Proseminar Linguistische Annotation

Ines Rehbein und Josef Ruppenhofer

SS 2010

Überblick über verschiedene Arten linguistischer Annotation

1. Linguistisch annotierte Korpora
   - Korpora - Geschichtlicher Überblick
   - Baumbanken - Überblick
   - Baumbanken fürs Deutsche
   - Treebanking

Überblick über verschiedene Arten linguistischer Annotation

1. Linguistisch annotierte Korpora
   - Korpora - Geschichtlicher Überblick
   - Baumbanken - Überblick
   - Baumbanken fürs Deutsche
   - Treebanking

Was genau sind Korpora?

- Sammlung von
  - Texten (z.B. Zeitungstexte, historische Texte, Kochrezepte, transkribierte gesprochene Sprache, ...)
    $\Rightarrow$ Textkorpora
  - Audiodateien (Sprachaufnahmen, evt. mit Transkription und phonetischer Annotation)
    $\Rightarrow$ Sprachkorpora
  - Videos (z.B. Gebärdensprache, evt. mit Transkription)
    $\Rightarrow$ multimodale Korpora
- meist mit linguistischen Annotationen versehen (Wortart, Syntax, Semantik, Diskurs, ...)
- Repräsentativität?
  - Ausgewogene (balancierte) versus opportunistische Korpora
  - Synchrone versus diachrone Korpora
  - Referenzkorpora (feste Größe) versus Monitorkorpora (anwachsend)

Entscheidung über "best geeignetes Korpus" abhängig von der jeweiligen Fragestellung

Was genau sind Korpora?

- Sammlung von
  - Texten (z.B. Zeitungstexte, historische Texte, Kochrezepte, transkribierte gesprochene Sprache, ...)
    $\Rightarrow$ Textkorpora
  - Audiodateien (Sprachaufnahmen, evt. mit Transkription und phonetischer Annotation)
    $\Rightarrow$ Sprachkorpora
  - Videos (z.B. Gebärdensprache, evt. mit Transkription)
    $\Rightarrow$ multimodale Korpora
- meist mit linguistischen Annotationen versehen (Wortart, Syntax, Semantik, Diskurs, ...)
- Repräsentativität?
  - Ausgewogene (balancierte) versus opportunistische Korpora
  - Synchrone versus diachrone Korpora
  - Referenzkorpora (feste Größe) versus Monitorkorpora (anwachsend)

Entscheidung über "best geeignetes Korpus" abhängig von der jeweiligen Fragestellung

Was genau sind Korpora?

- Sammlung von
  - Texten (z.B. Zeitungstexte, historische Texte, Kochrezepte, transkribierte gesprochene Sprache, ...)
    $\Rightarrow$ Textkorpora
  - Audiodateien (Sprachaufnahmen, evt. mit Transkription und phonetischer Annotation)
    $\Rightarrow$ Sprachkorpora
  - Videos (z.B. Gebärdensprache, evt. mit Transkription)
    $\Rightarrow$ multimodale Korpora
- meist mit linguistischen Annotationen versehen (Wortart, Syntax, Semantik, Diskurs, ...)
- Repräsentativität?
  - Ausgewogene (balancierte) versus opportunistische Korpora
  - Synchrone versus diachrone Korpora
  - Referenzkorpora (feste Größe) versus Monitorkorpora (anwachsend)

Entscheidung über "best geeignetes Korpus" abhängig von der jeweiligen Fragestellung

Erste Korpora

- Schon im 19. Jhdt. (und früher) Verwendung von Textsammlungen
  - zur Beschreibung von Sprachwandel
  - Wörterbucherstellung (z.B. Grimmsches Wörterbuch)
  - Dokumentation von Spracherwerb
  - Belege für grammatische Aussagen
- meist Belege aus der Literatur
- nicht repräsentativ

Erste digitale Korpora

- Anfang 60er:
  - Brown University Standard Corpus of Present-Day American English (Francis & Kucera)
    - ★ synchron, ausgewogen (balanced)
    - ★ ca. 1 Mio. Token (500 Samples mit je 2000 Token)
    - ★ geschriebene Sprache von 1961
    - ★ Korpus fertiggestellt in 1964
- Andere Korpora folgten:
  - Lancaster-Oslo/Bergen (LOB) Corpus (Leech)
    - ★ erstellt 1970-78
    - ★ englisches Gegenstück zum Brown Corpus (Größe, Design)
  - London-Lund Corpus (LLC, Swartvik)
    - ★ publiziert 1980
    - ★ gesprochenes Englisch, transkribiert
    - ★ ca. 50 000 Token
  - Kolhapur Corpus of Indian English (Shastri, 1988)
  - Australian Corpus of English (ACE)
  - Wellington Corpus of Written New Zealand English

Erste digitale Korpora

- Anfang 60er:
  - Brown University Standard Corpus of Present-Day American English (Francis & Kucera)
    - ★ synchron, ausgewogen (balanced)
    - ★ ca. 1 Mio. Token (500 Samples mit je 2000 Token)
    - ★ geschriebene Sprache von 1961
    - ★ Korpus fertiggestellt in 1964
- Andere Korpora folgten:
  - Lancaster-Oslo/Bergen (LOB) Corpus (Leech)
    - ★ erstellt 1970-78
    - ★ englisches Gegenstück zum Brown Corpus (Größe, Design)
  - London-Lund Corpus (LLC, Swartvik)
    - ★ publiziert 1980
    - ★ gesprochenes Englisch, transkribiert
    - ★ ca. 50 000 Token
  - Kolhapur Corpus of Indian English (Shastri, 1988)
  - Australian Corpus of English (ACE)
  - Wellington Corpus of Written New Zealand English

Erste Reaktionen auf linguistische Korpora

- **Theoretische Linguistik:**
  - 1957: Noam Chomsky, *Syntactic Structures*
  - Empirismus als herrschendes Paradigma in der Linguistik (und anderen Kognitionswissenschaften) wird vom Rationalismus abgelöst
  - Fokus auf Sprachkompetenz, Sprachperformanz und quantitative Aspekte von Sprache gelten als uninteressant

    "It must be recognized that the notion 'probability of a sentence' is an entirely useless one, under any known interpretation of this term." (Chomsky, 1969)

  - Wenig Interesse an empirischen, korpus-linguistischen Projekten
  - Korpora als zufällige, nicht repräsentative Sammlungen von Texten, die keinen wirklichen Einblick in die Sprachkompetenz geben

Erste Reaktionen auf linguistische Korpora

- **Theoretische Linguistik:**
  - 1957: Noam Chomsky, *Syntactic Structures*
  - Empirismus als herrschendes Paradigma in der Linguistik (und anderen Kognitionswissenschaften) wird vom Rationalismus abgelöst
  - Fokus auf Sprachkompetenz, Sprachperformanz und quantitative Aspekte von Sprache gelten als uninteressant

    "It must be recognized that the notion 'probability of a sentence' is an entirely useless one, under any known interpretation of this term." (Chomsky, 1969)

  - Wenig Interesse an empirischen, korpus-linguistischen Projekten
  - Korpora als zufällige, nicht repräsentative Sammlungen von Texten, die keinen wirklichen Einblick in die Sprachkompetenz geben

Erste Reaktionen auf linguistische Korpora

- **Theoretische Linguistik:**
  - 1957: Noam Chomsky, *Syntactic Structures*
  - Empirismus als herrschendes Paradigma in der Linguistik (und anderen Kognitionswissenschaften) wird vom Rationalismus abgelöst
  - Fokus auf Sprachkompetenz, Sprachperformanz und quantitative Aspekte von Sprache gelten als uninteressant

    "It must be recognized that the notion 'probability of a sentence' is an entirely useless one, under any known interpretation of this term." (Chomsky, 1969)

  - Wenig Interesse an empirischen, korpus-linguistischen Projekten
  - Korpora als zufällige, nicht repräsentative Sammlungen von Texten, die keinen wirklichen Einblick in die Sprachkompetenz geben

Erste Reaktionen auf linguistische Korpora (2)

- **Computerlinguistik:**

  "Corpora crashed into computational linguistics at the 1989 ACL meeting in Vancouver: but they were large, messy, ugly objects clearly lacking in theoretical integrity in all sorts of ways, and many people were skeptical regarding their role in the discipline."
  (Kilgarriff and Grefenstette 2003; Seite 334)

Nach Überwindung der Anfangsschwierigkeiten: Korpora als wichtige Ressource für die anwendungsorientierte statistische Sprachverarbeitung

Korpora - Wozu?

