Einführung Computerlinguistik

Konstituentensyntax II

Hinrich Schütze & Robert Zangenfeind

Centrum für Informations- und Sprachverarbeitung, LMU München

2013-11-18

Take-away

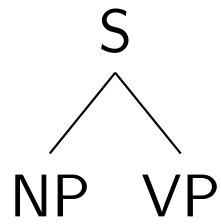
- Phrasenstrukturgrammatik: Formaler Ansatz zur Darstellung der Struktur von Sätzen
- Zwei Arten von Regeln: Phrasenstrukturregeln, Lexikonregeln
- Anwendungen von Syntax in der Computerlinguistik: Parsing, Maschinelle Übersetzung etc.

Overview

Outline

Phrasenstrukturgrammatik

- Beziehungen der unmittelbaren Dominanz werden als Regeln zur Erzeugung von Sätzen (Konstituentenstrukturen) aufgestellt , z.B.:



- Regel (rewriting rule, Phrasenstrukturregel) hierzu:
- $S \rightarrow NP \quad VP$
- zu lesen als: “expandiere S zu . . . / ersetze S durch . . . / S dominiert unmittelbar . . . / schreibe S um zu . . . ”
- S: Startsymbol (“Satz”)

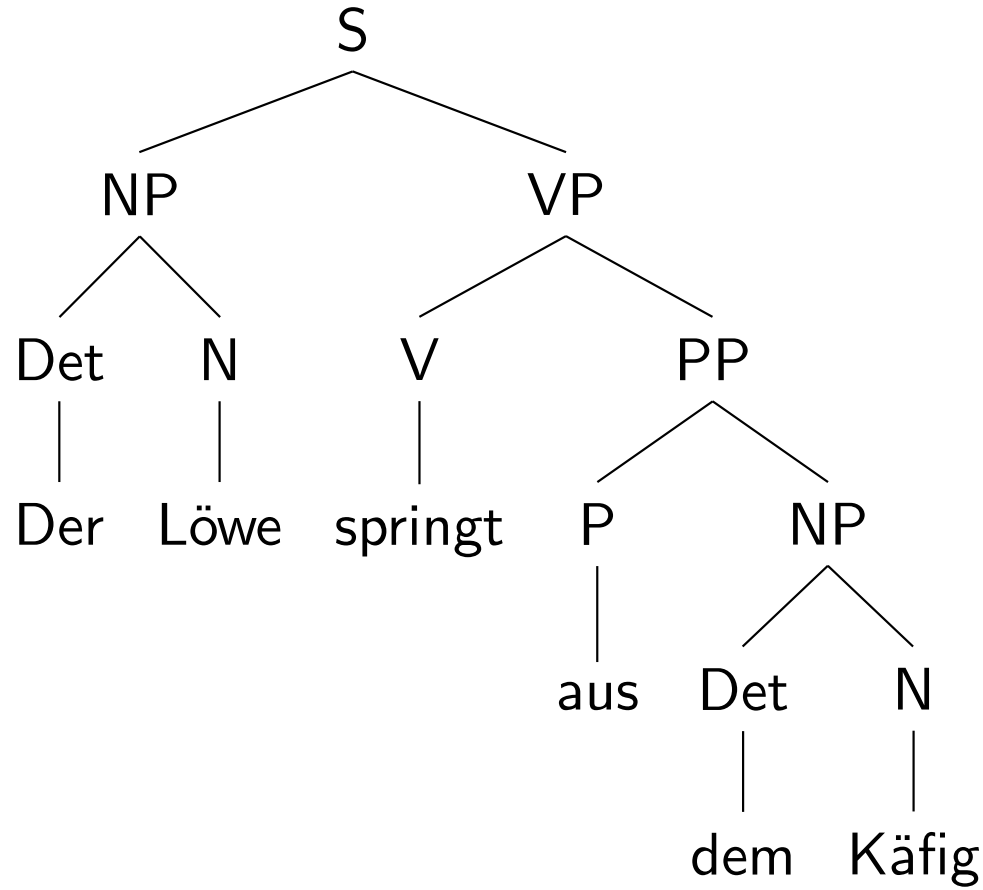
Schema für Phrasenstrukturregeln

Allgemeines Schema einer PS-Regel: X $\rightarrow$ (W) Y (Z) (W–Z: Kategorialsymbole)

- X: Eingabesymbol (genau ein solches steht links des Pfeils)
- W, Y, Z: Ausgabesymbole
- Y: obligatorisches Element
- W, Z: fakultative Elemente

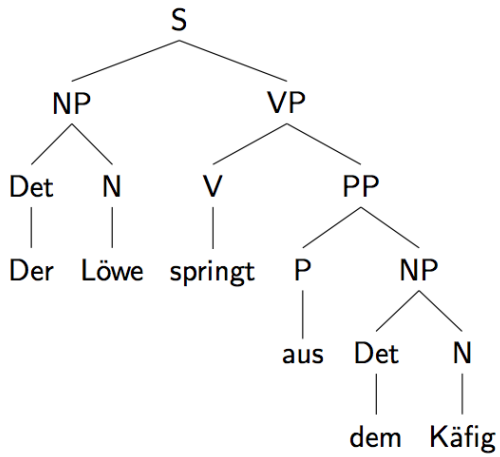
PS-Regeln sind kontextfrei, d.h. genau ein Nichtterminalsymbol (links) wird expandiert zu beliebiger Folge von Nichtterminal- und Terminalsymbolen

Beispiel für Phrasenstrukturregeln (1)



