Syntax natürlicher Sprachen

2: Phrasenstrukturgrammatik

A. Wisiorek

Centrum für Informations- und Sprachverarbeitung,
Ludwig-Maximilians-Universität München

24.10.2023

1. Lexikalische Kategorien

1.1. Wortarten-Klassifizierung

Lexikalische Kategorien

Lexikalische Kategorien = Wortarten / Parts-of-Speech

- Wort = **atomare syntaktische Einheit**
  $\rightarrow$ *terminale Konstituenten im Syntaxbaum*
- Wortart = Klasse von Wörtern mit gemeinsamen Eigenschaften
  $\rightarrow$ *sog. 'präterminale' Konstituenten im Syntaxbaum*

Lexikalische Regeln im Syntaxbaum

- Formalisierung der Zuordnung von Wörtern zu ihren Wortarten
- z.B. *ADJ* $\rightarrow$ *gut*
  $\rightarrow$ *Zuordnung des Wortes gut zu der POS-Kategorie Adjektiv*

Klassifikation nach verschiedenen Kriterien

- morphologische Klassifizierung
- syntaktische Klassifizierung
- semantische Klassifizierung

Morphologische Kriterien

Differenzierung über die Art ihrer grammatischen Merkmale

- **Flexionsparadigmen**: $\left\{ \begin{array}{l} \textit{Tür} \\ \textit{Welt} \end{array} \right\}$ *-en* vs. $\left\{ \begin{array}{l} \textit{geh} \\ \textit{steh} \end{array} \right\}$ *-e/st/t* *(*Welt-st)*

  $\rightarrow$ `Genus+Numerus` vs. `Person+Numerus`

flektierbar : deklinierbar : komparierbar : unflektierbar

- **Derivationsmorphologie**: $\left\{ \begin{array}{l} \textit{new} \\ \textit{beautiful} \end{array} \right\}$ *-ly*

  $\rightarrow$ *Adjektive bilden im Englischen in Kombination mit -ly Adverbien*

Syntaktische Kriterien

Differenzierung über Distribution

- Auftreten in gleichen Kontexten (distributionsäquivalent)
- z. B.: Adjektiv zwischen DET und NOUN oder nach Form von *sein*

Differenzierung über morphosyntaktisches Verhalten

- z.B. Präposition vs. Konjunktion (beide: unflektierbar)
  - Präposition: regiert Kasus in Umgebung
    → *wegen **des Hundes** ... (PP) vs. weil der Hund ... (CONJ)*

Differenzierung über syntaktische Funktion

- Prädikat : Subjekt : Objekt : Adverbial : Attribut

Semantische Kriterien

Differenzierung Wörter über ihre Bedeutung

- **Verb**: bezeichnet Zustände, Vorgänge, Tätigkeiten, Handlungen
- **Nomen**: bezeichnet Lebewesen, Sachen (Dinge), Begriffe (Abstrakta), Individuen
- **Adverb**: bezeichnet nähere Umstände von Sachverhalten
- **Adjektiv**: bezeichnet Eigenschaften und Merkmale von Sachen

weitere semantische Unterscheidungen

- **Auto- vs. Synsemantika**
  - **Inhaltswörter**: selbständige lexikalische Bedeutung; satzgliedfähig (Funktion als Phrasenkopf)
  - **Funktionswörter**: grammatische Bedeutung (abhängig von Bezugswort); nicht satzgliedfähig
- **offene vs. geschlossene Klassen**
  - endliche/abgeschlossene vs. potentiell unendliche Menge

Syntaktische vs. semantische Kategorisierung

traditionelle Grammatik: semantische und morphologische Wortklassifizierung

- Nomen, von lat. *nomen*: Namen einer Sache/Person/Ort
- Adjektiv: Eigenschaftswort, flektierbar (im Gegensatz zum Adverb)
- usw.

Prototypische Wortarten syntaktischer Funktionen

- Prädikat: Verb
- Subjekt/Objekt: Nomen
- Adverbial: Adverb
- Attribut: Adjektiv

Problematik der Wortarten-Klassifikation

keine direkte Entsprechung Semantik - syntaktische Funktion

- z. B.: prototypisches Nomen kann syntaktisch Teil des Prädikats sein, also eine andere syntaktische Funktion erfüllen (Prädikativum): *Er ist Lehrer.*
- z. B.: Wörter mit nicht-nominaler Semantik können eine prototypische nominale Strukturposition einnehmen (z.B. als Subjekt fungieren): *Blau ist eine Farbe.*
- Adjektive können attributiv, adverbial oder prädikativ gebraucht werden (s.u.)

Wortarten sind außerdem sprachabhängig

- es gibt Sprachen, die keine Eigenschaftswortklasse haben (Dyirbal, Lakhota; s. VanValin 2000, 12)
  - die typische syntaktische Funktion, die in indogermanischen Sprachen Adjektive übernehmen (Attributfunktion), wird hier von Nomen (Dyirbal) bzw. Verben (Lakhota) übernommen

Moderne Linguistik: Syntaktische Wortarten

Strukturalismus

- Definition Wortklassen über (morpho)syntaktische Eigenschaften
- Bestimmung Klassenmitglieder über **syntaktisches Verhalten**
- = **Distributionalismus**

Syntaktisches Wortartkriterium

- Kriterium: Auftreten in gleichen Kontexten (**distributionsäquivalent**)
  - **Distribution = Menge der Kontexte**
- Beispiel NLTK:
  `text.similar('bought') > made done put said found`

Generative Grammatik

- ebenfalls syntaktisches Wortartkriterium
- Kriterium: Besetzung gleicher **Strukturpositionen**