- Korpora erweisen sich als fruchtbare Hilfsmittel für linguistische Forschung:
  - ermöglichen die **Überprüfung linguistischer Theorien**
  - sinnvolle Ergänzung der Introspektion
  - Beispiele: Voranstellung von Partikelverben im Deutschen, Variabilität von Idiomen (siehe Folien vom 15.04.2010)
  - Ressourcen zum **Training von statistischen NLP-Systemen**
  - Linguistisch annotierte Daten zur **Evaluation** von NLP-Systemen (Goldstandard)
- Daher steigender Bedarf an
  - mehr Daten
  - mehr Annotation (Syntax, Semantik, Prosodie, Metadaten, ...)
  - mehr Sprachen

Korpora - Wozu?

- Korpora erweisen sich als fruchtbare Hilfsmittel für linguistische Forschung:
  - ermöglichen die **Überprüfung linguistischer Theorien**
  - sinnvolle Ergänzung der Introspektion
  - Beispiele: Voranstellung von Partikelverben im Deutschen, Variabilität von Idiomen (siehe Folien vom 15.04.2010)
  - Ressourcen zum **Training von statistischen NLP-Systemen**
  - Linguistisch annotierte Daten zur **Evaluation** von NLP-Systemen (Goldstandard)
- Daher steigender Bedarf an
  - mehr Daten
  - mehr Annotation (Syntax, Semantik, Prosodie, Metadaten, ...)
  - mehr Sprachen

Korpora - Wozu?

- Korpora erweisen sich als fruchtbare Hilfsmittel für linguistische Forschung:
  - ermöglichen die **Überprüfung linguistischer Theorien**
  - sinnvolle Ergänzung der Introspektion
  - Beispiele: Voranstellung von Partikelverben im Deutschen, Variabilität von Idiomen (siehe Folien vom 15.04.2010)
  - Ressourcen zum **Training von statistischen NLP-Systemen**
  - Linguistisch annotierte Daten zur **Evaluation** von NLP-Systemen (Goldstandard)
- Daher steigender Bedarf an
  - mehr Daten
  - mehr Annotation (Syntax, Semantik, Prosodie, Metadaten, ...)
  - mehr Sprachen

Überblick über verschiedene Arten linguistischer Annotation

1. Linguistisch annotierte Korpora
   - Korpora - Geschichtlicher Überblick
   - Baumbanken - Überblick
   - Baumbanken fürs Deutsche
   - Treebanking

Was sind und wofür braucht man Baumbanken?

- Baumbanken sind
  - Korpora mit syntaktischen Annotationen (über Part-of-Speech Ebene hinausgehend)
  - Syntax-Bäume a la Chomsky oder Dependenzen
  - manuell erstellt *oder*
  - automatisch erstellt und manuell korrigiert

- Baumbanken werden gebraucht zur
  - Untersuchung linguistischer Phänomene
  - Ressourcen zum Training von Methoden des Maschinellen Lernens/ für die Entwicklung von Sprachtechnologien:
    - ★ Training und Evaluation von Parsern
    - ★ Ressourcen für Maschinelle Übersetzung (Parallele Baumbanken)
    - ★ Extraktion von Subkategorisierungsrahmen für die Erstellung von Lexika
    - ★ ...

Erste Baumbanken

- Ellegård (Englisch, Uni Göteborg, 1978)
  - 128 000 Token des Brown Corpus of American English
  - **manuell** annotiert mit einer Dependenzgrammatik (Francis and Kucera, 1979)
- Lancaster-Leeds Treebank (Englisch, 80er Jahre)
  - 4.5% des LOB Korpus (45 000 Token)
  - ausgewogen (15 Textgenre)
  - automatisch annotiert mit POS-tags
  - **handgeparst** von G. Sampson
  - basierend auf einer Phrasenstrukturgrammatik
  - detaillierte Annotation
- LOB Corpus Treebank (English, Anfang 90er)
  - **automatisch und probabilistisch** geparste, handkorrigierte Texte vom LOB Corpus
  - 144 000 Token
  - weniger detallierte Annotation als die Lancaster-Leeds Treebank

Die Penn Treebank (PTB)

- Penn Treebank (Englisch, 1989-1995)
- Phase I (1989-1992)
  - Wall Street Journal (50 000 Sätze, 1 Mio. Worte)
  - Zusätzlich: geparste Version des Brown Korpus (1 Mio. Worte),
  - Automatisch getagged (POS)
  - Manuell annotiert mit Phrasen-Struktur (skeletal parse)

```
(SBARQ (WHNP Who)
       (SQ (NP T)
           will
           (VP come
               (PP to
                   (NP the party))))
?)
```

Penn Treebank - Phase II (1993-1995)

- Zusätzliche Annotation von grammatikalischen Funktionen
  - "tiefe" linguistische Information, ermöglicht die Extraktion von Prädikat-Argument-Struktur
  - 3 Arten von grammatikalischen Funktionen:
    - ★ GFs, die auf rein syntaktischer Ebene definiert sind: DTV, LGS, PRD, PUT, SBJ, TPC, VOC
    - ★ Form/Funktions-Tags: ADV (klausale und NP-Adverbiale), NOM (non-NP, die die Funktion einer NP hat)
    - ★ Semantische Rollen: BNF, DIR, EXT, LOC, MNR, PRP, TMP, MNR, TMP, PRD, ...
  - Annotation von Spuren, Null-Elementen, Koreferenz
    - ★ * "Ungesprochenes" Subjekt von Infinitiven oder Imperativen
    - ★ 0 Null-Variante von *that* in subordinierten Sätzen
    - ★ T markiert die Position, wo eine vorangestellte *wh*-Konstituente interpretiert wird

Penn Treebank - Phase II

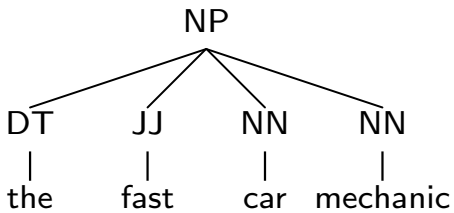
```
(SBARQ (WHNP-1 Who)
          (SQ was
             (NP-SBJ-2 *T*-1)
             (VP believed
               (S (NP-SBJ-3 *-2)
                  (VP to
                    (VP have
                      (VP been
                         (VP shot
                         (NP *-3)))))))
?)
```

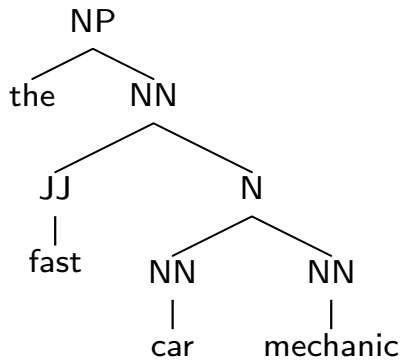
Penn Treebank - Phase II

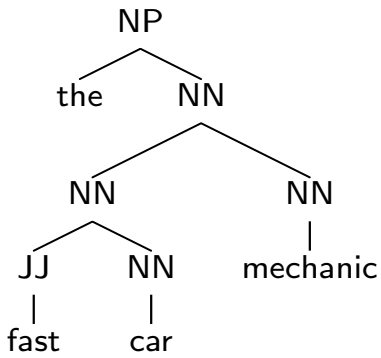
```
( (S (PP-TMP (IN On)
          (NP (NP (NNP Nov.) (CD 15) )
             (, ,)
             (SBAR-TMP (WHADVP-1 (WRB when) )
                (S (NP-SBJ (NNS Brazilians) )
                   (VP (VBP elect)
                      (NP (DT a) (NN president) )
                      (PP (IN for)
                         (NP (NP (DT the) (JJ first) (NN time) )
                            (PP-TMP (IN in)
                               (NP (CJ 29) (NNS years) ))))
                      (ADVP-TMP (-NONE- *T*-1) ))))))
    (, ,)
    (NP-SBJ (NP (DT the) (NN country) (POS 's) )
       (QP (CD 82) (CD million) )
       (NNS voters) )
    (VP (MD will)
       (VP (VB have)
          (NP (NP (CD 22) (NNS candidates) )
             (SBAR (WHNP-2 (-NONE- 0) )
                (S (NP-SBJ (-NONE- *) )
                   (VP (TO to)
                     (VP (VB choose)
                        (PP-CLR (IN from)
                           (NP (-NONE- *T*-2) ))))))))
    (. .) ))
```

