

# **Introduction to Corpus Resources, Annotation and Access: *Syntactic Annotation***

---

Sabine Schulte im Walde  
Universität Stuttgart

*Foundational Course*  
Departament de Traducció i Filologia  
Universitat Pompeu Fabra  
April 16-20, 2007

# Types of Syntactic Information



# Treebanks

---

- **Treebank**: a linguistically annotated corpus that includes some grammatical analysis beyond the part-of-speech
- Treebank vs. parsed corpus:  
strict manual annotation or post-editing
- Treebank information depends on linguistic theory
- Common treebanks:
  - » Penn Treebank (English), WSJ, Brown, Switchboard
  - » TIGER Treebank (German), newspaper texts
  - » TüBa Treebank (German), written and spoken texts

# Treebanks and Linguistic Theory

---

Three main kinds of annotation:

- **Constituent structure**
  - » Lancaster Parsed Corpus
  - » Penn Treebank 1 (skeletal parsing)
- **Dependency structure**
  - » Prague Dependency Treebank (analytical level)
  - » METU Treebank
- **Theory-specific annotation**
  - » Prague Dependency Treebank (tectogrammatical level: Functional Generative Description)
  - » BulTreebank, LinGo Redwoods Treebank (HPSG)
  - » CCG-Bank (Combinatory Categorical Grammar)

# Trebanks and Linguistic Theory

---

## Hybrid approaches

- combine constituents with functional/dependency information
  - » SUSANNE
  - » Penn Treebank II
  - » Penn Chinese Treebank
  - » NEGRA / TIGER Treebank
  - » TüBa Treebanks
  - » ARBORETUM

# Constituent Structure

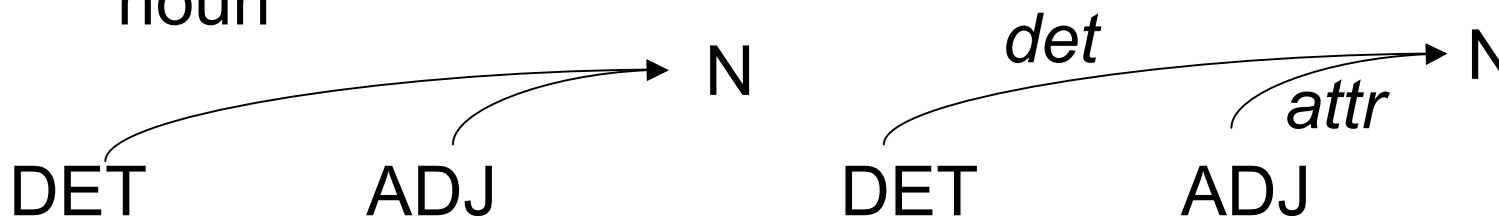
---

- ‘Bracketing’: sentences consist of hierarchically embedded subparts → **constituents**
  - » strings of words that belong together
  - » constituency tests:  
substitution, movement, stand-alone test, ...
- Part-whole relations
  - » e.g. an NP **consists of** determiner, adjective and noun  
[NP [DET ] [ADJ ] [N ]]

