, determiners, non-personal pronouns, prepositions with incorporated articles	case, number, gender
prepositions, postpositions	case
personal pronouns	case, number, gender, person
imperative verbs	person, number
other finite verbs	person, number, mood, tense
truncated words	number, gender

Feature	Values
case	n (nominative), g (genitive), d (dative), a (accusative), * (underspecified)
gender	m (masculine), f (feminine), n (neuter), * (underspecified)
number	s (singular), p (plural), * (underspecified)
mood	i (indicative), k (subjunctive)
person	1 (first), 2 (second), 3 (third)
tense	s (present), t (past)

*) The poster is based on work described in Hinrichs et al. (2004)



clausal level
'SIMPX' = simplex clause
etc.

level of topological fields
'VF' = initial field
'LK' = left sent. bracket
'MF'= middle field
etc.

lexical level
morphology 'dsn' = dative singular neuter
part of speech 'NE' = proper noun

phrasal level
'PX' = prepositional phrase
'NX' = noun phrase
etc.

grammatical functions
'HD' = head
'PRED' = predicate
etc.

Annotation of named entities

Names consisting of one lexical element are PoS-tagged as NE if they belong to one of the categories of proper nouns defined in the STTS guidelines (Schiller et al. 1995). Otherwise they are tagged according to their distribution and assigned the additional node label EN-ADD (e.g. names of products such as 'Opel', STTS-tag: NN, or compounds which consist of NE+NN such as street names like 'Auguststraße', STTS-tag: NN).
Complex named entities are also identified by EN-ADD. If the named entity spans only part of a more complex phrase the syntactic structure is not altered. A secondary edge EN is employed to make the named entity recoverable.

The annotation is conservative and monotonic:

- it respects all syntactic boundaries that have been imposed on the elements of named entity expressions by existing layers of syntactic annotation
- it has full compliance with the STTS labeling

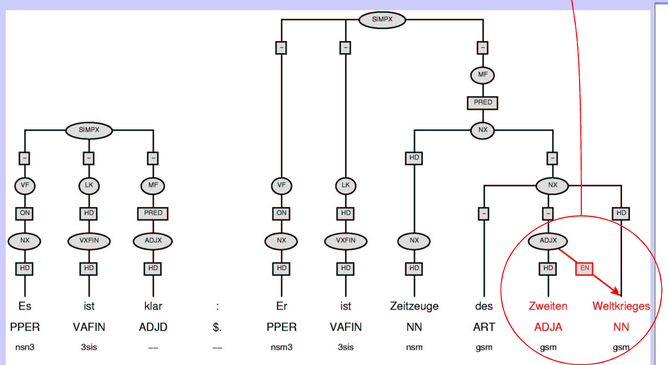
Anaphoric and coreference relations

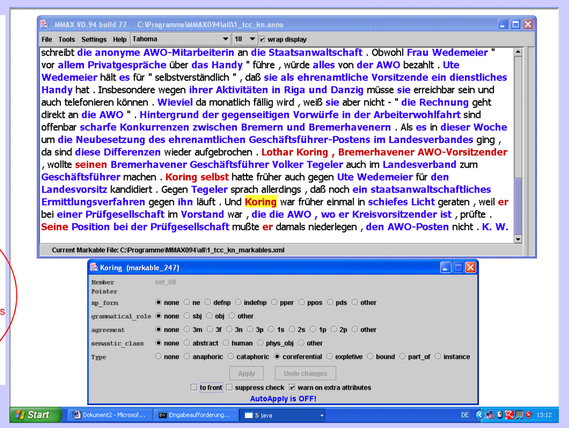
We annotate anaphoric and coreferential expressions referring to a nominal or pronominal antecedent. The markables, i.e. those entities in the text between which the relations to be annotated can possibly occur, are extracted automatically from TüBa-D/Z.