Besonderheiten der Penn Treebank

- Flache Annotation z.B. für Nomen-Prämodifizierer







- VP-Argumente und Adjunkte auf der gleichen Ebene
- Keine detaillierte Analyse von NPs (Zeitersparnis, Konsistenz)

  (NP (NP the defense and electronics group Thomson-CSF S.A.)
  and
  (NP the bank group Credit Lyonnais))

Bedeutung der Penn Treebank für NLP

- Bislang: Parser mit handgeschriebenen Regeln (zeitaufwendig, geringe Abdeckung)
- Ende 80er: Erste Baumbanken für linguistische Forschung
- Anfang 90er: Penn Treebank ermöglicht eine neue Herangehensweise an Parsing:
  - Machine Learning Algorithmen
  - Probabilistische Parser (robust, Ranking nach Häufigkeit der vorkommenden Strukturen)
- PTB als Benchmark für die Evaluation von Parsern fürs Englische
- PTB war verantworlich für große Fortschritte im Bereich des statistischen Parsens in den letzten 20 Jahren
  - F-score: 0.84 (Magerman, 1994) → 0.91 (Charniak & Johnson, 2005)

(aus einem Vortrag von Stefan Oepen, ULA, Bergen 2008)

Bedeutung der Penn Treebank für NLP

- Bislang: Parser mit handgeschriebenen Regeln (zeitaufwendig, geringe Abdeckung)
- Ende 80er: Erste Baumbanken für linguistische Forschung
- Anfang 90er: Penn Treebank ermöglicht eine neue Herangehensweise an Parsing:
  - Machine Learning Algorithmen
  - Probabilistische Parser (robust, Ranking nach Häufigkeit der vorkommenden Strukturen)
- PTB als Benchmark für die Evaluation von Parsern fürs Englische
- PTB war verantworlich für große Fortschritte im Bereich des statistischen Parsens in den letzten 20 Jahren
  - F-score: 0.84 (Magerman, 1994) $\rightarrow$ 0.91 (Charniak & Johnson, 2005) *(aus einem Vortrag von Stefan Oepen, ULA, Bergen 2008)*

Bedeutung der Penn Treebank für NLP (2)

- Heute: viel Kritik an der PTB:
  - begrenzte Domäne
  - veraltete Texte
  - keine (explizite) Kodierung von Phrasenköpfen
  - keine Unterscheidung von Argumenten und Modifikatoren
  - keine Multi Word Expressions (MWE, z.B. Verbpartikeln), flache NPs
  - die meisten Parser ignorieren einen großen Teil der in der Baumbank vorhandenen Informationen

> In the past decade or so, Computational Linguistics has degenerated into the science of the Wall Street Journal. (Ron Kaplan, 2007)

(aus einem Vortrag von Stefan Oepen, ULA, Bergen 2008)

- Neu erwachtes Interesse an "tiefen" Grammatiken und ausdrucksstarken Repräsentationen
- LFG Parsing, CCG Parsing, Tree Adjoining Grammars, HPSG

Bedeutung der Penn Treebank für NLP (2)

- Heute: viel Kritik an der PTB:
  - begrenzte Domäne
  - veraltete Texte
  - keine (explizite) Kodierung von Phrasenköpfen
  - keine Unterscheidung von Argumenten und Modifikatoren
  - keine Multi Word Expressions (MWE, z.B. Verbpartikeln), flache NPs
  - die meisten Parser ignorieren einen großen Teil der in der Baumbank vorhandenen Informationen

> In the past decade or so, Computational Linguistics has degenerated into the science of the Wall Street Journal. (Ron Kaplan, 2007)

(aus einem Vortrag von Stefan Oepen, ULA, Bergen 2008)

- Neu erwachtes Interesse an "tiefen" Grammatiken und ausdrucksstarken Repräsentationen
- LFG Parsing, CCG Parsing, Tree Adjoining Grammars, HPSG

Digitale Korpora / Baumbanken - Zwischenfazit

- Erste digitale Korpora seit Mitte 60er, erste syntaktisch annotierte digitale Korpora seit Anfang 80er
- Wichtige Hilfsmittel für linguistische und computerlinguistische Forschung
- Penn Treebank als erstes großes, syntaktisch annotiertes Korpus ermöglicht neue Herangehensweisen in NLP
  $\longrightarrow$ **probabilistische Methoden gewinnen an Bedeutung**
- Bedeutung von linguistisch annotierten Korpora wächst:
  - Erstellung von Korpora für andere Sprachen
  - Ausweitung der Annotation (Syntax, Semantik, Koreferenzen, ...)

Baumbanken für weitere Sprachen

- Phrasenstruktur-Baumbanken
  - BulTreebank (Bulgarisch, HPSG)
  - Penn Chinese Treebank
  - Alpino Treebank (Niederländisch)
  - Floresta sinta(c)tica (Portugiesisch)
  - Cast3LB (Spanisch, Katalanisch)
  - Eus3LB (Baskisch)
  - Talbanken (Schwedisch)
  - Penn Arabic Treebank
- Dependenz-Baumbanken:
  - Tschechisch: Prague Dependency Treebank
  - Prague Arabic Dependency Treebank
  - Danish Dependency Treebank
  - Slovene Dependency Treebank
  - METU-Sabanci Turkish Treebank
  - Kyoto Text Corpus (Japanisch)
- Hybride Baumbanken:
  - NEGRA, TIGER, TüBa-D/Z (Deutsch)