1.2. Traditionelle Grammatik als Wortartengrammatik

Traditionelle Grammatik als Wortartengrammatik

traditionelle Grammatik

- **histor.: Acht-Wortarten-Lehre (Dionysios Thrax, 2. Jhd. v. Chr.):**
  - Nomen, Verb, Partizip, Adverb, Pronomen, Artikel, Präposition, Konjunktion
- reine Wortarten-Syntax (**nur lexikalische Kategorien**)
- keine Phrasenebene (**keine (höheren) syntaktischen Kategorien**)

Problem

- ohne Phrasenebene: sehr viele Satzschemata: $S \rightarrow \left\{ \begin{array}{l} \text{Satzschema 1} \\ \ldots \end{array} \right\}$

Lösung: Phrasenstrukturgrammatik

- wenige Phrasenstruktur-Regeln können große Anzahl an Satzschemata generieren
- durch rekursive Regeln: unendlich viele Satzschemata generierbar

Beispiel Generierung Satzschemata

Phrasenstrukturregeln (6)

$S \rightarrow NP\ VP$
$VP \rightarrow V$
$VP \rightarrow V\ NP$
$VP \rightarrow V\ NP\ NP$
$NP \rightarrow DET\ N$
$NP \rightarrow N$

POS-Satzschemata (14)

$S \rightarrow DET\ N\ V$
$S \rightarrow DET\ N\ V\ DET\ N$
$S \rightarrow DET\ N\ V\ N$
$S \rightarrow DET\ N\ V\ DET\ N\ DET\ N$
$S \rightarrow DET\ N\ V\ DET\ N\ N$
$S \rightarrow DET\ N\ V\ N\ DET\ N$
$S \rightarrow DET\ N\ V\ N\ N$
$S \rightarrow N\ V$
$S \rightarrow N\ V\ DET\ N$
$S \rightarrow N\ V\ N$
$S \rightarrow N\ V\ DET\ N\ DET\ N$
$S \rightarrow N\ V\ DET\ N\ N$
$S \rightarrow N\ V\ N\ DET\ N$
$S \rightarrow N\ V\ N\ N$

Motivation für Klassifizierung in Syntaxanalyse

Wort- und Phrasenklassen-basierte Schemata

- ökonomische und adäquate **Modellierung hierarchische Struktur**:
  → *ökonomisch: viele Satzschemata durch wenige Regeln generierbar*
  → *beschreibungsadäquat: Phrasen empirisch feststellbar*

Lexikalische Regeln: `N → 'Hund' | 'Katze'`

- Zuordnung lexikalischer Einheiten (Wörter) zu ihren **lexikalischen Kategorien** (Wortarten, Part-of-Speech-Kategorien)
  → *POS = Klassen sich syntaktisch gleichverhaltender Wörter*
  → *Wörter austauschbar im selben Kontext*

Syntaktische Regeln: `NP → DET N | N`

- Regeln der Kombination von lexikalischen Kategorien (Wortarten) zu komplexeren syntaktischen Einheiten (Konstituenten, Phrasen, Sätze)
  → *Wortgruppen austauschbar im selben Kontext*

2. Syntaktische Kategorien

1. Lexikalische Kategorien
   - Wortarten-Klassifizierung
   - Traditionelle Grammatik als Wortartengrammatik
2. Syntaktische Kategorien
   - Konstituentenstruktur
   - Eigenschaften von Phrasen
   - Phrasenstruktur
3. Phrasenstrukturgrammatik
   - Formale Grammatik
   - Kontextfreie Grammatik
4. Tagsets

2.1. Konstituentenstruktur

1. Lexikalische Kategorien
   - Wortarten-Klassifizierung
   - Traditionelle Grammatik als Wortartengrammatik
2. Syntaktische Kategorien
   - Konstituentenstruktur
   - Eigenschaften von Phrasen
   - Phrasenstruktur
3. Phrasenstrukturgrammatik
   - Formale Grammatik
   - Kontextfreie Grammatik
4. Tagsets

Eigenschaften der Konstituentenstruktur

Konstituentenanalyse

auch IC-Analyse (Analyse der *immediate constituents*)

1. **Zerlegung** syntaktischer Einheit in ihre Teile (Konstituenten)
2. Bildung von **Konstituentenklassen** (lexikalische und syntaktische Kategorien)
   - Ermittlung über **Konstituententests**
   - Ergebnis ist eine **hierarchisch gegliederte Struktur**

unmittelbare Konstituenten (*immediate constituents*)

unmittelbare Konstituenten sind die **maximalen Konstituenten** einer Einheit (aus denen sie unmittelbar zusammengesetzt ist)

Konstituenz-Relation

Konstituenz

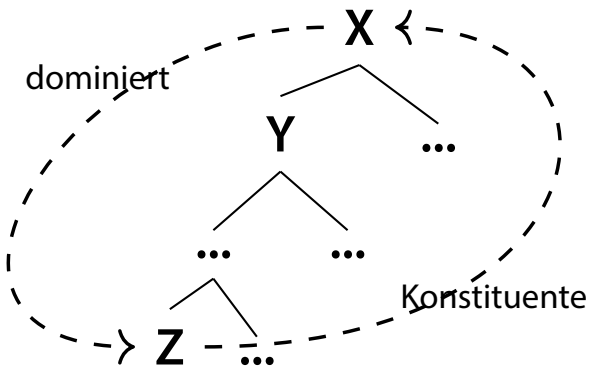
Teil-Ganzes-Beziehung zwischen sprachlichen Einheiten (Konstituenten)