# Dependency Structure

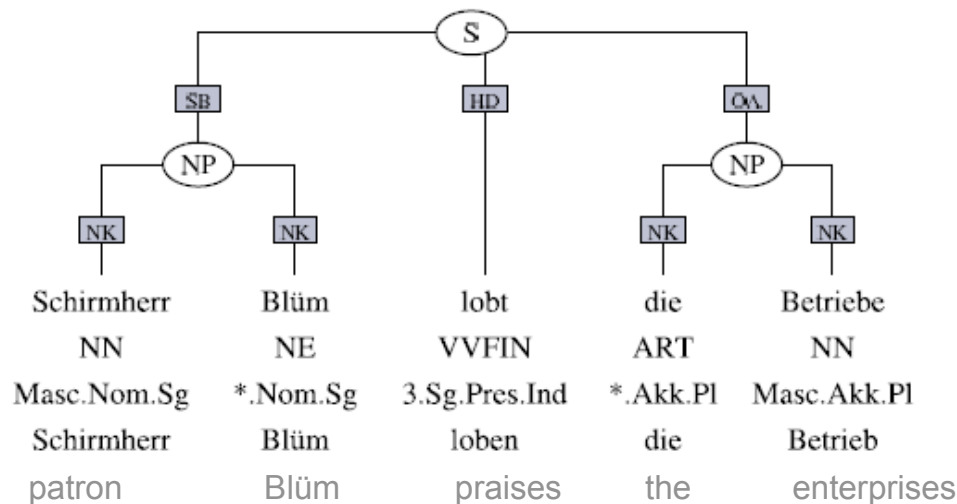
---

- First comprehensive theory: Lucien Tesnière (1959)
- Sentences consist of hierarchically structured asymmetric, binary **relations between word forms**  
→ dependency relations ('connexions')
  - » governor, dependent(s)
  - » closely related to functional analysis
- Relations
  - » e.g. determiner and adjective are **subordinated to the noun**



# Hybrid Models

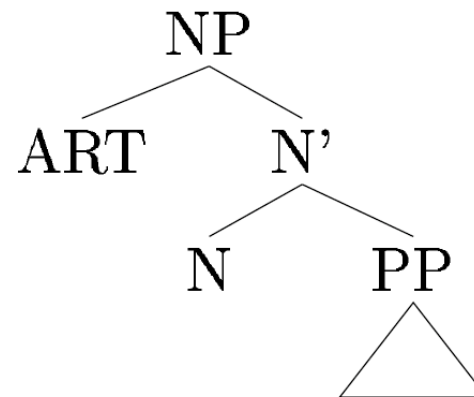
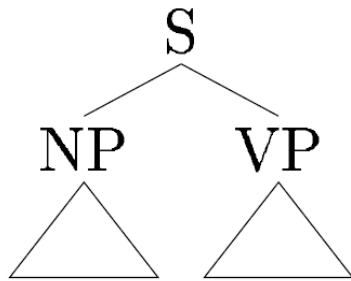
- Combine constituent and functional (dependency) information.
  - » function added as additional sub-label to daughter category, e.g. [S [NP-SB .... ]] in Penn Treebank II
  - » constituent label as node label, function as edge label, e.g. in TIGER, TüBA



# Phrases and Chunks

---

- A **phrase** is a constituent of a particular category
  - » exocentric phrase vs. endocentric phrase