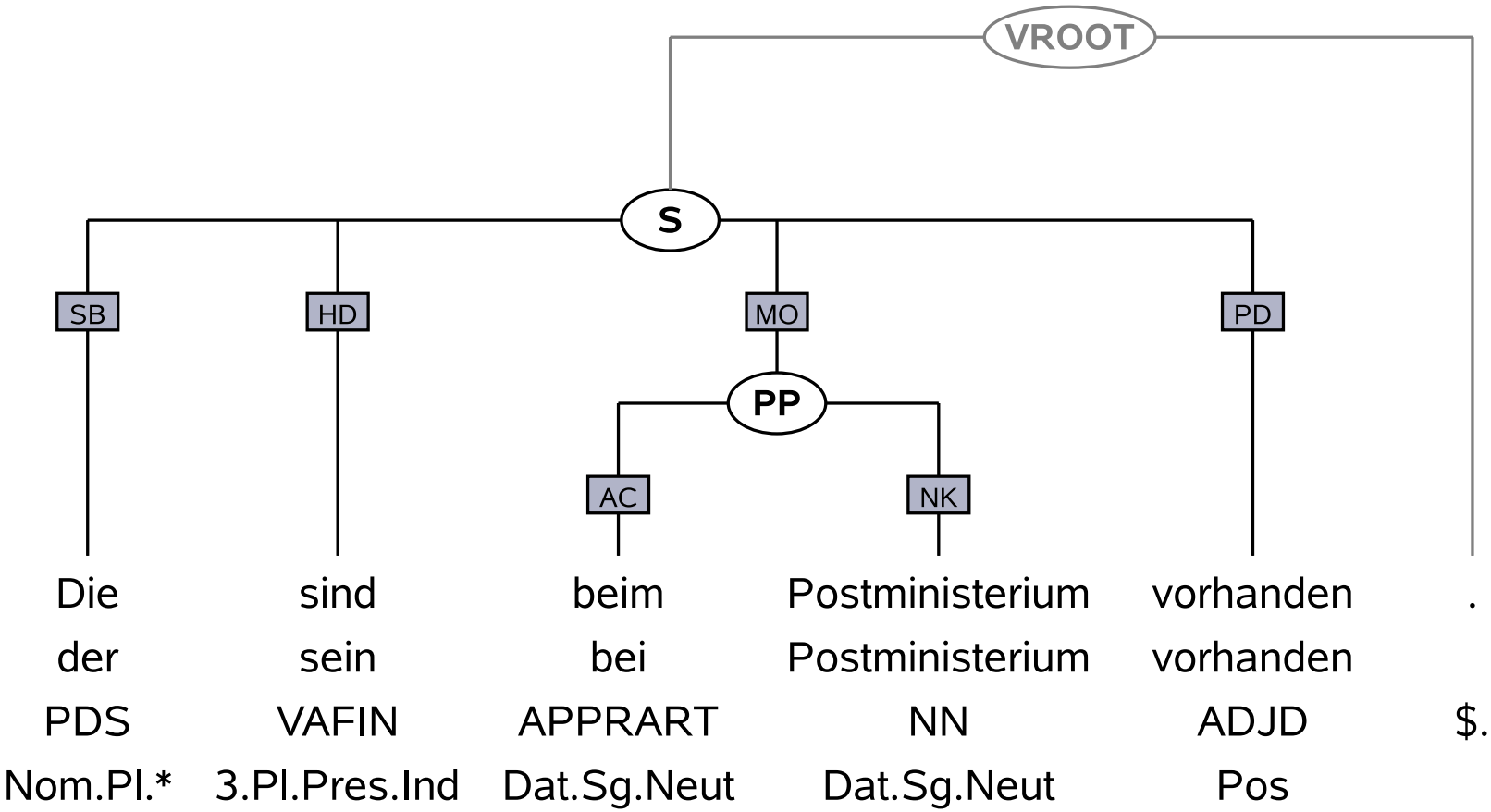
Überblick über verschiedene Arten linguistischer Annotation

1. Linguistisch annotierte Korpora
   - Korpora - Geschichtlicher Überblick
   - Baumbanken - Überblick
   - Baumbanken fürs Deutsche
   - Treebanking

Baumbanken fürs Deutsche

- NEGRA (Uni Saarbrücken)
  - 350 000 Token (20 000 Sätze) der Frankfurter Rundschau
    - ★ POS-Tags, Stuttgart-Tübingen-Tagset (STTS)
    - ★ syntaktischen Kategorien
    - ★ grammatischen Funktionen
    - ★ morphologische Analyse (für die ersten 60 000 Token)
- TIGER (Uni Stuttgart)
  - 900 000 Token (50 000 Sätze) der Frankfurter Rundschau
    - ★ POS-Tags (STTS), syntaktische Kategorien, grammatische Funktionen
    - ★ morphologische Analyse
    - ★ Frame-semantische Annotation
- TüBa-D/Z (Uni Tübingen)
  - 794 000 Token (45 200 Sätze) der taz
    - ★ POS-Tags (STTS), syntaktische Kategorien, grammatische Funktionen
    - ★ morphologische Analyse
    - ★ Named Entities (NE), Ko-Referenzen

Beispielbaum - TIGER Treebank



General Bracketing Format

```
(
    (S
        (PDS-SB Die)
        (VAFIN-HD sind)
        (PP-MO
            (APPRART-AC beim)
            (NN-NK Postministerium)
        )
        (ADJD-PD vorhanden)
    )
    ($. .)
)
```

General Bracketing Format

```
(
    (S
        (PDS-SB Die)
        (VAFIN-HD sind)
        (PP-MO
            (APPRART-AC beim)
            (NN-NK Postministerium)
         )
        (ADJD-PD vorhanden)
    )
    ($. .)
)
```

- Nichtterminale Knoten: S, VP, NP, PP, ...

General Bracketing Format

```
(
    (S
        (PDS-SB Die)
        (VAFIN-HD sind)
        (PP-MO
            (APPRART-AC beim)
            (NN-NK Postministerium)
        )
        (ADJD-PD vorhanden)
    )
    ($. .)
)
```

- Nichtterminale Knoten: S, VP, NP, PP, ...
- Terminale Knoten: Die, sind, beim, ...

General Bracketing Format

```
(
    (S
        (PDS-SB Die)
        (VAFIN-HD sind)
        (PP-MO
            (APPRART-AC beim)
            (NN-NK Postministerium)
        )
        (ADJD-PD vorhanden)
    )
    ($. .)
)
```

- Nichtterminale Knoten: S, VP, NP, PP, ...
- Terminale Knoten: Die, sind, beim, ...
- Part-of-Speech (POS) Tags: PDS, VAFIN, APPRART, NN, ...

General Bracketing Format

```
(
    (S
        (PDS-SB Die)
        (VAFIN-HD sind)
        (PP-MO
            (APPRART-AC beim)
            (NN-NK Postministerium)
        )
        (ADJD-PD vorhanden)
    )
    ($. .)
)
```

- Nichtterminale Knoten: S, VP, NP, PP, ...
- Terminale Knoten: Die, sind, beim, ...
- Part-of-Speech (POS) Tags: PDS, VAFIN, APPRART, NN, ...
- Grammatikalische Funktionen: SB, HD, OA, DA, AG, ...

NEGRA export format

Begin Of Sentence

#BOS 8021 0 1066832867 175 %% PO2AV

Terminale Knoten:

Wortform	Lemma	POS	Morph. Inf.	Label	Elternknoten
Die	der	PDS	Nom.Pl.*	SB	501
sind	sein	VAFIN	3.Pl.Pres.Ind	HD	501
beim	bei	APPRART	Dat.Sg.Neut	AC	500
Postministerium	Postministerium	NN	Dat.Sg.Neut	NK	500
vorhanden	vorhanden	ADJD	Pos	PD	501
.	–	$.	–	–	0
Nicht-Terminale Knoten:					
#500	–	PP	–	MO	501
#501	–	S	–	–	0