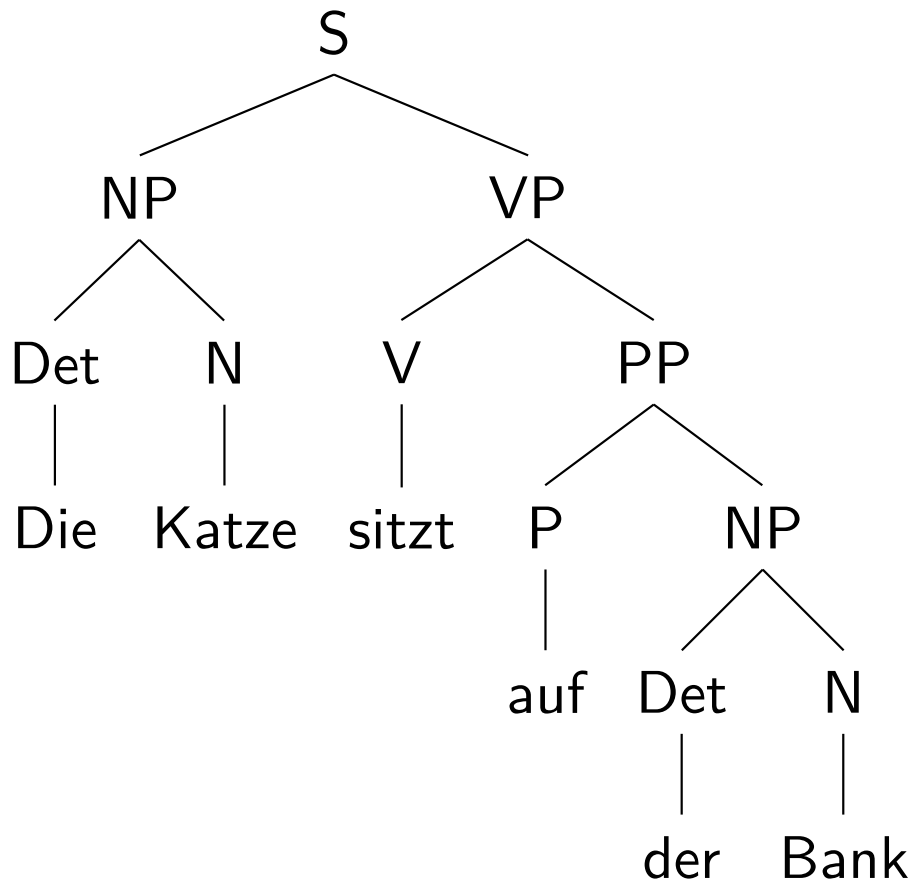
$S \rightarrow NP \quad VP; \qquad NP \rightarrow Det \quad N \; VP \rightarrow V \quad PP; \qquad PP \rightarrow P \; NP$

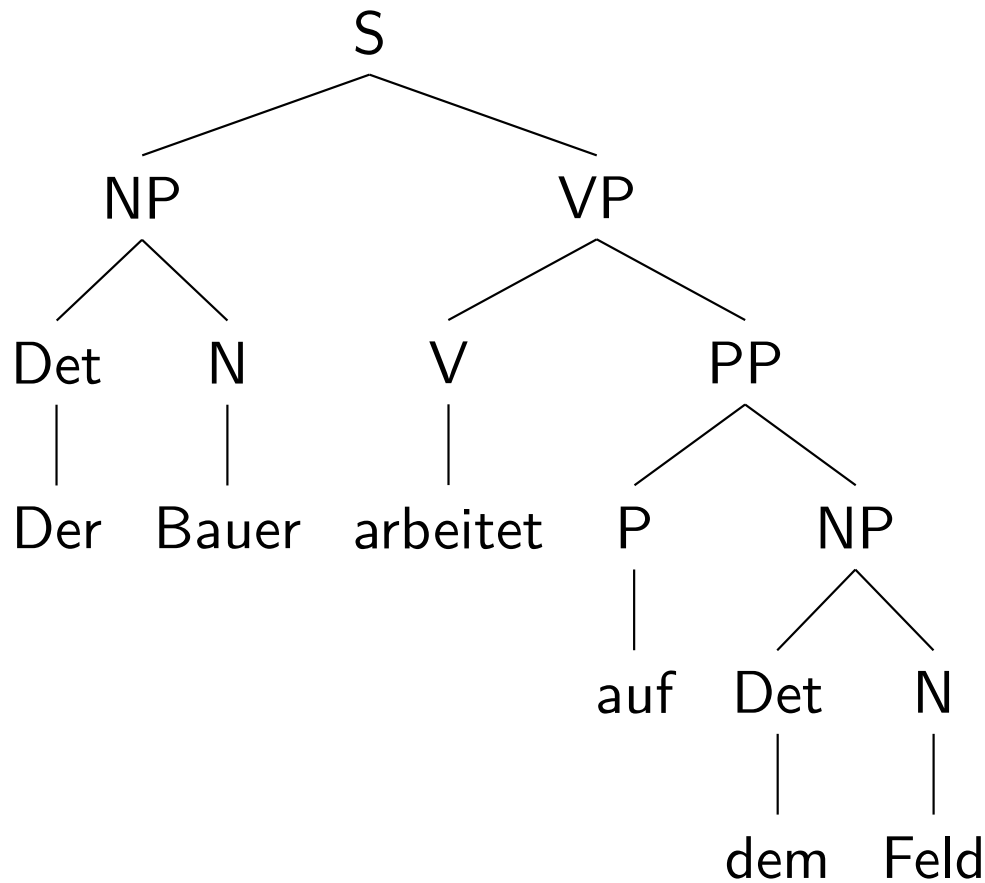
Beispiel für Phrasenstrukturregeln (2)



$S \rightarrow NP \quad VP$; $NP \rightarrow Det \quad N$ $VP \rightarrow V \quad PP$; $PP \rightarrow P \quad NP$ Mit diesen Regeln kann obige Baumstruktur bis zur Ebene der Wortkategorien (präterminale Kette) generiert werden. Außerdem möglich:

- Die Katze sitzt auf der Bank
- Der Bauer arbeitet auf dem Feld etc.