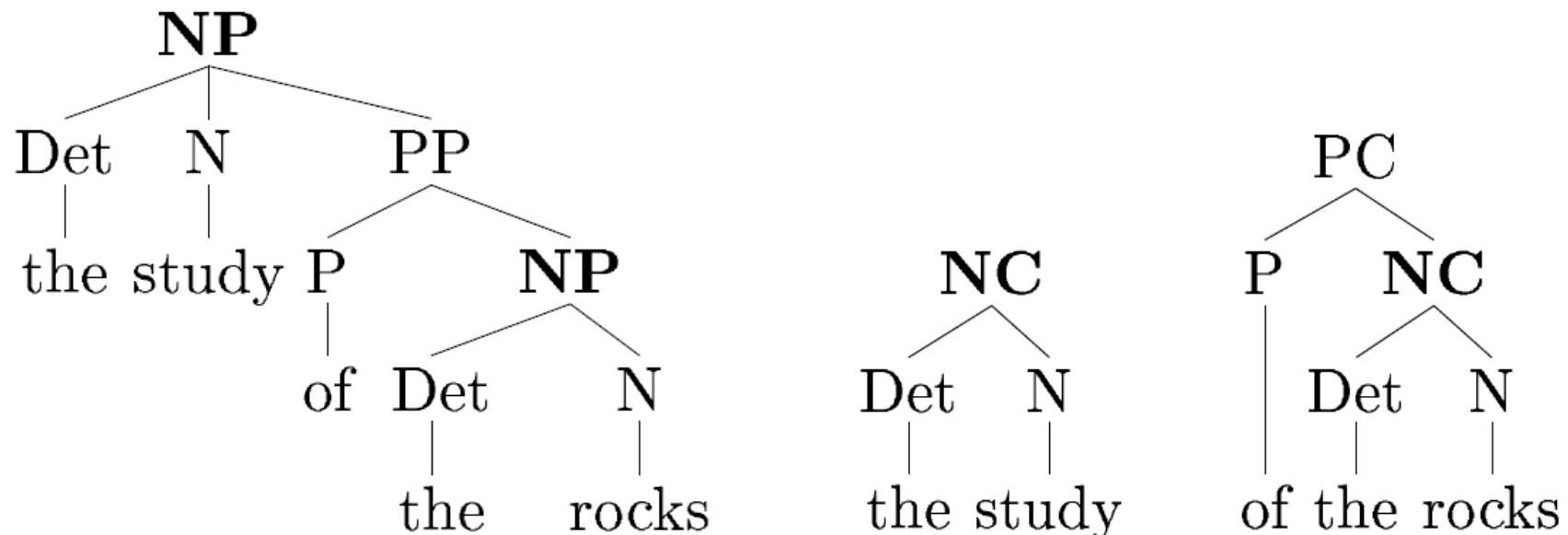


- A typical **chunk** consists of a single content word and surrounding constellation of function words
- The non-recursive core of a constituent which spans the beginning of the constituent **up to its lexical head**.

[the bold man][was sitting][on his suitcase]

# Phrases and Chunks

- Recursive phrase structure vs. non-recursive chunking



- Fully-fledged analysis vs. chunking (also 'partial parsing').

# Syntactic Annotation: References

---

- Steven Abney (1991): „[Parsing By Chunks.](#)” In: Robert Berwick, Steven Abney and Carol Tenny (eds.): “[Principle-Based Parsing.](#)” Kluwer Academic Publishers, Dordrecht.
- Geoffrey Leech, Ruthanna Barnett, Peter Kahrel (1996): “[EAGLES. Recommendations for the Syntactic Annotation of Corpora.](#)” EAGLES Document EAG-TCWG-SASG/1.8.
- Anne Abeillé, ed. (2003): “[Treebanks. Building and Using Parsed Corpora.](#)” Dordrecht, Boston, London: Kluwer Academic Publishers.
- Lothar Lemnitzer & Heike Zinsmeister (2006): “[Korpuslinguistik. Eine Einführung.](#)” Tübingen: Narr Verlag, chapter 4.
- Joakim Nivre (to appear): “[Treebanks.](#)” In: Anke Lüdeling and Merja Kytö (eds.): “[Corpus Linguistics: An International Handbook.](#)” Berlin: Mouton de Gruyter.

# Syntactic Annotation: References

---

- Michell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, Britta Schasberger (1994): „[The Penn Treebank: Annotating predicate argument structure](#)“. In: *Proceedings of the ARPA Human Language Technology Workshop*.
- Geoffrey Leech and Elizabeth Eyes (1997): „[Syntactic Annotation: Treebanks](#).“ In: Richard Garside, Geoffrey Leech and Anthony McEnery, editors: „[Corpus Annotation](#).“ London, New York: Longman, pp. 34-52.
- John Carroll, Guido Minnen, and Ted Briscoe (1999): „[Corpus annotation for parser evaluation](#).“ In: *Proceedings of Linguistically Interpreted Corpora*.“
- Catherine Lai and Steven Bird (2004): „[Querying and updating treebanks: A critical survey and requirements analysis](#).“ In: *Proceedings of the Australasian Language Technology Workshop*.

# The Penn Treebank



# Penn Treebank

---

- English treebank built at the University of Pennsylvania  
<http://www.cis.upenn.edu/~treebank>
- Distributed by the Linguistic data consortium (LDC)  
<http://www ldc.upenn.edu>
- **Phase I (1989 – 1992):** skeletal parse
  - 2.6 mill words tagged (PoS) material from Dow Jones News Service (Wall Street Journal)
  - thereof over 1,7 mill word hand-parsed material
  - first fully parsed version of Brown Corpus (1 million words)
  - tagged and parsed data from Department of Energy abstracts, IBM computer manuals, MUC-3 and ATIS.