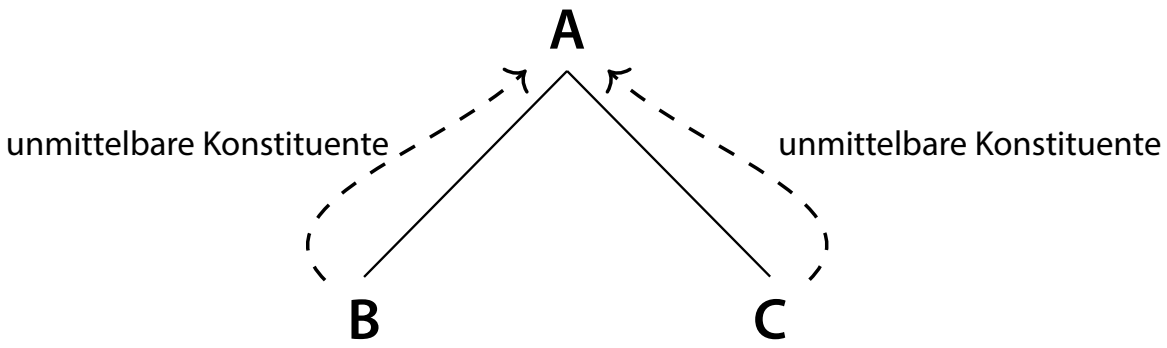
unmittelbare Dominanz

Beziehung der **unmittelbaren Dominanz** zwischen Einheit und ihren unmittelbaren Konstituenten

Dominanz

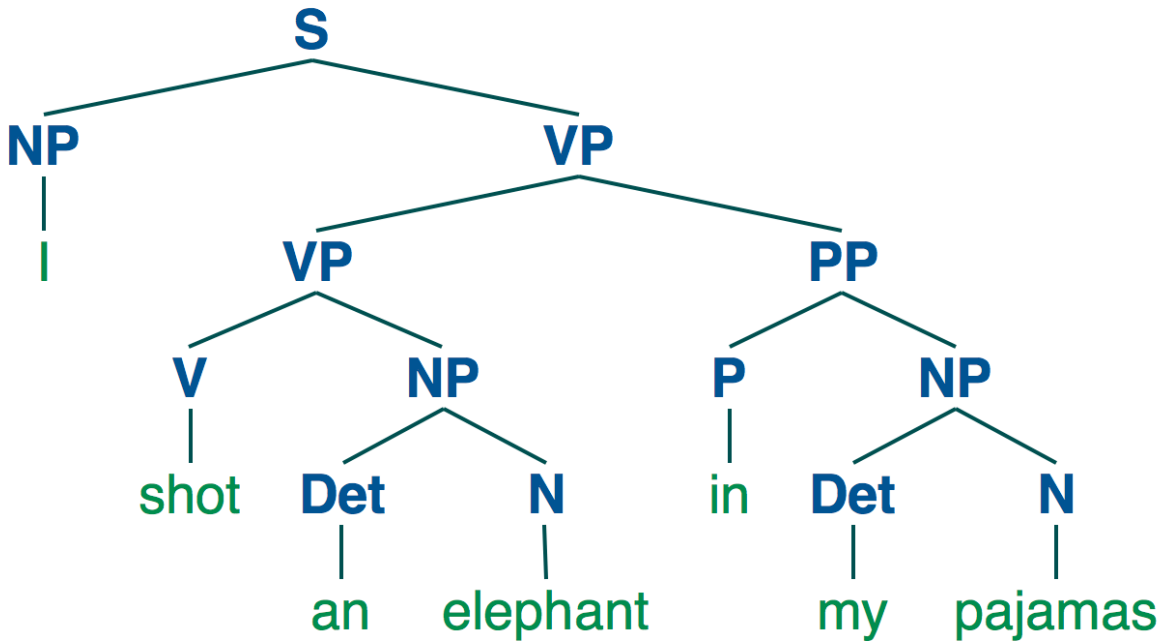
Beziehung der **Dominanz** zwischen Einheit X und der unmittelbare Konstituente Y; sowie zwischen X und Z, wenn Y Z dominiert (transitive Relation)





Konstituentenstruktur

- Menge der **durch die Relation der unmittelbaren Dominanz verbundenen Konstituenten**
- durch Bezug auf **Konstituentenklassen** (lexikalische und syntaktische Kategorien als Knoten) und Abstraktion von der Wortebene ergeben sich **Konstituentenschemata**



Übersicht Konstituentenstruktur

Elemente der Struktur (Knoten)

- **Wörter** $\rightarrow$ *terminale Knoten*
- **lexikalische Kategorien** $\rightarrow$ *präterminale Knoten*
- **syntaktische Kategorien** $\rightarrow$ *nichtterminale Knoten*

Relationen der Struktur (Kanten)

- Teil-Ganzes-Beziehung
- unmittelbare Dominanz des Mutterknotens über Tochterknoten

Strukturinformationen in Knoten des Syntaxbaums!

2.2. Eigenschaften von Phrasen

Klassifizierung von Konstituenten

Phrasen als spezielle Konstituentenklassen

- im gleichen Kontext austauschbare Konstituenten bilden **Konstituentenklassen**
- Phrasen sind spezifische Klassen von Konstituenten, die im Satz eine ähnliche syntaktische Funktion erfüllen
- die syntaktische Funktion wird primär vom sog. Phrasenkopf bestimmt, dem Kern der Phrase

Phrase

- Phrase = Konstituente, in der ein (Inhalts-)Wort als **Phrasenkopf um Wörter oder Phrasen erweitert** ist
  $\rightarrow$ *in Terminologie der Generativen Grammatik:* ***maximale Projektion***
- Eine Phrase ist also eine Konstituente, die eine sinnvolle Einheit bilden, wobei der Phrasenkopf das zentrale Wort ist, das die **grammatische Rolle und Bedeutung der gesamten Phrase** bestimmt