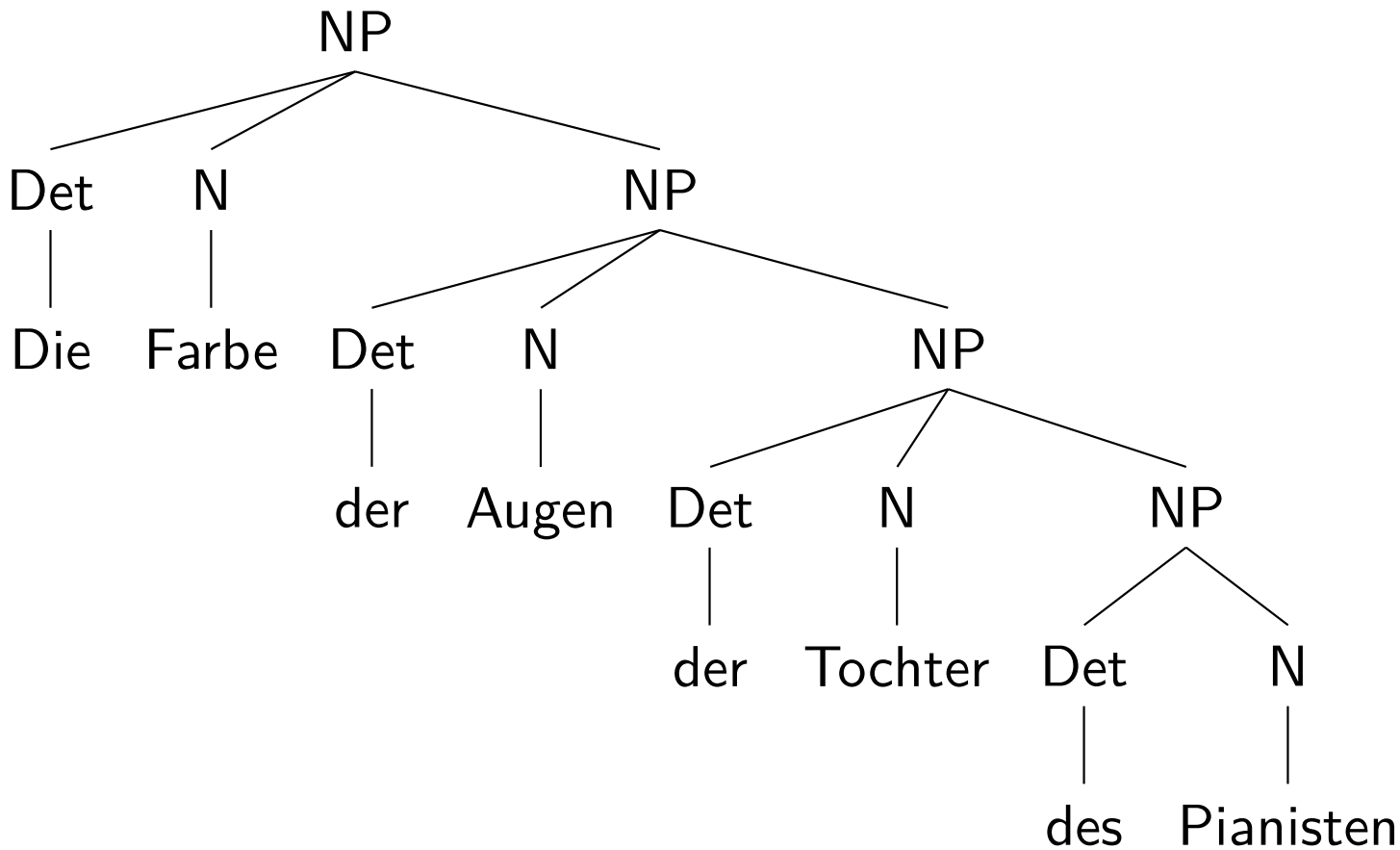

S
NP
VP
Det
N
V
PP
Die
Katze
sitzt
P
NP
auf
Det
N
der
Bank


S
NP
VP
Det
N
V
PP
Der
Bauer
arbeitet
P
NP
auf
Det
N
dem
Feld

Rekursion: Nichtterminal dominiert Instanz seiner selbst