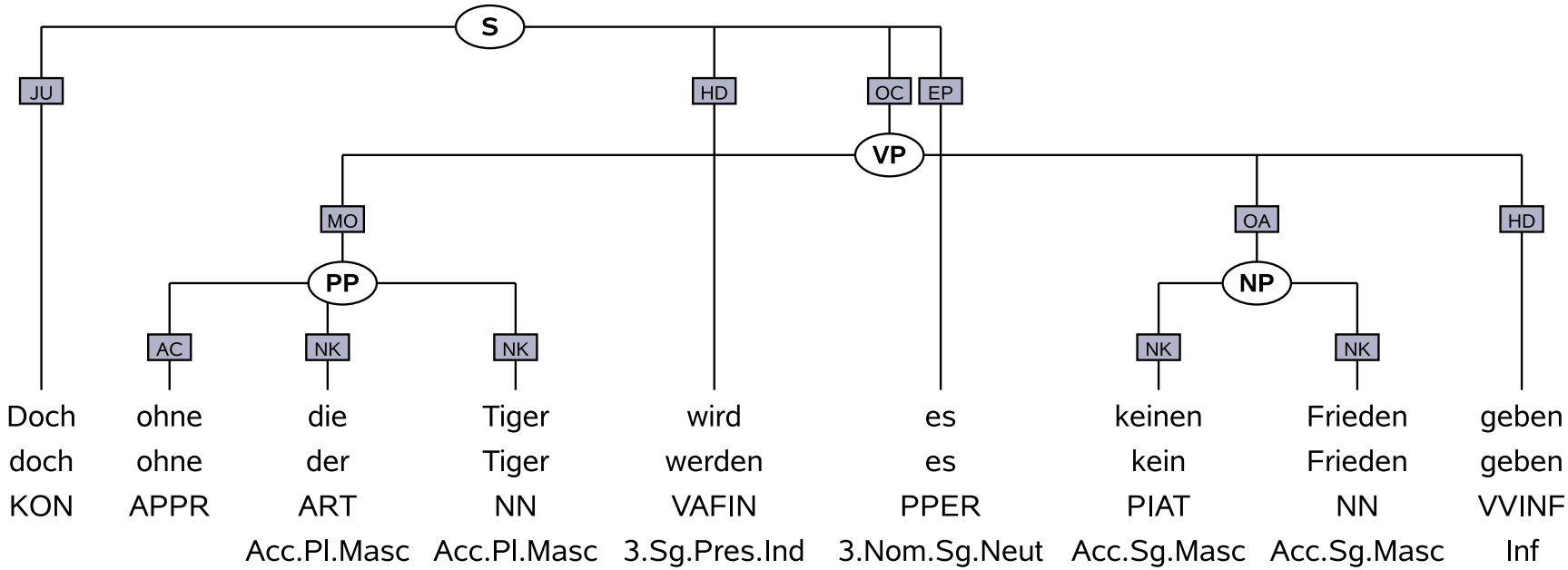
End Of Sentence

#EOS 8021

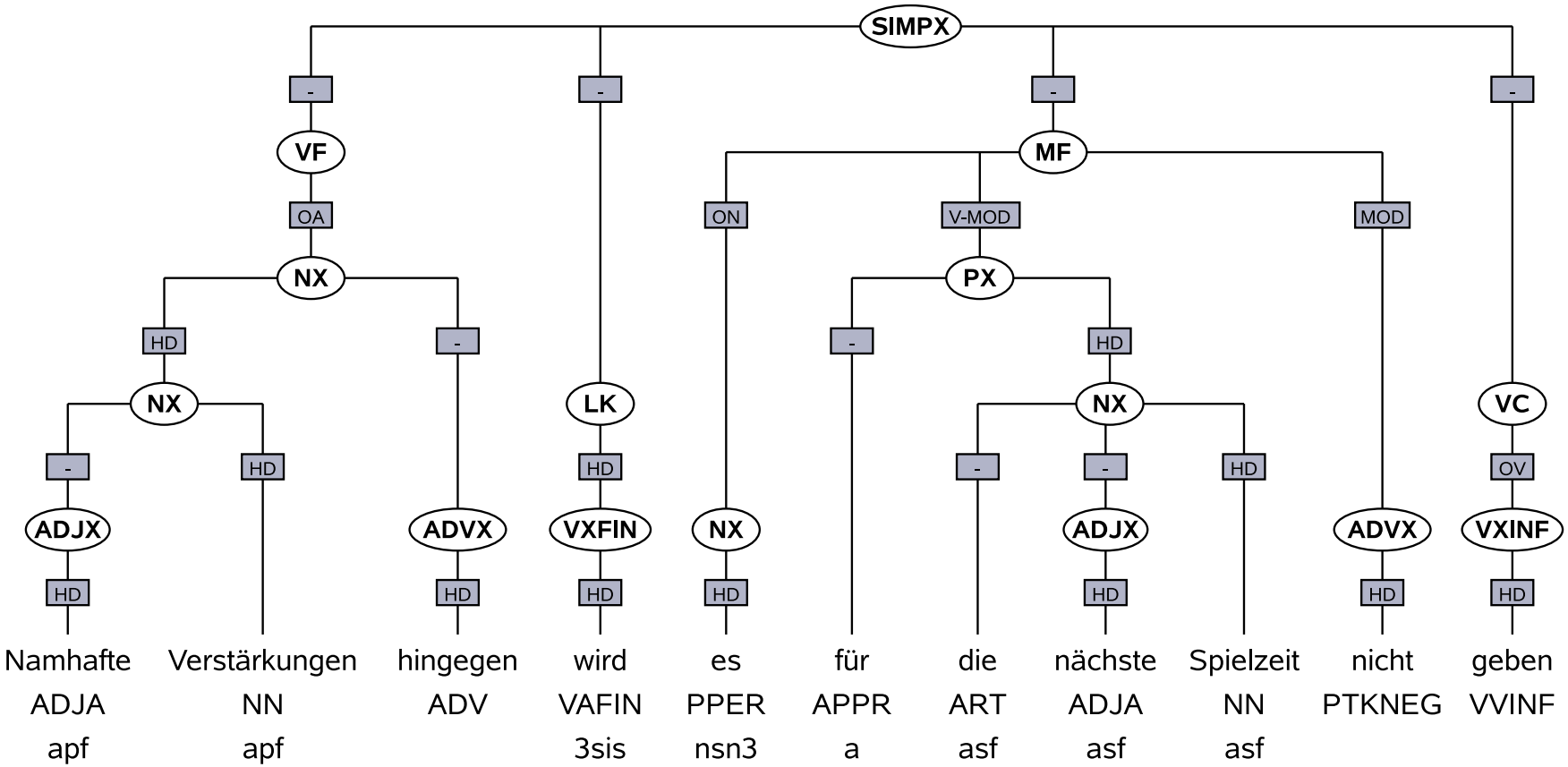
TIGER XML

```
<s id="s8021">
  <graph root="s8021_VROOT">
   <terminals>
    <t id="s8021_1" word="Die" lemma="der" pos="PDS" morph="Nom.Pl.*" />
    <t id="s8021_2" word="sind" lemma="sein" pos="VAFIN" morph="3.Pl.Pres.Ind" />
    <t id="s8021_3" word="beim" lemma="bei" pos="APPRART" morph="Dat.Sg.Neut" />
    <t id="s8021_4" word="Postministerium" lemma="Postministerium" pos="NN" morph="Dat.Sg.Neut" />
    <t id="s8021_5" word="vorhanden" lemma="vorhanden" pos="ADJD" morph="Pos" />
    <t id="s8021_6" word="." lemma="--" pos="$." morph="--" />
   </terminals>
   <nonterminals>
    <nt id="s8021_500" cat="PP">
      <edge label="AC" idref="s8021_3" />
      <edge label="NK" idref="s8021_4" />
      </nt>
    <nt id="s8021_501" cat="S">
      <edge label="SB" idref="s8021_1" />
      <edge label="HD" idref="s8021_2" />
      <edge label="MO" idref="s8021_500" />
      <edge label="PD" idref="s8021_5" />
    </nt>
    <nt id="s8021_VROOT" cat="VROOT">
      <edge label="--" idref="s8021_501" />
      <edge label="--" idref="s8021_6" />
    </nt>
   </nonterminals>
  </graph>
</s>
```

TiGer


S
JU
HD
OC
EP
VP
MO
OA
HD
PP
NP
AC
NK
NK
NK
NK
Doch doch KON
ohne ohne APPR
die der ART Acc.Pl.Masc
Tiger Tiger NN Acc.Pl.Masc
wird werden VAFIN 3.Sg.Pres.Ind
es es PPER 3.Nom.Sg.Neut
keinen kein PIAT Acc.Sg.Masc
Frieden Frieden NN Acc.Sg.Masc
geben geben VVINF Inf

TüBa-D/Z


SIMPX
VF
OA
NX
HD
NX
ADJX
HD
Namhafte
ADJA
apf
HD
Verstärkungen
NN
apf
ADVX
HD
hingegen
ADV
LK
HD
VXFIN
HD
wird
VAFIN
3sis
ON
NX
HD
es
PPER
nsn3
MF
V-MOD
PX
HD
NX
HD
für
APPR
a
ADJX
HD
die
ART
asf
ADJX
HD
nächste
ADJA
asf
HD
Spielzeit
NN
asf
MOD
ADVX
HD
nicht
PTKNEG
VC
OV
VXINF
HD
geben
VVINF