Eigenschaften der Phrasenstruktur

Hierarchischer rekursiver Strukturaufbau

- minimal besteht die Phrase nur aus ihrem Kopf (Phrasenkern: $VP \rightarrow V$)
- Phrasen können andere Phrasen als Subkonstituenten enthalten, z.B.:
  - Verbalphrase (VP): $VP \rightarrow V\ NP$
  - Nominalphrase (NP): $NP \rightarrow DET\ N\ PP$
  - Präpositionalphrase (PP): $PP \rightarrow P\ NP$
- Phrasen können rekursiv aufgebaut sein (vgl. die obigen Regeln für NP und PP mit (indirekter) rekursiver PP-Einbettung), z.B.:
  - *(S The mailman ate his (NP lunch (PP with his friend (PP from the cleaning staff (PP of the building (PP at the intersection (PP on the north end (PP of town)))))))*
  - Beispiel von `https://people.cs.umass.edu/~mccallum/courses/inlp2007/lect5-cfg.pdf`

Beispiel

- In der Nominalphrase *das rote Auto* ist *rote Auto* eine Ko-Konstituente
- *rote Auto* ist aber selbst keine eigenständige Phrase (unvollständig)
  - ohne Artikel z.B. nicht im Satz als Subjekt verwendbar

Phrasentypen

Phrasenbildende Wortarten

- nur Autosemantika (Inhaltswörter) sind phrasenbildend
- als Kern (Kopf) einer Phrase kann also nur ein (Pro)nomen, ein Verb, eine Präposition, ein Adjektiv oder ein Adverb auftreten

Phrasentypen

- Nominalphrase (NP): *das kleine **Kind***
- Verbalphrase (VP): *nach Hause **fahren***
- Präpositionalphrase (PP): *mit dem **Kind***
- Adjektivphrase (ADJP): *sehr **klein***
- Adverbphrase (ADVP): *sehr **oft***

Kopfprinzip

Phrasenkopf

- alle Wörter und Phrasen in der Phrase sind zum Kopf **dependent**
- Kopf vererbt **morphosyntaktische Merkmale** an Phrase (Kasus usw.)
- Kopf steuert **syntaktisches Verhalten** der Konstituente im Satz
- Kopf bestimmt die **Phrasenkategorie** (Wortart X $\rightarrow$ Phrasenkat. XP)

Beispiel Phrasenkopf

- in der Verbalphrase *fährt im rote Auto davon* ist das Verb *fährt* der Kopf der Phrase
- die PP *im roten Auto* und die Adverbialphrase ADVP *davon* sind Adjunkte der VP
- *fährt* kann auch alleine eine VP bilden: *Er fährt.*
- das Subjektagreement wird alleine vom Verb gesteuert:
  Er fährt davon. vs. *Sie fahren davon.*

Kopf-Perkolation

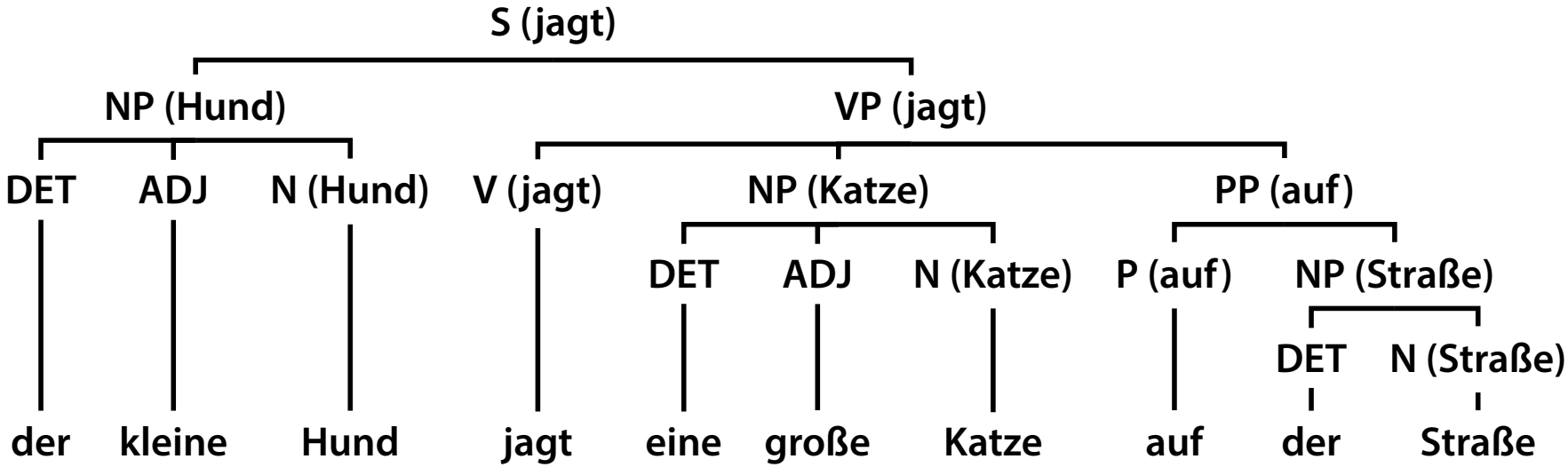
Kopf-Perkolation

- Köpfe werden im Syntaxbaum nach oben weitergereicht (da hierarchische Struktur, Phrasen in andere eingebettet)
- wichtig u.a. für **lexikalisierte Grammatiken** sowie die **Transformation einer Phrasenstruktur- in eine Dependenzgrammatik**

Kopf-Perkolations-Regeln

- head(S) = head(VP)
  $\rightarrow$ *Kopf von S ist Kopf von VP*
- head(VP) = head(V)
- head(V) = *jagt*

Kopf-annotierter Beispielsatz



2.3. Phrasenstruktur

1. Lexikalische Kategorien
   - Wortarten-Klassifizierung
   - Traditionelle Grammatik als Wortartengrammatik
2. Syntaktische Kategorien
   - Konstituentenstruktur
   - Eigenschaften von Phrasen
   - Phrasenstruktur
3. Phrasenstrukturgrammatik
   - Formale Grammatik
   - Kontextfreie Grammatik
4. Tagsets