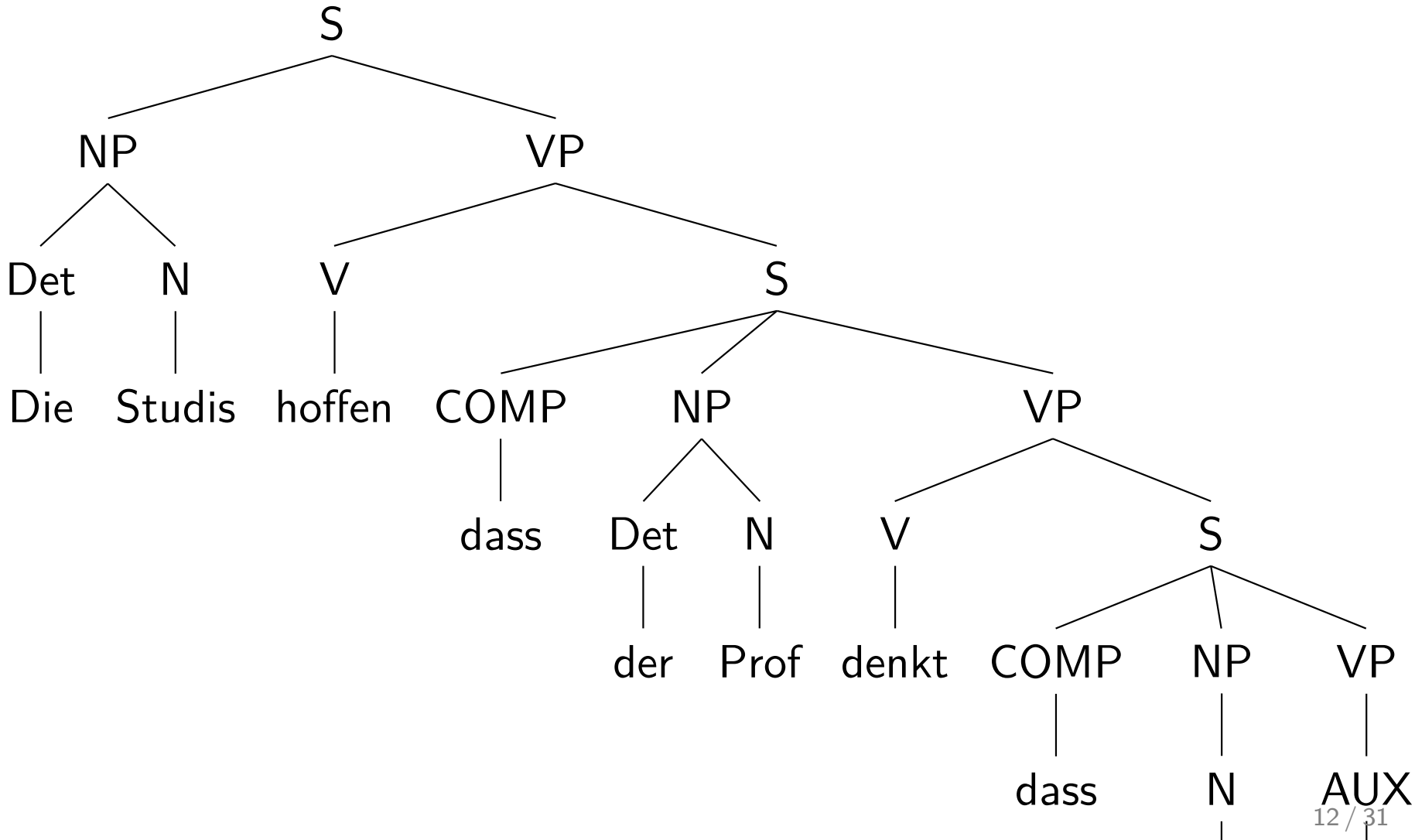
Fall 1: Ausgabe einer Regel dient als Eingabe der gleichen Regel

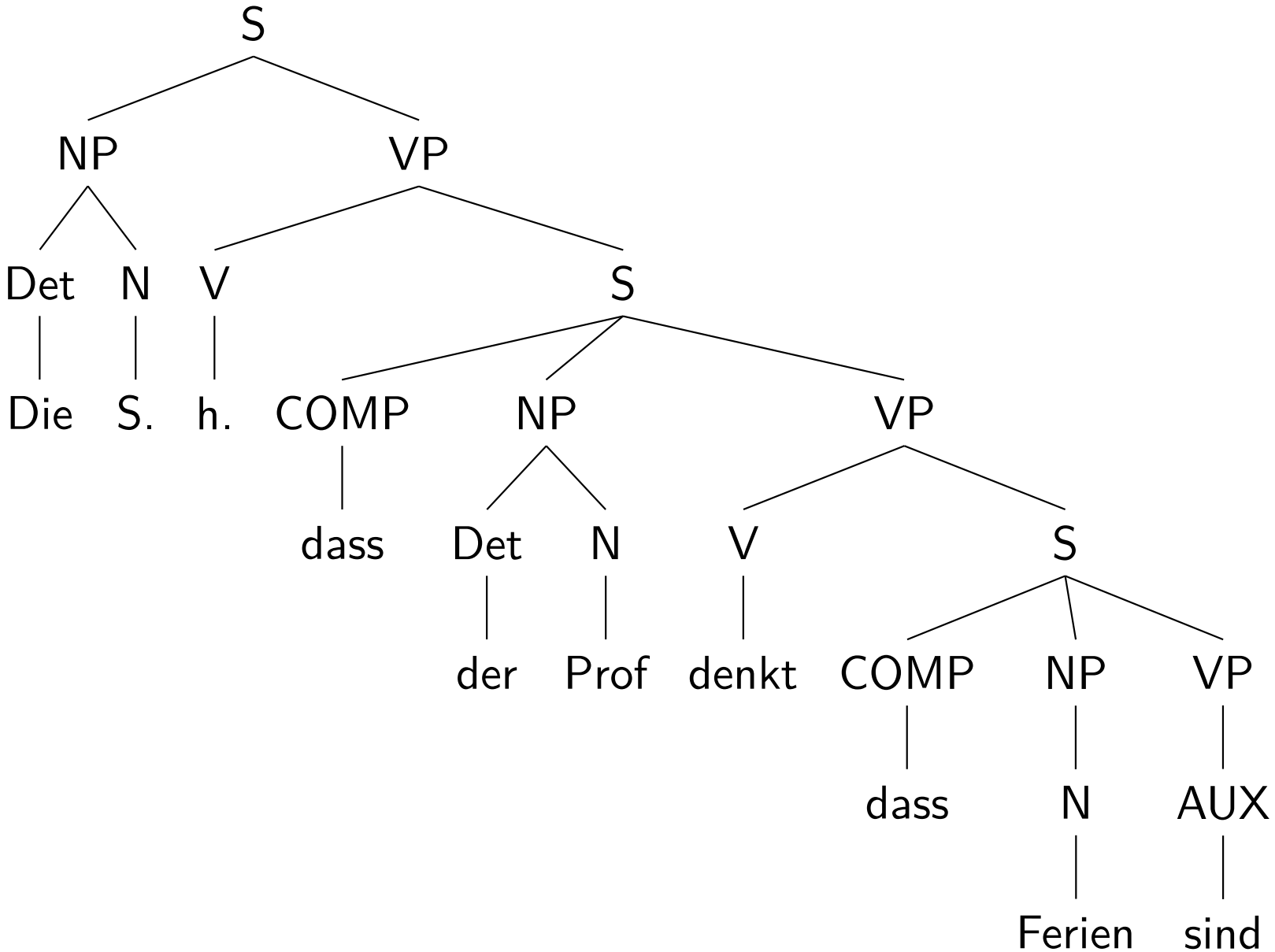
z.B. NP $\rightarrow$ Det N NP: “die Farbe der Augen der Tochter des Pianisten”