2 Deutsche Baumbanken: TIGER und TüBa-D/Z

- Textsorte: Deutscher Zeitungstext
- POS-Tagset: STTS

- Unterschiede in der Annotation:
  - TiGer: keine unären Knoten
  - TiGer: kreuzende Kanten
  - TüBa-D/Z: Topologische Felder
  - TiGer: flach, TüBa-D/Z: mehr hierarchisch
  - TiGer: keine Kopf-Markierung in NPs, PPs

	# Sätze	∅ Satz-länge	Syntakt. Kat.	Gramm. Funktion	Nicht-Term. /Term. Knoten
TiGer	50474	17.5	25	44	0.47
TüBa-D/Z	27125	17.6	26	40	1.20

(Die Zahlen für TüBa-D/Z beziehen sich auf Release 1 der Baumbank)

Unterschiede TiGer - TüBa-D/Z

- Erlernbarkeit:
  - flache Annotation resultiert in vielen langen Grammatikregeln mit niedriger Frequenz
  - lange, niedrig-frequente Regeln sind nur schwer lernbar für einen statistischen Parser
- Unterschiede in der Annotation von nicht-lokalen Abhängigkeiten
  - in TIGER annotiert mittels kreuzender Kanten
  - müssen aufgelöst werden, bevor PCFG extrahiert wird
    $\longrightarrow$ wichtige Information geht verloren
- Annotation von Topologischen Feldern in TüBa-D/Z vermindert die Anzahl an Regeln in der Grammatik
  $\longrightarrow$ für den Parser leichter zu lernen?

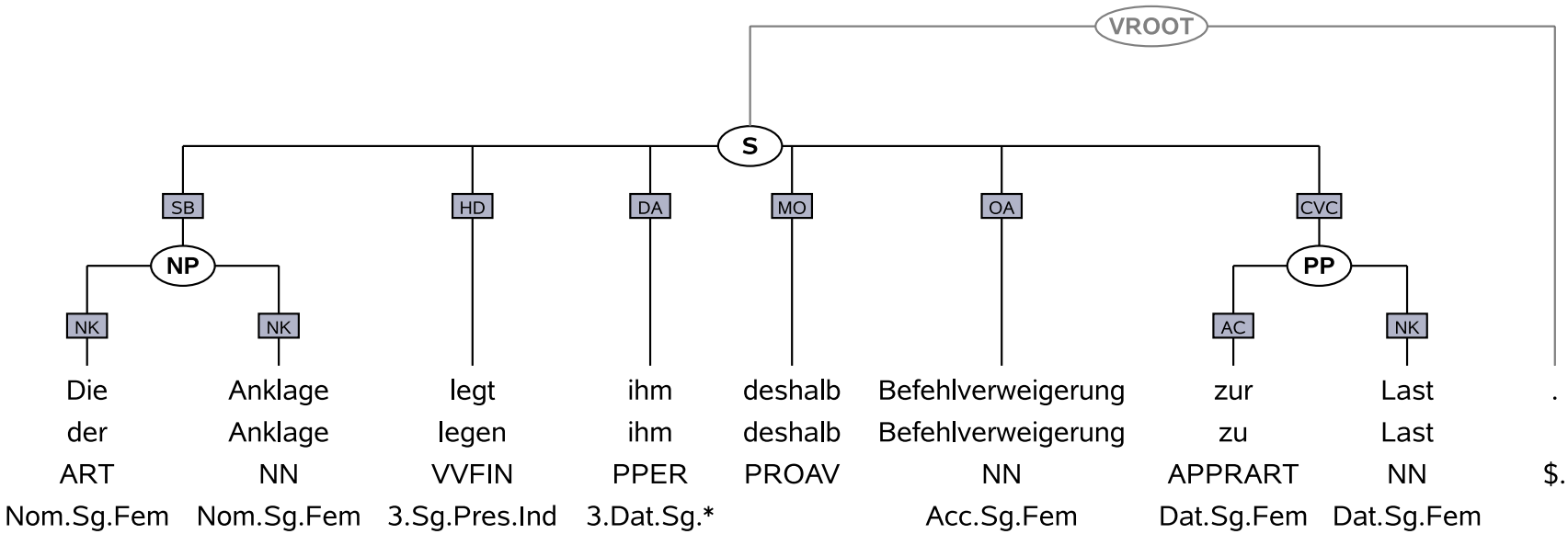
Wortfolge im Deutschen

Die Anklage legt ihm deshalb Befehlsverweigerung zur Last.

Die Anklage	legt	ihm	deshalb	Befehlsverweigerung	zur Last.
Die Anklage	legt	deshalb	Befehlsverweigerung	ihm	zur Last.
Die Anklage	legt	deshalb	ihm	Befehlsverweigerung	zur Last.
Die Anklage	legt	deshalb	ihm	zur Last	Befehlsverweigerung.
Befehlsverweigerung	legt	ihm	deshalb	die Anklage	zur Last.
Befehlsverweigerung	legt	deshalb	ihm	die Anklage	zur Last.
Befehlsverweigerung	zur Last	legt	ihm	deshalb	die Anklage.
Befehlsverweigerung	zur Last	legt	deshalb	ihm	die Anklage.
Befehlsverweigerung	zur Last	legt	deshalb	ihm	die Anklage.
Ihm	legt	die Anklage	deshalb	Befehlsverweigerung	zur Last.
Ihm	zur Last	legt	deshalb	die Anklage	Befehlsverweigerung.
Ihm	zur Last	legt	die Anklage	deshalb	Befehlsverweigerung.
Zur Last	legt	ihm	deshalb	die Anklage	Befehlsverweigerung.
Zur Last	legt	ihm	die Anklage	deshalb	Befehlsverweigerung.
Zur Last	legt	die Anklage	ihm	deshalb	Befehlsverweigerung.
Deshalb	legt	ihm	die Anklage	Befehlsverweigerung	zur Last.
...	...	...	...	...	...