Schema Einfacher Satz

allgemeines Satzschema: S = NP + VP

- Ergebnis von Konstituententests (Reduktion auf Zweiwort-Satz)
- **Subjekt-NP und Verb interdependent**, also gegenseitig abhängig (sichtbar am **Verb-Agreement**)
- Subjekt (Satzgegenstand) - Prädikat (Satzaussage)
- abstrahiert von linearer Ordnung: **Wortstellung sprachabhängig**

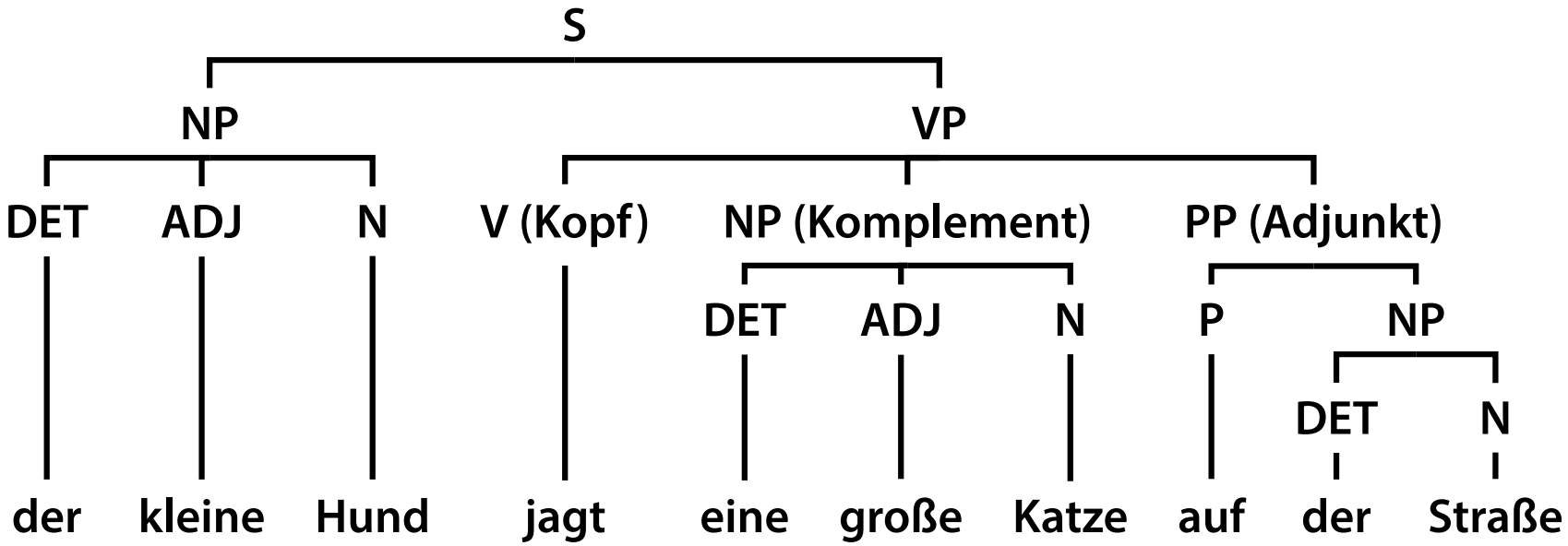
VP = VERB + Komplemente + Adjunkte

- Komplemente = obligatorische (valenzgeforderte) Erweiterungen
- Adjunkte = nicht-obligatorisch Erweiterungen, Anzahl nicht begrenzt

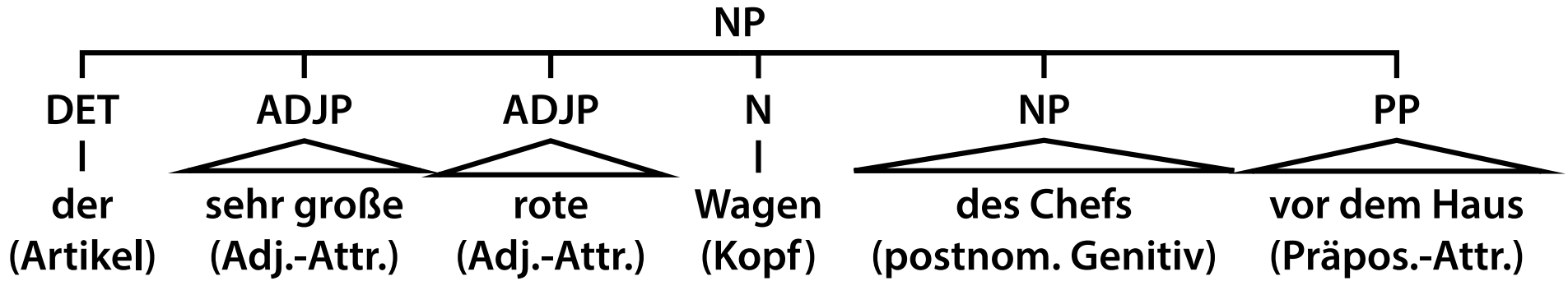
NP = NOUN + nominale Adjunkte (Attribute)