Rekursion: Nichtterminal dominiert Instanz seiner selbst (2)

Fall 2: Ausgabe einer Regel ist Eingabe einer "früheren" Regel: S $\rightarrow$ NP VP,
VP $\rightarrow$ V S: "Die Studis hoffen, dass der Prof denkt, dass Ferien sind"




S
NP
VP
Det
N
V
S
Die
S.
h.
COMP
NP
VP
dass
Det
N
V
S
der
Prof
denkt
COMP
NP
VP
dass
N
AUX
Ferien
sind

Lexikonregeln (1)

Zur Generierung der terminalen Kette *Der Löwe springt aus dem Käfig* sind noch weitere Regeln nötig: Det $\rightarrow$ der N $\rightarrow$ Löwe V $\rightarrow$ springt P $\rightarrow$ aus Det $\rightarrow$ dem N $\rightarrow$ Käfig

- zu lesen als: *Käfig* ist ein Exemplar der Menge der Nomen
- diese Regeln entsprechen einem Lexikon
- Mittels dieser Lexikonregeln werden die Wörter in die durch Phrasenstrukturregeln erzeugte Strukturbäume eingesetzt (lexikalische Einsetzung).

Lexikonregeln (2)

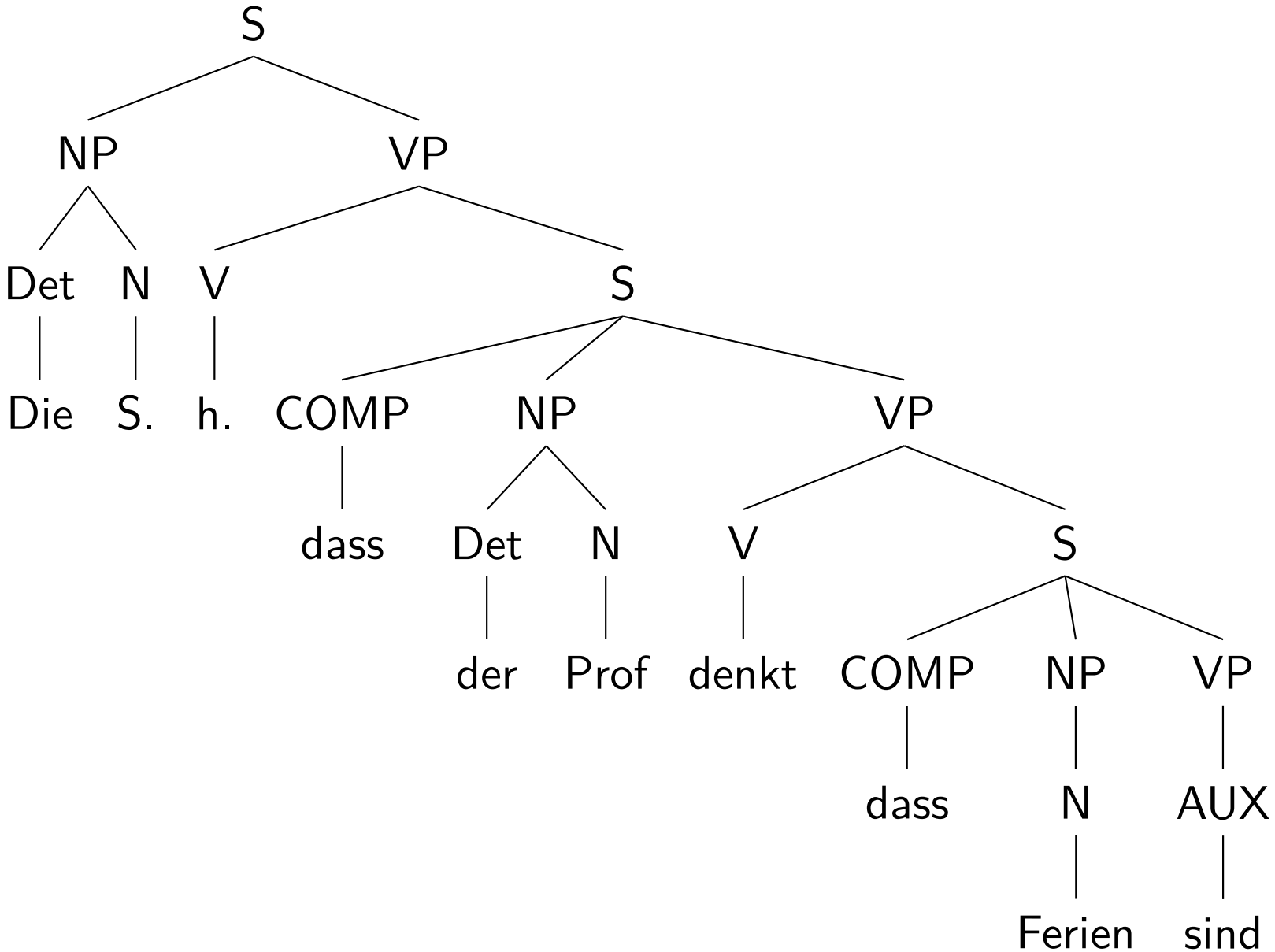
mehr Information nötig, weil: z.B. **Der Löwe schläft aus dem Käfig.* → Lexikoneintrag muss Informationen über den zugelassenen syntaktischen Rahmen des Verbs enthalten → Subkategorisierung (vgl. Rektionsmodell): z.B.: a. schlafen V, [___] b. helfen V, [___ NP] c. schenken V, [___ NP, NP]