Freie Wortfolge im Deutschen

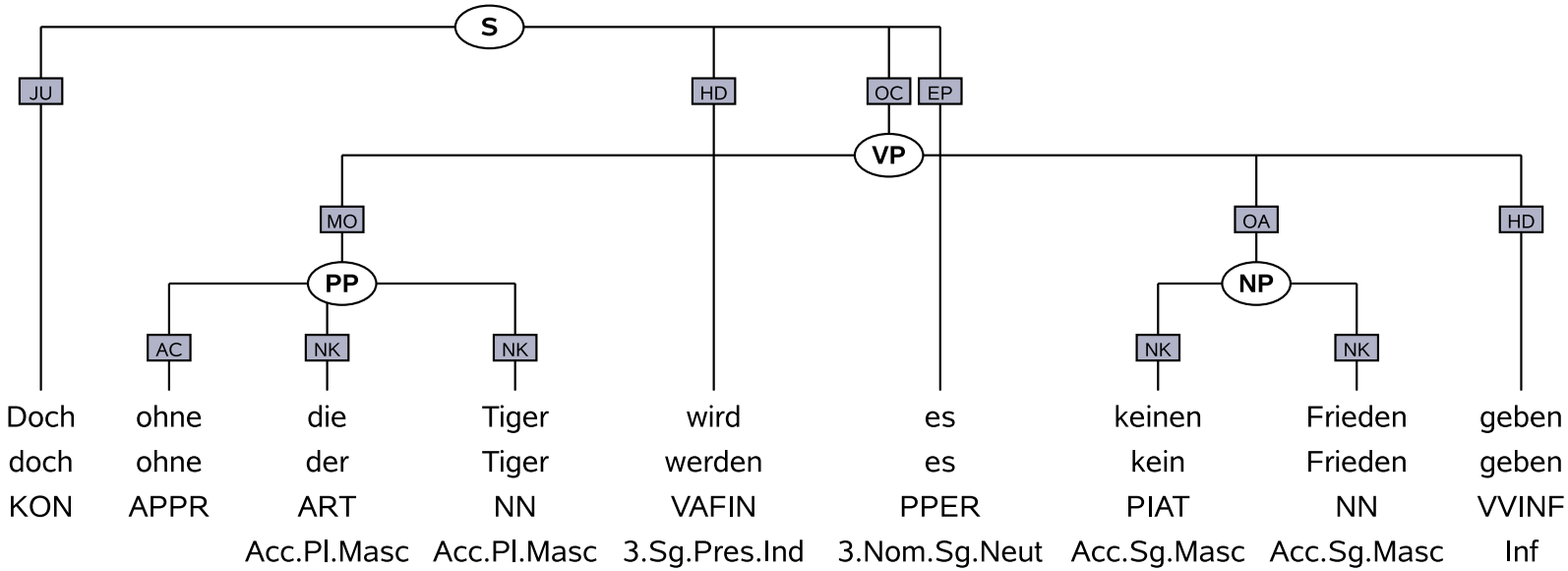
- Regel: S → NP VVFIN PPER PROAV NN PP



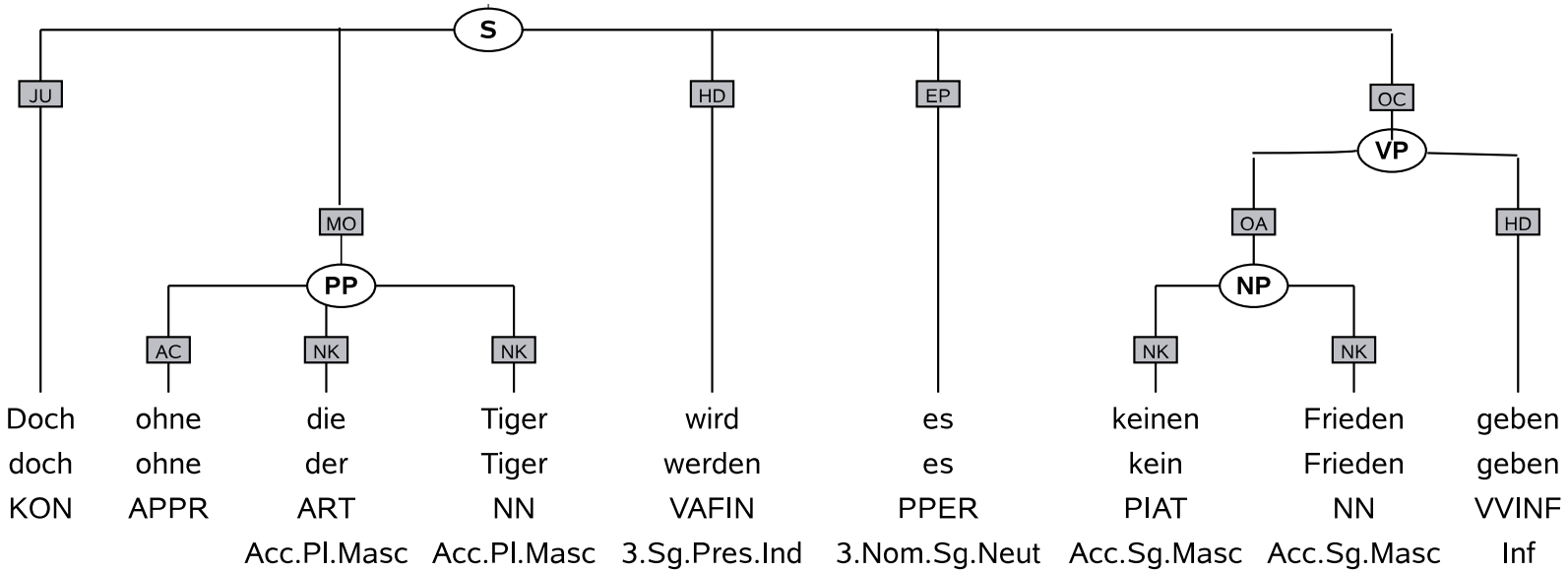
Unterschiede zwischen den Baumbanken

- Erlernbarkeit:
  - flache Annotation resultiert in vielen langen Grammatikregeln mit niedriger Frequenz
  - lange, niedrig-frequente Regeln sind nur schwer lernbar für einen statistischen Parser
- Unterschiede in der Annotation von nicht-lokalen Abhängigkeiten
  - in TIGER annotiert mittels kreuzender Kanten
  - müssen aufgelöst werden, bevor PCFG extrahiert wird
    $\longrightarrow$ wichtige Information geht verloren
- Annotation von Topologischen Feldern in TüBa-D/Z vermindert die Anzahl an Regeln in der Grammatik
  $\longrightarrow$ für den Parser leichter zu lernen?

TIGER


S
JU
HD
OC
EP
VP
MO
OA
HD
PP
NP
AC
NK
NK
NK
NK
Doch doch KON
ohne ohne APPR
die der ART Acc.Pl.Masc
Tiger Tiger NN Acc.Pl.Masc
wird werden VAFIN 3.Sg.Pres.Ind
es es PPER 3.Nom.Sg.Neut
keinen kein PIAT Acc.Sg.Masc
Frieden Frieden NN Acc.Sg.Masc
geben geben VVINF Inf

TIGER


S
JU
HD
EP
OC
VP
MO
OA
HD
PP
NP
AC
NK
NK
NK
NK
Doch doch KON
ohne ohne APPR
die der ART Acc.Pl.Masc
Tiger Tiger NN Acc.Pl.Masc
wird werden VAFIN 3.Sg.Pres.Ind
es es PPER 3.Nom.Sg.Neut
keinen kein PIAT Acc.Sg.Masc
Frieden Frieden NN Acc.Sg.Masc
geben geben VVINF Inf

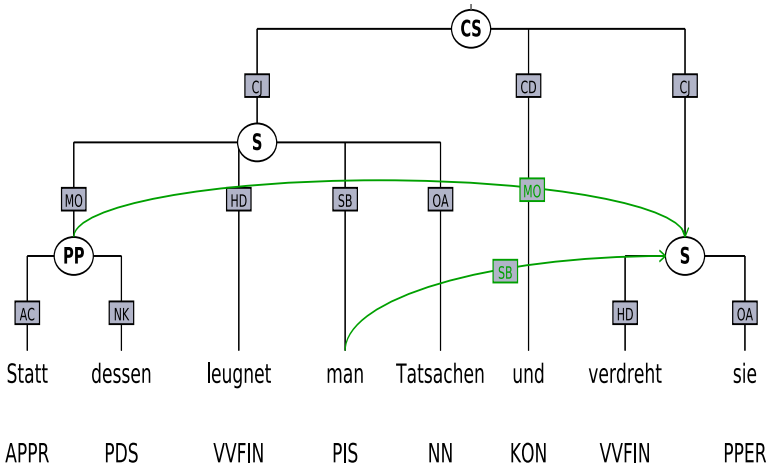
Unterschiede zwischen den Baumbanken

- Erlernbarkeit:
  - flache Annotation resultiert in vielen langen Grammatikregeln mit niedriger Frequenz
  - lange, niedrig-frequente Regeln sind nur schwer lernbar für einen statistischen Parser
- Unterschiede in der Annotation von nicht-lokalen Abhängigkeiten
  - in TIGER annotiert mittels kreuzender Kanten
  - müssen aufgelöst werden, bevor PCFG extrahiert wird
    $\longrightarrow$ wichtige Information geht verloren
- Annotation von Topologischen Feldern in TüBa-D/Z vermindert die Anzahl an Regeln in der Grammatik
  $\longrightarrow$ für den Parser leichter zu lernen?