- potential anaphoric and coreferential expressions:
  - definite NPs and pronouns (personal, relative, reflexive, reciprocal, demonstrative, indefinite, possessive)
- the linking relations are annotated in MMAX
  - coreferential: two NPs refer to the same extralinguistic referent
  - anaphoric: referring back to a contextual antecedent
  - other relations
    cataphoric, bound, part-of, instance, expletive

Ex.: **[**Der neue Vorsitzende der Gewerkschaft Erziehung und Wissenschaft**]** heißt **[**Uli Thöne**]**. **[**Er**]** wurde gestern mit 217 von 255 Stimmen gewählt.

coreferent — anaphoric





XML-representation

```
<sentence>
   <node cat="SIMPX" func="--" parent="0">
     <node cat="VF" func="-">
       <node id="s11976n513" cat="NX" func="OA">
         <node cat="NX" func="APP">
           <word form="Ihre" pos="PPOSAT" morph="asf" func="-"/>
           <word form="Schulkameradin" pos="NN" morph="asf" func="HD"/>
         </node>
         <node cat="EN-ADD" func="APP">
           <node cat="NX" func="-">
             <word form="Cassie" pos="NE" morph="asf" func="-"/>
             <word form="Bernall" pos="NE" morph="asf" func="-"/>
           </node> </node> </node> </node>
     <node cat="LK" func="-">
       <node cat="VXFIN" func="HD">
         <word form="fragten" pos="VVFIN" morph="3pit" func="HD"/>
       </node> </node>
     <node cat="MF" func="-">
       <node cat="NX" func="ON">
         <word form="sie" pos="PPER" morph="np*3" func="HD"/>
           <car type="ana" ante="s11975n517"/>
           </word> </node> </node>
     <word form="," pos="$," morph="--" func="--" parent="0"/>
     <node cat="NF" func="-">
       <node cat="SIMPX" func="OS">
         <node cat="C" func="-">
           <word form="ob" pos="KOUS" morph="--" func="-"/>
         </node>
         <node cat="MF" func="-">
           <node cat="NX" func="ON">
             <word form="sie" pos="PPER" morph="nsf3" func="HD">
               <car type="ana" ante="s11976n513"/>
             </word> </node>
...
         <node cat="VC" func="-">
           <node cat="VXFIN" func="HD">
             <word form="glaube" pos="VVFIN" morph="3sks" func="HD"/>
           </node> </node> </node> </node> </node>
</sentence>
```

blue = morphology
red = named entities
green = anaphora and coreference

Availability

- license: free of charge for scientific use (you need a 'taz' license), see **http://www.sfs.uni-tuebingen.de/en_tuebadz.shtml**
- release 1 (December 2003): 15.260 sentences, 266.441 tokens, syntactic and functional annotation, named entities
- release 2 (May 2005): 22.087 sentences, 381.565 tokens, morphologic, syntactic, and functional annotation, named entities

Related work

The Tübingen Partially Parsed Corpus of Written German (TüPP-D/Z) comprises more than 200 million tokens of automatically annotated text (Müller 2004, Ule 2004). See **http://www.sfs.uni-tuebingen.de/en_tuepp.shtml**

References

- annotate: http://www.coli.uni-sb.de/sfb378/negra-corpus/annotate.html
- Brants, T. & O. Plaehn. 2000. Interactive Corpus Annotation. In Proceedings of *LREC 2000*.
- Hinrichs, E., S. Kübler, K. Naumann, J. Trushkina. 2004. Recent developments in Linguistic Annotations of the TüBa-D/Z Treebank. In Proceedings of TLT 2004.
- KIT: http://www.ims.uni-stuttgart.de/projekte/KCL/
- MMAX: http://mmax.eml-research.de
- Schiller, A., S. Teufel & C. Thielen. 1995. Guidelines für das Tagging deutscher Textcorpora mit STTS. Technical report. Stuttgart & Tübingen.
- Stegmann, R., H.Telljohann & E. Hinrichs. 2000. Stylebook for the German Treebank in Verbmobil. Technical report 239.
- Telljohann, H., E. Hinrichs, S. Kübler. 2003. Stylebook for the Tübingen Treebank of written German (TüBa-D/Z). Technical report. Tübingen.
- Trushkina, J. 2004. Morpho-Syntactic Annotation and Dependency Parsing of German. PhD thesis. University of Tübingen.
- XRCE morphological analyzer: http://www.xrce.xerox.com/competencies/content-analysis/demos/german