Zwei Typen von Konstituenten

- Wortkategorien
  - z.B.: N, V, DET
  - schreiben ein Wortsymbol um als ein Wort
  - z.B.: DET $\rightarrow$ der, N $\rightarrow$ Hund
  - setzt Klassifizierung der Woerter voraus
- Phrasenkategorien
  - z.B.: NP, VP, S
  - schreiben ein Phrasensymbol um als Folge von einem oder mehreren Phrasensymbolen
  - z.B.: S $\rightarrow$ NP VP

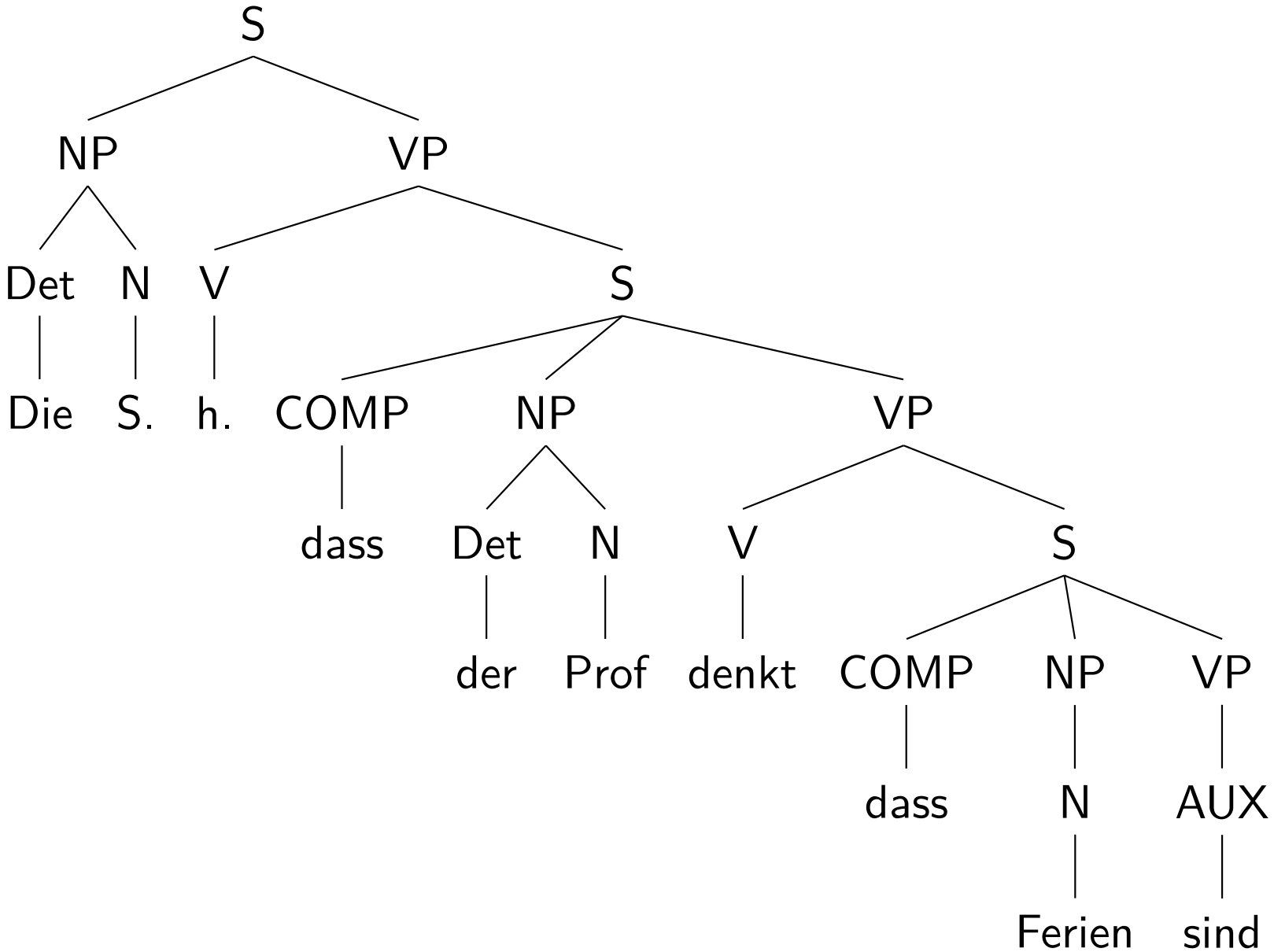
Wortkategorien / Wortarten

- terminale Kategorien
- dargestellt in präterminaler Kette
- bezieht sich auf terminale Kette: Wortformen (Elemente, die auf Satzebene nicht weiter zerteilt werden)