# Penn Treebank

---

- Phase II (1993 – 1995)
  - » enriching part of the original material with
    - grammatical functions and semantic relations
    - null elements, coreference symbols
    - information about non-continuous constituents / dependencies (traces, coreference symbols)
- Phase III (1996 - 2000)
  - » additional material
    - Switchboard Corpus (telephone conversations):  
parsed and disfluency-annotated

# Penn Treebank: POS Annotation

---

- Modified **BROWN** tagset
  - avoids lexical redundancies: no tags that are unique to particular lexical items (exception: 'TO').
  - encodes word's syntactic function when possible, e.g.
  - one\_CD apple vs. the ones\_NN
  - allows for multiple tagging: word's POS cannot be decided or annotator is unsure → avoid arbitrary decisions
  - 36 POS tags, 12 other tags (punctuation, currency symbols)

# Penn Treebank: Skeletal parsing

---

```
( (S
  (NP Martin Marietta Corp.)
  was
  (VP given
    (NP a
      $ 29.9
      million Air Force contract
      (PP for
        (NP low-altitude navigation
          and
          targeting equipment))))))
.)
```

# Penn Treebank: Syntactic Tagset II

---

## Null elements

1. \*        ``Understood'' subject of infinitive or imperative
2. 0        Zero variant of that in subordinate clauses
3. T        Trace---marks position where moved wh-constituent  
            is interpreted
4. NIL      Marks position where preposition is interpreted in  
            pied-piping contexts

# Penn Treebank: Functional Tagset

---

## Text categories

- HLN headlines and datelines
- LST list markers
- TTL titles

## Grammatical functions

- CLF true clefts
- NOM non NPs that function as NPs
- ADV clausal and NP adverbials
- LGS logical subjects in passives
- PRD non VP predicates
- SBJ surface subject
- TPC topicalized and fronted constituents
- CLR closely related

# Penn Treebank: Functional Tagset

---

## Semantic roles

-VOC	vocatives
-DIR	direction and trajectory
-LOC	location
-MNR	manner
-PRP	purpose
-TMP	temporal phrases

## Pseudo-attachment

*ICI*	Interpret Constituent Here
*PPA*	Permanent Predictable Ambiguity
*RNR*	Right Node Raising
*EXP*	Expletive

# Penn Treebank: WH-Question

---

Predicate argument structure and empty categories

```
(SBARQ (WHNP-1 What
        (SQ is
          (NP-SBJ Tim)
          (VP eating
            (NP *T*-1)))
        ?)
```

Predicate Argument Structure:  
eat(Tim, what)

# Penn Treebank: References

---

- Mitchell Marcus, Beatrice Santorini, Mary Ann Marcinkiewicz (1993): “Building a large annotated corpus of English: the Penn Treebank.” In: *Computational Linguistics* 19.
- Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, Britta Schasberger (1994): “The Penn TREEBANK: “Annotating predicate argument structure.”
- **Overview:** Ann Taylor, Mitchell Marcus, Beatrice Santorini (2003): “The Penn Treebank: An overview.” In: Anne Abeillé (ed.) *Treebanks: building and using parsed corpora*. Dordrecht: Kluwer, 5-22.