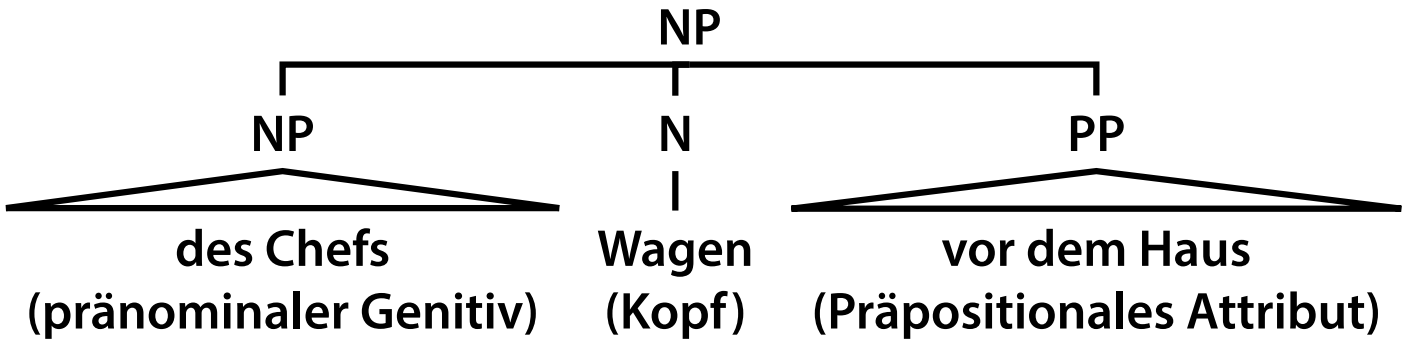
- Links- und Rechtserweiterungen um Nomen (als NP-Kopf)

Phrasenstruktur Aussagesatz (mit flachen Syntaxregeln)

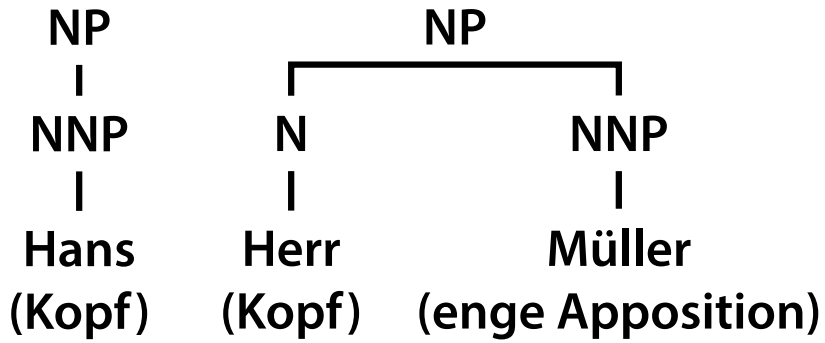


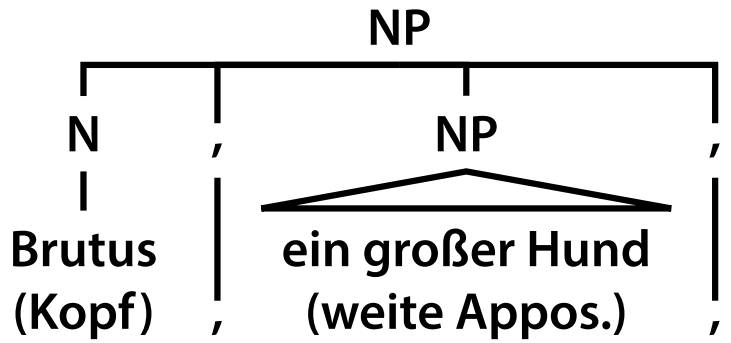
Links- und Rechtsattribute der NP im Deutschen

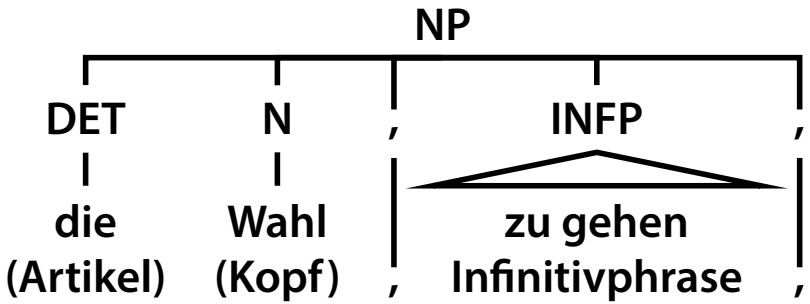


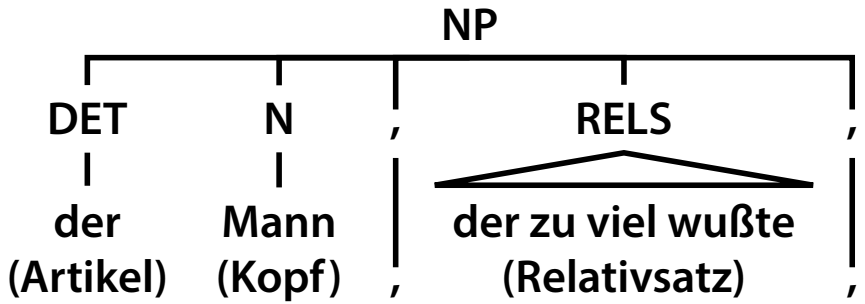


Weitere NP-Phrasenstrukturen des Deutschen









3. Phrasenstrukturgrammatik

3.1. Formale Grammatik

Mögliche Methoden für Syntaxanalyse

Beschreibung des Sprachsystems

- traditionelle Buch-Grammatik
- nicht-computational!

Aufzählung aller grammatischen Sätze

- Problem 1: natürliche Sprachen sind unendlich
- Problem 2: Struktur nicht repräsentiert

Beschreibung durch *formale Grammatik*

- mathematisches Modell des syntaktischen Regelsystems
- computational!
- ermöglicht die Analyse der Struktur einer unendlichen Menge an Sätzen mit endlichen Mitteln

Formale Grammatik

Formale Grammatik

- **mathematisches Regelsystem**, das verwendet wird, um eine **formale Sprache** eindeutig zu beschreiben und zu erzeugen
- dient zur **Generierung aller wohlgeformten Ausdrücke** einer Sprache und wird oft als *generative Grammatik* bezeichnet

Ableitungsregeln

- eine formalen Grammatik verwendet Ableitungsregeln (auch **Produktionsregeln**), die von einem **Startsymbol** ausgehend
  - die linke Regelseite (**LHS** = *left-hand side*)
  - durch die rechte Regelseite (**RHS** = *right-hand side*) ersetzen