S
NP
VP
Det
N
V
S
Die
S.
h.
COMP
NP
VP
dass
Det
N
V
S
der
Prof
denkt
COMP
NP
VP
dass
N
AUX
Ferien
sind

Phrasenkategorien

- Konstituenten, die mehr als ein Wort enthalten: Phrasen (stellen grammatische Klassen dar)
- Phrase besteht aus Kopf (obligatorischer Kern) und optional aus weiteren (dominierten) Wörtern
- Klassifizierung nach der Wortart, zu der der Kopf einer Phrase gehört
- z.B. *großes Haus*: grammatische Eigenschaften wie *Haus* $\rightarrow$ *Haus* ist Kopf
- "Haus" ist N, also ist die Phrase NP
- $\rightarrow$ Kategoriensymbole ersetzen die Knoten im Baum $\rightarrow$ Konstituentenstruktur


S
NP
VP
Det
N
V
S
Die
S.
h.
COMP
NP
VP
dass
Det
N
V
S
der
Prof
denkt
COMP
NP
VP
dass
N
AUX
Ferien
sind

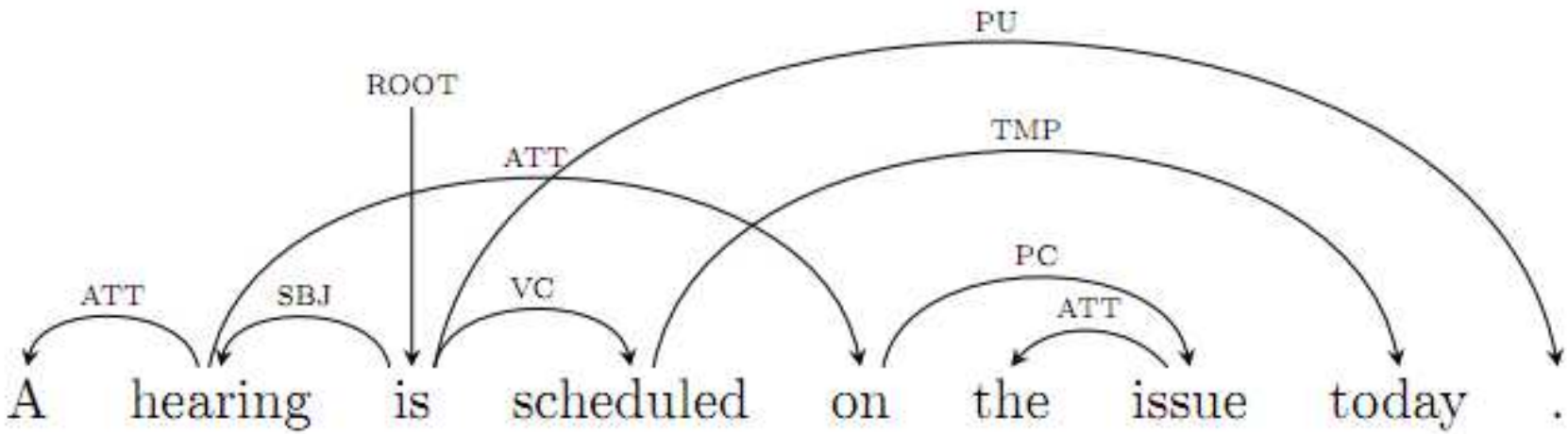
how would you extend this to model (i) agreement (“Peter rennen.”, “das Schule”) (ii) government (“vor den Schule”, “er half ihn”)?