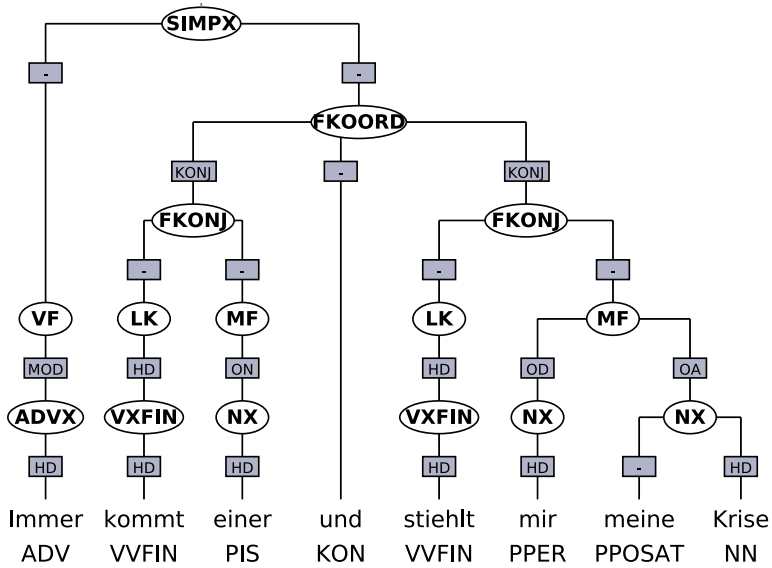
Einfluss der Annotationsschemata auf das Parsen spezifischer syntaktischer Phänomene

- Syntaktische Konstruktionen:
  - PP-Attachment
  - Extraponierte Relativsätze
  - Subjektlücken mit vorangestellten finiten Verben
  - Forward Conjunction Reduction
  - Koordinationen mit ungleichen Konstituenten
- Welches Annotationsschema ist besser geeignet, diese Konstruktionen zu parsen?

Subjektlücken mit vorangestellten finiten Verben (TiGer vs. TüBa-D/Z)



Statt dessen leugnet **man** Tatsachen und verdreht sie.



Immer kommt **einer** und stiehlt mir meine Krise.

Ergebnis

- TIGER profitiert von der flachen Annotation, die die Struktur für den Parser transparenter macht (z.B. für ERC, FCR und SGF)
- Die tiefere Baumstruktur in TüBa-D/Z macht es dem Parser schwer, Hinweise zu entdecken, da diese oft zu tief eingebettet sind
- Die Kodierung von nicht-lokalen Abhängigkeiten mit Hilfe von Funktionslabeln wird vom Parser oft nicht bzw. falsch gelernt
- Die zusätzliche Annotationsebene der topologischen Felder in TüBa-D/Z erhöht die Anzahl an möglichen Anhängungspositionen und so die Fehlerwahrscheinlichkeit
- Gleichzeitig reduziert sie die Anzahl an Regeln in der Grammatik und verbessert so die Lernbarkeit besonders für kleine Trainingssets

Überblick über verschiedene Arten linguistischer Annotation

1. Linguistisch annotierte Korpora
   - Korpora - Geschichtlicher Überblick
   - Baumbanken - Überblick
   - Baumbanken fürs Deutsche
   - Treebanking

Treebanking: Wie baue ich eine Baumbank

- Treebanking ist extrem zeitaufwendig und kostenintensiv:

"Creating the requisite training corpus, or treebank, is a Herculean task" *Eugene Charniak (1997)*

"Maximum working time on this task: 4-5 hours per day.
Else danger of going crazy." *Martin Volk (2006)*

Daher: gutes Design und breite Anwendbarkeit der Baumbank wichtig!

Treebanking: Wie baue ich eine Baumbank

- Treebanking ist extrem zeitaufwendig und kostenintensiv:

"Creating the requisite training corpus, or treebank, is a Herculean task" *Eugene Charniak (1997)*

"Maximum working time on this task: 4-5 hours per day. Else danger of going crazy." *Martin Volk (2006)*

Daher: gutes Design und breite Anwendbarkeit der Baumbank wichtig!

Treebanking: Erstellen von Baumbanken

- Design der Baumbank hängt ab vom beabsichtigen Zweck (sollte möglichst breit definiert sein aufgrund der hohen Kosten)
- Wichtige Designpunkte:
  - Textauswahl (Textsorte, gesprochen/geschrieben, Repräsentativität, ...)
  - Linguistische Theorie hinter der Annotation
    - ★ sollte möglichst theorieneutral sein, damit für viele nutzbar
    - ★ aber: wie sieht theorie-neutrale Syntax aus?
    - ★ außerdem: Theorie erhöht die Konsistenz
  - Was wird annotiert? (Detailliertheit der linguistischen Annotation vs. Konsistenz)
    - ★ Set mit nur 3 verschiedenen non-terminalen Kategorien ermöglicht hohe Konsistenz, ist aber nur begrenzt nützlich
- Extrem wichtig:
  - Inter-Annotator-Agreement (Konsistenz)
  - Dokumentation (Was wurde wie annotiert, wie wurde mit linguistischen Zweifelsfällen umgegangen?)

Fazit

- Syntaktisch annotierte Korpora (Baumbanken) als wichtige Ressource für
  - die theoretische Linguistik
  - die Computerlinguistik
- Syntaktische Annotation eines Korpus ist extrem zeitaufwändig und teuer, deshalb sollte man bei Design und Erstellung die nötige Sorgfalt walten lassen
- Das gewählte Annotationsschema hat einen großen Einfluss auf die Nützlichkeit der Ressource
- Treebank Trends
  - Baumbanken für immer mehr Sprachen
  - immer größerer Schwerpunkt auf Funktion der Annotation
  - mehr Ebenen der Annotation (Semantik, Diskurs, ...)
  - Parallele Baumbanken (z.B. für Maschinelle Übersetzung)

Referenzen I

- Baumbanken
  - Penn Treebank: `http://www.cis.upenn.edu/~treebank`
  - Susanne: `http://www.grsampson.net/RSue.html`
  - NEGRA: Skut, Wojciech, Brigitte Krann, Thorsten Brants, and Hans Uszkoreit. 1997. An annotation scheme for free word order languages. *In Proceedings of ANLP 1997*, Washington, D.C.
  - TIGER:
    - Brants, Sabine, and Silvia Hansen. 2002. Developments in the TIGER Annotation Scheme and their Realization in the Corpus. *In Proceedings of the Third Conference on Language Resources and Evaluation (LREC 2002)* pp. 1643-1649 Las Palmas.
    - Dipper, S., T. Brants, W. Lezius, O. Plaehn, and G. Smith. 2001. The TIGER Treebank. *In Third Workshop on Linguistically Interpreted Corpora LINC-2001*, Leuven, Belgium.
  - TüBa-D/Z: Telljohann, Heike, Erhard W. Hinrichs, Sandra Kübler, and Heike Zinsmeister. 2005. *Stylebook for the Tübingen Treebank of Written German (TüBa-D/Z).* Seminar für Sprachwissenschaft, Universität Tübingen, Germany.
  - POS-Tagging
    - Schiller, Anne, Simone Teufel, and Christine Thielen. 1995. *Guidelines für das Tagging deutscher Textcorpora mit STTS.* Technical Report, IMS-CL, University Stuttgart, 1995.

Referenzen II

- Treebanking
  - Volk (2006)
    www.ling.su.se/dali/education/courses/corp_ling$_o$n/Lect_07b$_T$reebank_Intro.p
    http://www.ling.su.se/DaLi/education/courses/treebank_course_2006/index.ht
  - Eugene Charniak. 1997. Statistical Parsing with a Context-free Grammar and Word Statistics. *Proceedings of the Fourteenth National Conference on Artificial Intelligence AAAI*.Menlo Park, Ca.
  - M. Marcus, B. Santorini, and M. Marcinkiewicz. 1993. *Building a large annontated corpus of english: the penn treebank.* Computational Linguistics, 19(2):313–330.
  - Natalie Schluter and Josef van Genabith. 2007. *Preparing, Restructuring and Augmenting a French Treebank: Lexicalised Parsing or Coherent Treebanks?*. The 10th Conference of the Pacific Association of Computational Linguistics PACLING 2007, 19-SEP-07 - 21-SEP-07, Melbourne Australia.
  - Anne Abeillé, François Toussenel, and Martine Chéradame. 2004. Corpus le monde: Annotations en constituants. guide pour les correcteurs. Technical report, LLF and UFRL and Université Paris 7.