Formale Grammatik zur Syntaxanalyse

- neben der Erzeugung *formaler Sprachen* kann eine formale Grammatik auch als Modell zur **Erkennung** und **Analyse** der syntaktischen Struktur ***natürlicher* Sprachen** dienen

Formale Sprache

Formale Sprache

- Menge aller Ausdrücke, die mithilfe der Regeln einer formalen Grammatik aus **Nichtterminalsymbolen** abgeleitet werden können
- wobei die Ausdrücke aus den **Grundsymbolen** der Sprache bestehen

Grundsymbole (Terminalsymbole)

- z. B. $\{a, b, c\}$
- in Modellierung natürlicher Sprache: **Wörter** des Lexikons, z. B. $\{die, der, den, Hund, Katze, jagt\}$

Ableitbare Ausdrücke

- z. B. $\{a, aa, aba, abcc, \ldots\}$
- = **formalsprachliche Wörter** (die Blätter des Ableitungsbaums)
- in Modellierung natürlicher Sprache: natürlichsprachliche **Sätze**, z. B. $\{$*"der Hund jagt die Katze"*, *"die Katze jagt den Hund"*, $\ldots\}$

Nichtterminalsymbole

Nichtterminale

- werden in den Regeln der formalen Grammatik definiert
- kommen nur in Zwischenschritten der Ableitung vor

Nitterminale in der Syntaxanalyse

- in der Syntaxanalyse entsprechen sie den syntaktischen Kategorien (Satz, Phrasen, Wortarten)
- Beispiele sind: S, NP, VP, DET, N, V
- Wortarten: werden auch Präterminale genannt

Vorteile einer Modellierung mit formaler Grammatik

Mächtigkeit

Unendliche Menge an Sätzen mit endlichen Mitteln beschreibbar.

Automatisierung

Rechnergestützt verarbeitbar durch Parsingalgorithmen.

Sprachkomplexität

Beantwortung von Fragen zur Komplexität natürlicher Sprache (Chomsky-Hierarchie).

Psycholinguistik

Anwendung als Modell menschlicher Sprachverarbeitung.

3.2. Kontextfreie Grammatik

Chomsky-Hierarchie

Typ 0: Unbeschränkte Grammatiken (Rekursiv aufzählbar)

- Regeln der Form $\alpha \rightarrow \beta$, wobei α, γ beliebige Folge von Terminal- und Nichtterminalsymbolen

Typ 1: Kontextsensitive Grammatiken

- Regeln der Form $\alpha A \beta \rightarrow \alpha \gamma \beta$, wobei A ein Nichtterminal ist (α, β, γ beliebig)

Typ 2: Kontextfreie Grammatiken (CFG)

- Regeln der Form $A \rightarrow \gamma$, wobei A ein Nichtterminal ist (γ beliebig)
- **Beschränkung:** links (LHS) genau ein Nichtterminal, rechts (RHS) beliebig

Typ 3: Reguläre Grammatiken

- Einfachste Art von Grammatiken (am stärksten beschränkt)
- Regeln der Form $A \rightarrow \alpha B$, wobei A und B Nichtterminale und α eine Folge von Terminalsymbolen sind.
- **Weitere Beschränkung:** rechts (RHS) nur ein Nichtterminal möglich!

Asdrucksfähigkeit formaler Grammatiken

- Typ 0 (rekursiv aufzählbar, Turingmaschinen) ist am ausdrucksfähigsten
- die Ausdrucksfähigkeit nimmt mit jedem Typ ab
- Typ 3 (regulär, endliche Automaten) ist am stärksten eingeschränkt

Phrasenstrukturgrammatik

Kontextfreie Grammatik als Phrasenstrukturgrammatik (PSG)

- Chomsky: die **Konstituenten- bzw. Phrasenstruktur** natürlicher Sprache ist formal beschreibbar durch **kontextfreie Grammatiken**
- Strukturregeln der (auch rekursiven) **Kombination von lexikalischen und phrasalen Kategorien** zu phrasalen Kategorien und Sätzen
  - z.B. $S \rightarrow NP\ VP$, $NP \rightarrow DET\ ADJ\ N$

Syntaktische Regeln ($NP \rightarrow DET\ N\ PP$)

bestimmen, zu welchen Klassen die unmittelbaren Konstituenten (RHS) einer syntaktischen Kategorie (LHS) gehören (*LHS+RHS jeweils **Nichtterminalsymbole***)

Lexikalische Regeln ($N \rightarrow Hund$)

bestimmen die Zugehörigkeit einer elementaren Konstituente (RHS) (= *Wort, **Terminalsymbol***) zu einer lexikalischen Kategorie (LHS) (= *Wortart, Nichtterminalsymbol, manchmal auch **Präterminalsymbol** genannt*)

Beispiel einer CFG-Phrasenstrukturgrammatik

```
1 | ###### Syntaktische Regeln ######
2 | S → NP VP
3 | NP → DET N
4 | VP → V NP NP
5 |
6 | ###### Lexikalische Regeln ######
7 | DET → "der" | "die" | "das"
8 | N → "Mann" | "Frau" | "Buch"
9 | V → "gibt" | "schenkt"
```

Parsingergebnis NLTK

```
1
2  for tree in parser.parse(sent):
3      print(tree);tree.pretty_print()
4  #(S
5  #  (NP (DET der) (N Mann))
6  #  (VP (V gibt) (NP (DET der) (N Frau)) (NP (DET das) (
       N Buch))))
7
8                        S
9           _____________|_______
10         |                     VP
11         |            _________|________
12        NP           |         NP       NP
13     ___|___         |      ___|___   ___|___
14  DET      N      V   DET      N    DET     N
15   |       |      |    |       |     |      |
16  der     Mann   gibt  der    Frau  das    Buch
```