Outline

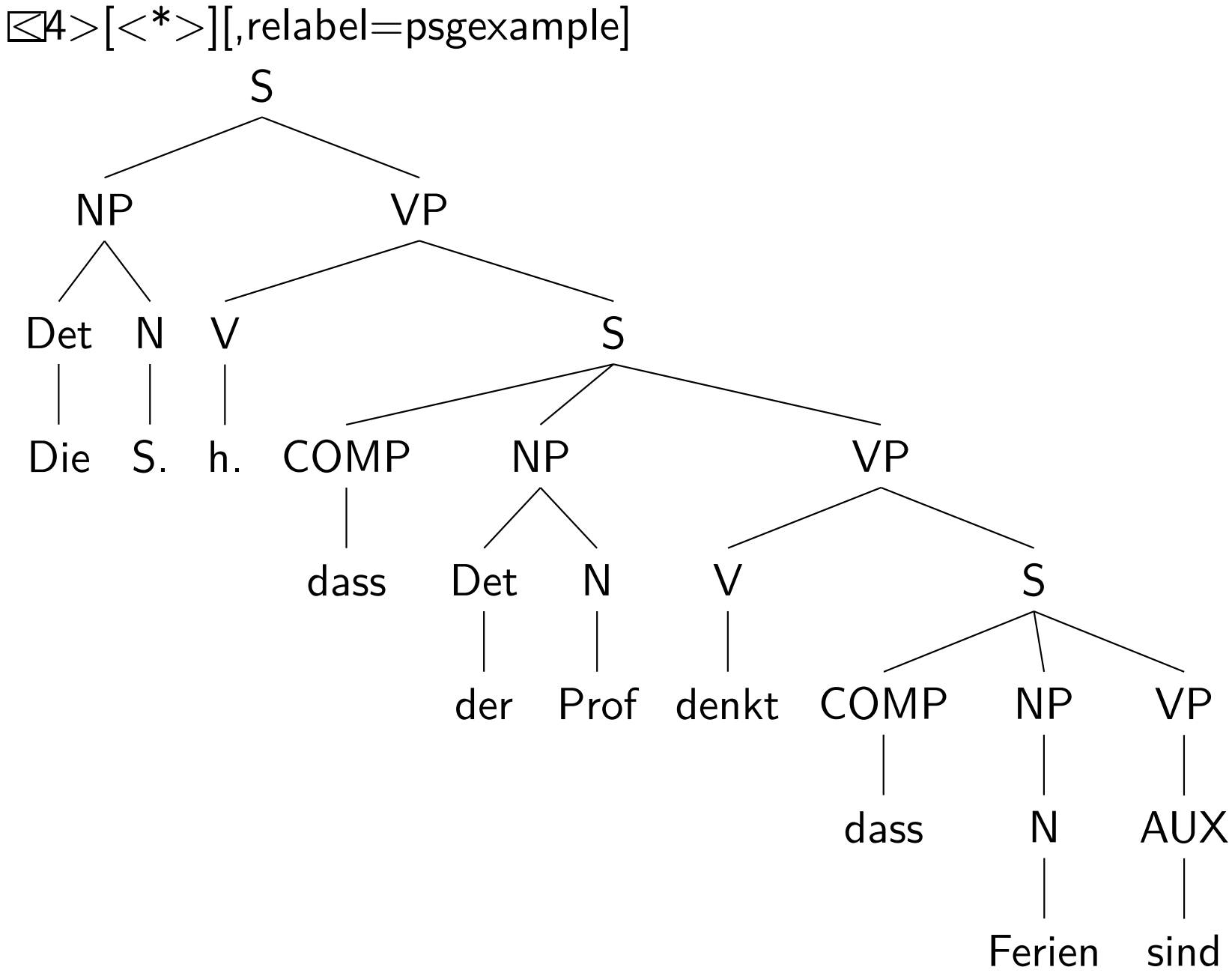
Parsing, Parsen

- syntaktische Analyse eines Textes, d.h. den Sätzen wird syntaktische Struktur zugeordnet
- Zwei grundlegende Typen von Strukturen: Dependenzbäume und Phrasenstrukturbäume
- Dependenzbäume zeigen die syntaktischen Abhängigkeiten der Wörter untereinander.
- Phrasenstrukturbäume zeigen die Konstituentenstruktur des Satzes.
  - setzt Analyse von Syntax als Konstituentenstruktur voraus
  - Ursprung: generative Grammatik (Chomsky)

Dependenzbaum



Phrasenstrukturbaum



Satzendeerkennung

- Feststellen des jeweiligen Satzendes
- wichtige Voraussetzung für das Parsen eines Textes
- Schwierigkeiten:
  - Punkt kennzeichnet keineswegs immer das Satzende , sondern auch Abkürzungen oder Nummerierungen (z.B. *Abschnitt 5.3*) u.ä.
  - nicht nur Punkt als Kennzeichen des Satzendes

Lokale Grammatiken

- Beschreibung der syntaktischen Struktur von jeweils ganz bestimmten Mustern von Wortfolgen (Phrasen).
- direkte, lokale Umgebung von Wörtern, die durch die Syntax dieser Wörter bestimmt ist (in welcher Form folgen die Wörter aufeinander?)
- durch Graphen als endliche Automaten darstellbar
- Unitex zur Entwicklung und Anwendung lokaler Grammatiken (http://www-igm.univ-mlv.fr/~unitex/download.html)

Anwendungen

- Informationsextraktion (insbesondere lokale Grammatiken)
- Information Retrieval (protein-protein interactions)
- Maschinelle Übersetzung
  - (oft) nicht einzelne Wörter aus der Quellsprache übersetzen
  - sondern größere Einheiten von zwei oder mehr Wörtern
  - z.B. bei Kollokationen (*schwer verletzen* – engl. *to heart badly* und nicht **to heart heavily*)
  - oder bei Idiomen (engl. *to kick the bucket* – *ins Gras beißen* und nicht **den Eimer treten*)
  - unterschiedliche Wortordnung richtig übersetzen

Parser im Netz

- Stanford Parser (Phrasenstruktur, Dependenzbaum; Demo):
  http://nlp.stanford.edu:8080/parser/
- ETAP (Dependenzbaum; Demo):
  http://proling.iitp.ru/etap3
- MSTParser (Dependenzbaum; download):
  http://sourceforge.net/projects/mstparser/
- MaltParser (Dependenzbaum; download):
  http://www.maltparser.org/

write down parse tree for “I saw a man with an umbrella.”

Take-away

- Phrasenstrukturgrammatik: Formaler Ansatz zur Darstellung der Struktur von Sätzen
- Zwei Arten von Regeln: Phrasenstrukturregeln, Lexikonregeln
- Anwendungen von Syntax in der Computerlinguistik: Parsing, Maschinelle Übersetzung etc.