# **TIGER and TüBa – two German treebanks**

---

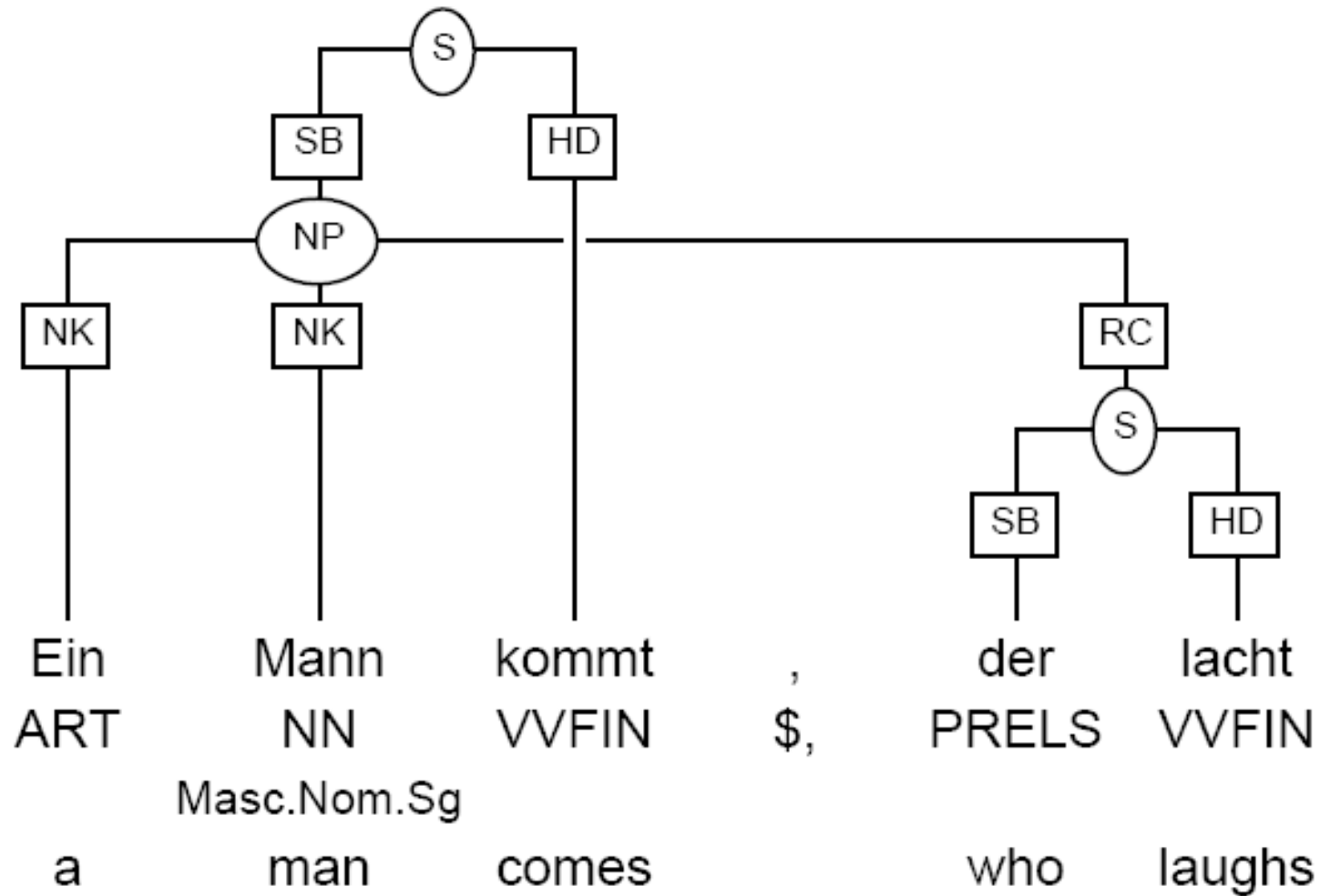
# TIGER Treebank

---

„Linguistic Interpretation of a **GER**man Corpus“

- 50.000 sentences
- Follow-up of **NEGRA** corpus (20.000 sentences)
- German newspaper texts („Frankfurter Rundschau“)
- Free license
- Hybrid annotation
- **Crossing branches** for discontinuous constituents
- Stylebook (in German)
- POS-Tags: STTS-Tagset (Schiller et al. 1999)