Aufbau Phrasenstrukturgrammatik

Grammatikdefinition

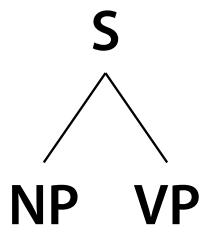
- **Startsymbol:** S
- **Nichtterminalsymbole:** NP, VP, DET, N, V
- **Terminalsymbole:** *der, Hund, schläft*
- **Produktionsregeln:** $S \rightarrow NP\ VP, NP \rightarrow DET\ N, VP \rightarrow V$

Hinweise zur Grammatik

- Ersetzungsregeln (linke mit rechter Seite)
- CFG-Regel-Einschränkung (Chomsky-Hierarchie):
  - links nur ein Nichtterminalsymbol
    $\rightarrow$ *Ersetzung unabhängig von Kontext (Kontextfreiheit)*

PSG-Regeln als Produktionsregeln

- PSG-Regeln können als **Konstruktionsanweisung für Syntaxbäume** interpretiert werden:
  $S \rightarrow NP\ VP$ als *'expandiere S zu Folge NP + VP'*



- PSG-Regel definiert **Relation der unmittelbaren Dominanz zwischen Mutterknoten und Tochterknoten**
  → *'S dominiert unmittelbar NP und VP'*
  → *'S dominiert vollständig die Folge NP + VP'*
  → *und: 'NP und VP sind Ko-Konstituenten'* (***Geschwisterknoten***)
- PSG **erkennt durch Ableitung** Sätze als zur Sprache gehörig und weist ihnen die ihren Regeln entsprechende **Strukturbeschreibung** zu
  → *Strukturbeschreibung = die auf Kategorien bezugnehmende* ***Konstituentenstruktur***
  → *'Die Folge NP + VP ist ein S'*

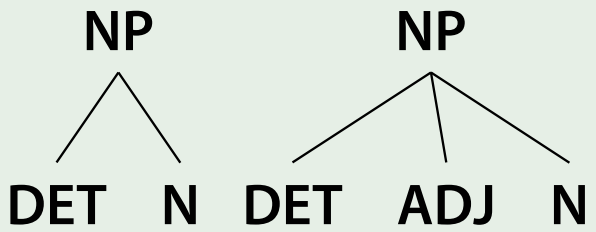
Disjunktionsoperator

Disjunktionsoperator

- der Disjunktionsoperator | wird verwendet, um alternative Ableitungen auszudrücken
- Abkürzung für zwei Regeln mit der selben LHS
- kann in den CFGs von NLTK verwendet werden

Beispiel

- $NP \rightarrow DET\ N \mid DET\ ADJ\ N$
- äquivalent zu:
  $NP \rightarrow DET\ N$
  $NP \rightarrow DET\ ADJ\ N$



Konvention für fakultative Elemente

- im Beispiel ist ADJ fakultativ
- kann auch folgendermaßen geschrieben werden: $NP \rightarrow DET\ (ADJ)\ N$

Ableitung als top-down Erzeugung

Formale Definition einer kontextfreien Grammatik

$G\,(Grammatik) = < T\,(Terminale), N\,(Nichtterminale), S\,(Startsymbol), R\,(Regeln) >$

Beispielableitung

$G = < \{das, Tier, Futter, sieht\}, \{S, NP, VP, DET, N, V\}, S, R >$

$R = \{S \rightarrow NP\ VP, NP \rightarrow DET\ N, VP \rightarrow V\ NP, DET \rightarrow das, N \rightarrow Tier, N \rightarrow Futter, V \rightarrow sieht\}$

S	$\Rightarrow$	NP VP	(S $\rightarrow$ NP VP)
	$\Rightarrow$	DET N VP	(NP $\rightarrow$ DET N)
	$\Rightarrow$	das N VP	(DET $\rightarrow$ das)
	$\Rightarrow$	das Tier VP	(N $\rightarrow$ Tier)
	$\Rightarrow$	das Tier V NP	(VP $\rightarrow$ V NP)
	$\Rightarrow$	das Tier sieht NP	(V $\rightarrow$ sieht)
	$\Rightarrow$	das Tier sieht DET N	(NP $\rightarrow$ DET N)
	$\Rightarrow$	das Tier sieht das N	(DET $\rightarrow$ das)
	$\Rightarrow$	das Tier sieht das Futter	(N $\rightarrow$ Futter)

4. Tagsets

1. Lexikalische Kategorien
   - Wortarten-Klassifizierung
   - Traditionelle Grammatik als Wortartengrammatik
2. Syntaktische Kategorien
   - Konstituentenstruktur
   - Eigenschaften von Phrasen
   - Phrasenstruktur
3. Phrasenstrukturgrammatik
   - Formale Grammatik
   - Kontextfreie Grammatik
4. Tagsets

POS-Tagsets

- **Tagset** = Sammlung von Kategorienlabels
- traditionelle Wortart-Analysen: wenige lexikalische Kategorien
- in Korpuslinguistik/Computerlinguistik: umfangreichere Tagsets
  $\rightarrow$ *umfassen z. T. auch morphologische Kriterien*

Bekannte POS-Tagsets

- **Universal POS-Tagset**: 17 POS-Tags
  $\rightarrow$ *erste Angabe bei Wortarten oben*
- **Penn Treebank POS-Tagset**: 45 POS-Tags
  $\rightarrow$ *vereinfachtes **Brown Corpus POS-Tagset** (87 POS-Tags)*
  $\rightarrow$ *zweite Angabe bei Wortarten oben*
- **TIGER/STTS-POS-Tagset**: 53 POS-Tags (deutsch)

Syntaktische Tagsets

Konstituenten-Tagesets

- **Penn-Treebank Constituent Tags**
- **TIGER Konstituenten Labels**

Dependency-Tagsets

- **UD-Tagset (Universal Dependencies)**
- **TIGER-Dependencies**