# TIGER Discontinuous Constituents



# TüBa-D/Z

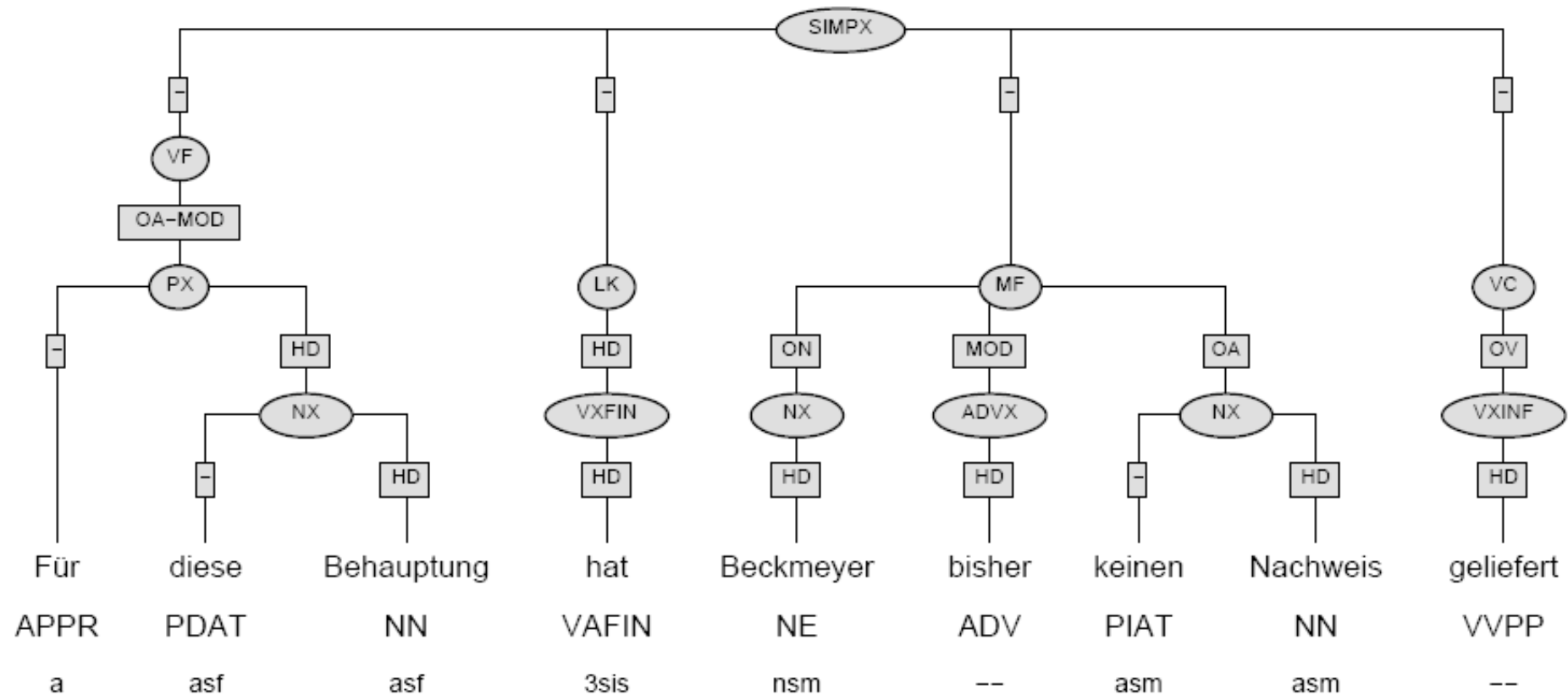
---

## TüBa-D/Z treebank *family*

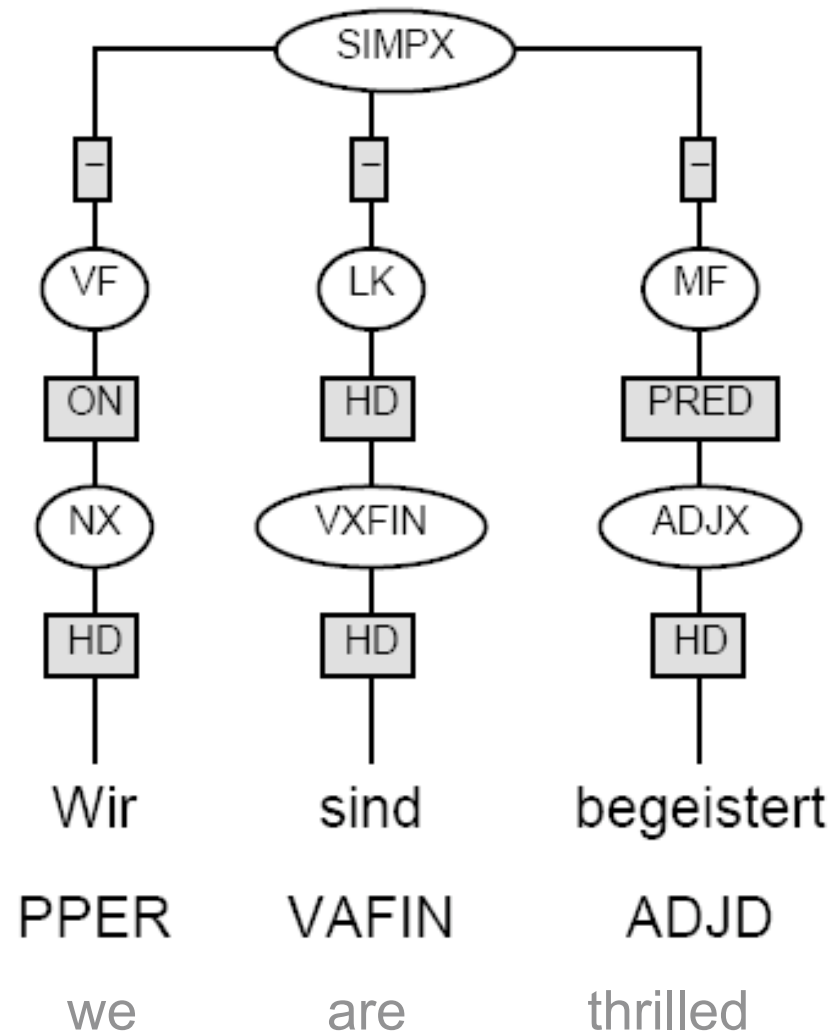
„Tübinger Baumbank des Deutschen/Zeitungssprache“

- **Written** German: newspaper texts (‘tageszeitung’ (taz))
  - 27.000 sentences, 470.000 words (by 2006/07)
  - requires licence for taz-CD ~ EUR 50,-
  - [http://www.sfs.uni-tuebingen.de/en\\_tuebadz.shtml](http://www.sfs.uni-tuebingen.de/en_tuebadz.shtml)
- **Spoken** German: Verbmobil dialogues
  - 38.000 sentences, 360.000 words
  - licence free of charge
  - [http://www.sfs.uni-tuebingen.de/en\\_tuebads.shtml](http://www.sfs.uni-tuebingen.de/en_tuebads.shtml)
- POS-Tags: STTS-Tagset (see TIGER)

# TüBA-D/Z: Discontinuous Constituents



# TüBA-D/Z: Representation



# TüBa-D/Z: Bracketing Format

---

```
%%Sent 1630
( (SIMPX
  (VF
    (NX-ON
      (PPER-HD Wir)))
  (LK
    (VXFIN-HD
      (VAFIN-HD sind)))
  (MF
    (ADJX-PRED
      (ADJD-HD begeistert))))
($. !))
```

# TüBa-D/Z: Column Format

---

‚NEGRA export format‘ of the Annotate tool.

Wir	PPER	--	HD	500
sind	VAFIN	--	HD	501
begeistert	ADJD	--	HD	502
!	\$.	--	--	0
#500	NX	--	ON	503
#501	VXFIN	--	HD	504
#502	ADJX	--	PRED	505
#503	VF	--	-	506
#504	LK	--	-	506
#505	MF	--	-	506
#506	SIMPX	--	--	0

# TüBa-D/Z: Export XML

```
<sentence>
  <node cat="SIMPX" func="--" parent="0" comment="">
    <node cat="VF" func="-" comment="">
      <node cat="NX" func="ON" comment="">
        <word form="Wir" pos="PPER" func="HD" comment=""/>
      </node>
    </node>
    <node cat="LK" func="-" comment="">
      <node cat="VXFIN" func="HD" comment="">
        <word form="sind" pos="VAFIN" func="HD" comment=""/>
      </node>
    </node>
    <node cat="MF" func="-" comment="">
      <node cat="ADJX" func="PRED" comment="">
        <word form="begeistert" pos="ADJD" func="HD" comment=""/>
      </node>
    </node>
  </node>
  <word form="!" pos="$." func="--" parent="0" comment=""/>
</sentence>
```